



**Central Asian
Scientific
Journal**

**VOL 4(32)
2026**



ASTANA

ISSN: 3135-3061



Central Asian Scientific Journal

Халықаралық электрондық ғылыми журнал
Международный электронный научный журнал

ENGLISH

Published since 2021
One issue per quarter, 3 monthly
volumes

ҚАЗАҚША

2021 жылдан бері шығарылады
Тоқсанына 1 шығарылым, 3 том
— ай сайын

РУССКИЙ

Издается с 2021 года
1 выпуск в квартал, 3 тома —
ежемесячно

VOLUME 2 · ISSUE 4 (32) · NOVEMBER 2026

Том 2 · №4 (32) · Ноябрь / Қараша 2026

IN THIS ISSUE

В НОМЕРЕ · НӨМІРДЕ

Reported bribe requests among firms in kazakhstan, 2019-2024: direct and indirect questions in the world bank enterprise surveys

8

Duisen Nazym

How many open robot datasets can actually train a robot? A census of 400 lerobot datasets on hugging face

34

Azamat Sagadat

Морфологически ориентированная методика тематического моделирования казахоязычной поэзии: проектирование исследования на материале произведений М. Макатаева

79

Қ.Д. Сәрсембай

How many maps of Z/NZ preserve multiplication? A counting formula for OEIS sequence A117986

94

Onal Yerkanat

EDITORIAL BOARD

EDITOR-IN-CHIEF:

Talgat Zh. Baidildinov — Candidate of Pedagogical Sciences, Abai Kazakh National Pedagogical University

EDITORIAL BOARD:

Radwan Labban — Member of RINA (UK) and SNAME (USA), Plymouth College, United Kingdom

Madina A. Yermaganbetova — Candidate of Pedagogical Sciences, Associate Professor, L.N. Gumilyov Eurasian National University

Anar A. Mukasheva — Doctor of Law, Professor, L.N. Gumilyov Eurasian National University

Dametken S. Baigozhanova — Candidate of Pedagogical Sciences, Honored Professor of Kazakhstan, Academician of the International Academy of Informatization

Galym B. Teleuev — PhD, Campus Dean, Kazakh-American University

Aruzhan K. Mektepbayeva — MSc in Computer Science, Lecturer, Astana IT University

Alexey D. Fedenko — MSc in Computer Science, Lecturer, Astana IT University

Central Asian Scientific Journal

International electronic scientific journal

ISSN 3135-3061

Published since 2021 (registered 25.08.2021)

Periodicity: One issue per quarter, 3 monthly volumes

Languages of publication: Kazakh, Russian, English

Thematic focus: Pedagogical, social and humanitarian, technical, economic and legal sciences; information and communication technologies; theoretical and applied research

Owner: Sole Proprietor «DOC»

Registered with the Information Committee of the Ministry of Information and Public Development of the Republic of Kazakhstan, certificate No. KZ77VPY00147053 dated 17.04.2026.

Editorial office: 010010, Republic of Kazakhstan, Astana (closed-type office)

info@cajournal.kz · cajournal.kz

The authors are responsible for the accuracy of the published information, quotations and other statements.

© Central Asian Scientific Journal, 2026

РЕДАКЦИЯ

БАС РЕДАКТОР:

Байдильдинов Талгат Жарылкасынович — п.ғ.к., Абай атындағы Қазақ ұлттық педагогикалық университеті

РЕДАКЦИЯЛЫҚ АЛҚА:

Radwan Labban — RINA (Ұлыбритания) және SNAME (АҚШ) мүшесі, Plymouth College, Ұлыбритания

Ермаганбетова Мадина Аскарровна — п.ғ.к., доцент, Л.Н. Гумилев атындағы Еуразия ұлттық университеті

Мукашева Анар Абайханқызы — з.ғ.д., профессор, Л.Н. Гумилев атындағы Еуразия ұлттық университеті

Байгожанова Даметкен Сагидуллаевна — п.ғ.к., Қазақстанның құрметті профессоры, Халықаралық ақпараттандыру академиясының академигі

Телеуев Галым Байгазиевич — PhD, кампус деканы, Қазақ-Америка университеті

Мектепбаева Аружан Кайратовна — компьютерлік ғылымдар магистрі, оқытушы, Astana IT University

Феденко Алексей Демьянович — компьютерлік ғылымдар магистрі, оқытушы, Astana IT University

Central Asian Scientific Journal

Халықаралық электрондық ғылыми журнал

ISSN 3135-3061

Шығарылады: 2021 ж. бастап (тіркелген күні 25.08.2021)

Мерзімділігі: Тоқсанына 1 шығарылым, 3 том — ай сайын

Жарияланым тілдері: қазақ, орыс, ағылшын

Тақырыптық бағыты: Педагогикалық, әлеуметтік-қоғамдық, техникалық, экономикалық және заң ғылымдары, ақпараттық-коммуникациялық технологиялар, теориялық және ғылыми-практикалық зерттеулер

Меншік иесі: «DOC» ЖК

Қазақстан Республикасы Ақпарат және қоғамдық даму министрлігінің Ақпарат комитетінде тіркелген, 17.04.2026 ж. № KZ77VPY00147053 куәлік.

Редакция мекенжайы: 010010, Қазақстан Республикасы, Астана қ., (жабық үлгідегі кеңсе)

info@cajournal.kz · cajournal.kz

Жарияланатын ақпараттың, дәйексөздердің және өзге де мәліметтердің дұрыстығы үшін авторлар жауапты.

© Central Asian Scientific Journal, 2026

РЕДАКЦИЯ

ГЛАВНЫЙ РЕДАКТОР:

Байдильдинов Талгат Жарылкасынович — к.п.н., Казахский национальный педагогический университет им. Абая

РЕДАКЦИОННАЯ КОЛЛЕГИЯ:

Radwan Labban — Член RINA (Великобритания) и SNAME (США), Plymouth College, Великобритания

Ермаганбетова Мадина Аскарловна — к.п.н., доцент, Евразийский национальный университет им. Л.Н. Гумилёва

Мукашева Анар Абайханкызы — д.ю.н., профессор, Евразийский национальный университет им. Л.Н. Гумилёва

Байгожанова Даметкен Сагидуллаевна — к.п.н., почётный профессор Казахстана, академик МАИН

Телеуев Галым Байгазиевич — PhD, декан кампуса, Казахско-Американский университет

Мектепбаева Аружан Кайратовна — магистр компьютерных наук, преподаватель, Astana IT University

Феденко Алексей Демьянович — магистр компьютерных наук, преподаватель, Astana IT University

Central Asian Scientific Journal

Международный электронный научный журнал

ISSN 3135-3061

Издаётся с 2021 г. (регистрация 25.08.2021)

Периодичность: 1 выпуск в квартал, 3 тома — ежемесячно

Языки публикаций: казахский, русский, английский

Тематическая направленность: Педагогические, общественно-социальные, технические, экономические и юридические науки, информационно-коммуникационные технологии, теоретические и научно-практические исследования

Собственник: ИП «DOC»

Зарегистрирован в Комитете информации Министерства информации и общественного развития Республики Казахстан, свидетельство № KZ77VPY00147053 от 17.04.2026 г.

Адрес редакции: 010010, Республика Казахстан, г. Астана, (офис закрытого типа)

info@cajournal.kz · cajournal.kz

За достоверность публикуемой информации, цитат и иных сведений ответственность несут авторы.

© Central Asian Scientific Journal, 2026

МАЗМҰНЫ — СОДЕРЖАНИЕ — CONTENTS

ӘЛЕУМЕТТІК-ЭКОНОМИКАЛЫҚ ЖӘНЕ ҚҰҚЫҚ ҒЫЛЫМДАРЫ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИЕ И ЮРИДИЧЕСКИЕ НАУКИ SOCIO-ECONOMIC AND LEGAL SCIENCES

Duisen Nazym

REPORTED BRIBE REQUESTS AMONG FIRMS IN KAZAKHSTAN, 2019-2024: DIRECT AND
INDIRECT QUESTIONS IN THE WORLD BANK ENTERPRISE SURVEYS..... 8

Chen Jun

SYNDICATED AND BLENDED FINANCING FOR GREEN INFRASTRUCTURE IN
KAZAKHSTAN: RISK ALLOCATION AND BANKABILITY 12

Wang Liang

GREENFIELD DEVELOPMENT OF CANAL-BASED SMALL HYDROPOWER IN
UZBEKISTAN: SITE SCREENING, PPP RISK ALLOCATION AND CLIMATE RESILIENCE
THROUGH THE WATER-ENERGY-FOOD NEXUS..... 23

ТЕХНИКАЛЫҚ ҒЫЛЫМДАР ТЕХНИЧЕСКИЕ НАУКИ TECHNICAL SCIENCE

Azamat Sagadat

HOW MANY OPEN ROBOT DATASETS CAN ACTUALLY TRAIN A ROBOT? A CENSUS OF
400 LEROBOT DATASETS ON HUGGING FACE..... 34

Cheng Che

FROM 100 MW DEMONSTRATIONS TO A 1 GW WIND AND STORAGE PLATFORM:
SCALING RENEWABLE ENERGY DEVELOPMENT IN KAZAKHSTAN..... 39

Б.С. Сапарғалиев, Д.С. Байгожанова, Н.С. Ермакова

ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИЙ РАСПОЗНАВАНИЯ ЛИЦ ДЛЯ ПРЕДОТВРАЩЕНИЯ
СПИСЫВАНИЯ В СИСТЕМАХ ДИСТАНЦИОННОГО КОНТРОЛЯ ЗНАНИЙ..... 50

Zaikenov Temirlan

A HYBRID MODEL FOR AUTOMATIC NEWS MESSAGE CLASSIFICATION..... 56

ПӘНДЕР АРАЛЫҚ ҒЫЛЫМДАР МЕЖДИСЦИПЛИНАРНЫЕ НАУКИ INTERDISCIPLINARY SCIENCES

Қ.Д. Сәрсембай

МОРФОЛОГИЧЕСКИ ОРИЕНТИРОВАННАЯ МЕТОДИКА ТЕМАТИЧЕСКОГО
МОДЕЛИРОВАНИЯ КАЗАХОЯЗЫЧНОЙ ПОЭЗИИ: ПРОЕКТИРОВАНИЕ ИССЛЕДОВАНИЯ
НА МАТЕРИАЛЕ ПРОИЗВЕДЕНИЙ М. МАКАТАЕВА..... 79

Onal Yerkanat

HOW FAST IS KAZAKHSTAN LEAVING COAL? THE COAL SHARE OF ELECTRICITY IN
KAZAKHSTAN COMPARED WITH CHINA, RUSSIA, UZBEKISTAN AND KYRGYZSTAN,
2000-2025..... 87

ЕСТЕСТВЕННЫЕ НАУКИ ЖАРАТЫЛЫСТАНУ ҒЫЛЫМДАРЫ NATURAL SCIENCES

Onal Yerkanat

HOW MANY MAPS OF Z/NZ PRESERVE MULTIPLICATION? A COUNTING FORMULA FOR
OEIS SEQUENCE A117986..... 94

**ӘЛЕУМЕТТІК ЖӘНЕ ГУМАНИТАРЛЫҚ ҒЫЛЫМДАР
СОЦИАЛЬНО-ГУМАНИТАРНЫЕ НАУКИ
SOCIAL SCIENCES AND HUMANITIES**

Saltanat Mussina, Torgyn Yeralinova

LANGUAGE OF SCHOOLING AND PERCEPTIONS OF ENGLISH ACCENTED SPEECH IN
KAZAKHSTAN.....

101



ӘЛЕУМЕТТИК-ЭКОНОМИКАЛЫҚ ЖӘНЕ ҚҰҚЫҚ ҒЫЛЫМДАРЫ
СОЦИАЛЬНО-ЭКОНОМИЧЕСКИЕ И ЮРИДИЧЕСКИЕ НАУКИ
SOCIO-ECONOMIC AND LEGAL SCIENCES

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. No. 4 (32), Vol. 2, pp. 8–11
DOI: <https://doi.org/10.5281/zenodo.23763856>
Journal website: <https://cajournal.kz>

UDC 328.185

REPORTED BRIBE REQUESTS AMONG FIRMS IN KAZAKHSTAN, 2019-
2024: DIRECT AND INDIRECT QUESTIONS IN THE WORLD BANK
ENTERPRISE SURVEYS

Duisen Nazym

Nazarbayev Intellectual School of Natural Sciences and Mathematics, Abay District, Shymkent, Kazakhstan

Duisen Nazym — Grade 12 student, Nazarbayev Intellectual School of Natural Sciences and Mathematics, Abay District, Shymkent, Kazakhstan.

© Duisen Nazym, 2026

Abstract: In the World Bank Enterprise Surveys, the share of Kazakhstani firms asked for a bribe rose from 11.6% in 2019 to 16.4% in 2024. This can mean more bribery or more candid managers. Following the author's rule, the paper compares a direct question about the firm's own experience (J.5) with an indirect one about "establishments like this one" (J.7) in Kazakhstan, Uzbekistan and the Kyrgyz Republic. In Kazakhstan both rose (by 5.3 and 4.5 points), but the difference (+0.8, 95% interval -21.4 to +23.0) favours neither reading, also at the lower design effects seen in the neighbours' surveys. Changes by transaction cannot be confirmed with the published standard errors. The author's forecast for the neighbours held in the point estimates.

Keywords: corruption; bribery; respondent reticence; World Bank Enterprise Surveys; Kazakhstan; Central Asia.

1. Introduction

The World Bank Enterprise Surveys ask managers whether a gift or informal payment was expected from their firm when it dealt with the state. In Kazakhstan the share of firms that met at least one such request in six common transactions (bribery incidence) rose from 11.6% in 2019 to 16.4% in 2024 [1]. This looks like more bribery, but the figure would also rise if managers became more willing to admit requests that were already being made. This reluctance to answer truthfully is called respondent reticence [2].

The author set out both readings in her hypothesis: firms still face procedures where "opportunities for informal payments may still exist", but "businesses are now more willing to report bribery requests than before". Her rule for telling them apart: "If only the question about whether a firm was asked for a bribe increases, while other indicators of corruption and firms' interactions with government remain stable", this "could suggest" more candour. If several indicators rise together, "this would be more consistent with a real increase in bribery".

The surveys make this rule testable. Question J.5 asks about the firm's own experience: "In any of these inspections or meetings was a gift or informal payment expected or requested?" Question J.7 asks what "establishments like this one" pay, which does not accuse the respondent's own firm [3; 4]. If managers had only become more candid, J.5 should rise while J.7 stays flat. If bribery had become more common, both should rise.

Research question. Did the 2019-2024 rise in reported bribe requests in Kazakhstan reflect more bribery or more candid answers, and in which interactions with the state did it occur?



This paper argues that the published data cannot yet answer the first part: the two answers rose together, but too uncertainly to favour either reading.

2. Literature Review

The test rests on one finding: reluctant respondents distort answers about their own firm much more than answers about similar firms. Clausen, Kraay and Murrell, building on a Romanian study, identified 13.1% of respondents in the Enterprise Survey for Nigeria as highly likely to be reticent [2]. They admitted sensitive acts at 61% of the others' rate when asked about their own firm, but answered virtually like everyone else on the question corresponding to J.7. This study and [5] predate the data; they are used only for the method, and a non-exhaustive search found no later study with this test.

Reticence can also hide in a question's structure. J.5 is asked only of firms that report contact with tax officials, and denying the contact avoids the bribe question [5]. Tax contact must therefore be read with J.5.

A recent study of Kazakhstan shows where firms meet corruption, but not how it changes. In a 2023 survey of 702 entrepreneurs, Kudasheva and Rakhmatullayeva found construction, energy, extractive industries and logistics most exposed through licensing, taxation and customs [6]. As the author noted, such surveys differ from the Enterprise Surveys, which use the same method in every country and so allow comparison with neighbours.

3. Data and Method

The paper analyses published country-level indicators; no data were collected and no ethical approval was needed. Indicators, standard errors and numbers of firms were downloaded from the Enterprise Surveys portal on 13 October 2026 [1]. The surveys represent formal private firms with five or more employees and are stratified and weighted [7; 8]. The latest waves cover 1,446 and 1,013 firms in Kazakhstan (2019, 2024), 1,239 and 1,008 in Uzbekistan (2019, 2024) and 360 and 354 in the Kyrgyz Republic (2019, 2023) [1]. After 2019, Kazakh firms were surveyed only in two COVID-19 follow-ups (2021 and 2022), which did not ask about bribes, so 2019 is the only recent starting point.

Measures. Bribery incidence and J.5 are direct; J.7, the share saying similar firms pay "to get things done", is indirect; the share naming corruption a major obstacle measures perception [7]. The wording of J.5 and J.7 is essentially the same in both waves [3; 4].

The test. The main comparison is the change in J.5 against the change in J.7. J.5 is answered by the firms that met tax officials, a subset of those answering J.7, so the two estimates are correlated by an unpublished amount; intervals are shown for correlations of 0 and of 1, the second only as the narrowest possible width.

Precision. The standard error of a change combines the two published ones, treating the waves as independent. Kazakhstan's standard errors imply design effects of 3.5 to 17.8 (neighbours: 1.7 to 4.0), and their method is undocumented. They are used as published, except in four transaction cells where they are 0 or implausibly small. The main changes are also recomputed at assumed design effects of 1 to 3. No correction is made for multiple comparisons.

4. Results

4.1 The direct and indirect answers rose together

Between 2019 and 2024 every main corruption measure in Table 1 rose: J.5 by 5.3 points and J.7 by 4.5. This is the pattern the rule links to more bribery, not candour. But the change in J.5 minus the change in J.7 is +0.8 points, with a 95% interval of -21.4 to +23.0 (or -13.7 to +15.3 at the narrowest bound). Both intervals contain a J.5-only rise as large as the one observed (+5.3), so the test cannot choose. With these standard errors it could only detect a gap of about 21 to 32 points, so this is no evidence that the gap is zero.

**Table 1. Kazakhstan: share of firms, % (published standard error), and the 2019-2024 change in percentage points (95% interval).**

Measure	2019	2024	Change	p
Bribery incidence (direct)	11.6 (5.4)	16.4 (6.4)	+4.8 (-11.6; +21.2)	0.57
J.5 tax meetings (direct)	11.9 (6.3)	17.2 (8.8)	+5.3 (-15.9; +26.5)	0.62
J.7 similar firms pay (indirect)	7.9 (2.1)	12.4 (2.7)	+4.5 (-2.2; +11.2)	0.19
Corruption a major obstacle	12.0 (2.8)	16.1 (2.2)	+4.1 (-2.9; +11.1)	0.25
Met or visited by tax officials	30.9 (4.2)	27.7 (3.1)	-3.2 (-13.4; +7.0)	0.54
J.5 change minus J.7 change			+0.8 (-21.4; +23.0)	0.94
<i>By transaction (firms in 2024):</i>				
Operating licence (97)	8.0 (1.6)	11.0 (9.0)	+3.0 (-14.9; +20.9)	0.74
Import licence (24)	24.3 (*)	0.9 (*)	-23.4	
Construction permit (49)	14.6 (8.0)	0.1 (*)	-14.5	
Electrical connection (22)	4.4 (3.6)	22.1 (*)	+17.7	

Note: source [1]; own calculations. Firms answering in 2019 and 2024: incidence 615, 422; J.5 358, 327; J.7 1,356, 923. Test row: no correlation assumed. Tax meetings are J.5; water connections: no 2024 value. * Standard error 0 or implausibly small.

With the published standard errors, no single rise exceeds sampling error (Table 1). This rests on Kazakhstan's undocumented standard errors: at any design effect up to 3.1, which covers most of the neighbours' range, the J.7 rise would be significant ($p = 0.047$ at 3). The main test would not (+0.8, -9.4 to +11.0). J.5 rests on a quarter to a third of J.7's firms, which caps the correlation of the two estimates near 0.5-0.6. At the neighbours' design effects (1.7 and above) a correlation of at least 0.84 would be needed to exclude a J.5-only rise, so the test contains both readings there. Only at a design effect near 1, far below the 3.5 to 17.8 implied by Kazakhstan's published standard errors, would a correlation of 0.53, within that cap, exclude it. With the published standard errors and at the neighbours' design effects, the test is inconclusive.

4.2 Where the change sits

The point estimates rose for tax meetings, operating licences and electrical connections and fell for import licences and construction permits (Table 1). With the published standard errors none of these changes can be confirmed. At the neighbours' design effects the fall for construction permits (49 firms in 2024, with a published 2024 standard error of 0) would exceed sampling error. The 2024 questionnaire has no question on bribes at customs.

4.3 The neighbours and the author's forecast

In Uzbekistan J.5 rose from 6.7% to 8.3% while J.7 fell from 7.3% to 6.7%, the pattern of candour, but within sampling error (difference +2.2, -7.2 to +11.6). In the Kyrgyz Republic the bribery answers barely changed.

Before this paper's figures were computed, the author forecast: "I would predict that the Kyrgyz Republic may have the highest share of firms answering 'Yes' to the direct J.5 question, while Uzbekistan may have a lower share." It held in the point estimates: J.5 is 27.9% in the Kyrgyz Republic, 17.2% in Kazakhstan and 8.3% in Uzbekistan. The Kyrgyz lead over Uzbekistan is clear (19.6 points, 9.8 to 29.4); over Kazakhstan (10.7, -8.0 to +29.4) it is within sampling error.

5. Discussion

Beyond sampling error, four features of the surveys weaken the test. First, as the author anticipated, the questions differ: J.5 covers tax meetings among the 28-31% of firms that had them, J.7 any payment by similar firms. Second, denying a tax visit avoids the bribe question [5], so the slight fall in tax contact may be real or reticent. Third, refusals count as "yes" in bribery incidence and J.7 (for J.5 this is not stated) [7], so more refusals would raise both, the very pattern observed. Fourth, the 2024 questionnaire was revised and dropped J.6 and the customs question, which could also shift answers. By transaction, the point estimates rose for tax meetings and operating licences, procedures Kudasheva and Rakhmatullayeva name [6], and fell for construction permits, although construction is one of the sectors they name.



6. Conclusion

The 2019-2024 rise in reported bribery in Kazakhstan cannot yet be read as either more bribery or more candour. Firm-level files could show whether refusals changed, and a randomized response question in the next wave [2] would measure reticence directly.

Author contributions and use of AI tools

The author chose the topic and question from her mentor's options (14 and 16 September 2026). She wrote the hypothesis, rule and forecast, found studies [2; 6] and nine others left out here, and approved the analysis plan, which named three waves (two are used here). She had seen the Kazakh headline figures and the direction of change in some neighbours before writing the hypothesis. She used ChatGPT to check statistics and, in places, to phrase her thoughts. Her mentor turned the rule into the test and, with a generative AI tool (Claude, Anthropic), downloaded the data, wrote the analysis code and drafted the English text around her quoted words. No number, quotation or reference was produced by a language model without being checked.

Data availability

All data are public [1; 3; 4; 7]. The analysis scripts (Python) are available on request.

Conflicts of interest

The author declares no conflicts of interest. No funding was received.

References:

- [1] World Bank. Enterprise Surveys: data portal, Custom Query and survey list [Electronic resource]. Washington, DC: World Bank, 2026. URL: <https://www.enterprisesurveys.org/en/data> (accessed 13.10.2026).
- [2] Clausen B., Kraay A., Murrell P. Does respondent reticence affect the results of corruption surveys? Evidence from the World Bank Enterprise Survey for Nigeria. Policy Research Working Paper No. 5415. Washington, DC: World Bank, 2010. DOI: 10.1596/1813-9450-5415.
- [3] EBRD, EIB, World Bank Group. Enterprise Survey: services module (2018) [Electronic resource]. URL: <https://www.ebrd.com/home/what-we-do/office-of-the-chief-economist/beeps/business-environment-enterprise-performance-survey-data.html> (accessed 13.10.2026).
- [4] World Bank. The World Bank Enterprise Survey 2023; The World Bank Enterprise Survey 2024: questionnaires [Electronic resource]. URL: <https://www.enterprisesurveys.org/en/methodology> (accessed 13.10.2026).
- [5] Karalashvili N., Kraay A., Murrell P. Doing the survey two-step: the effects of reticence on estimates of corruption in two-stage survey questions. Policy Research Working Paper No. 7276. Washington, DC: World Bank, 2015. DOI: 10.1596/1813-9450-7276.
- [6] Kudasheva T., Rakhmatullayeva D. Evaluating the impact of corruption on the development of entrepreneurship in Kazakhstan // The Political Economy of Legal and Governance Reform in Central Asia / ed. by K. Ruziev, R. Urinboyev. Cham: Springer Nature Switzerland, 2026. P. 391-415. DOI: 10.1007/978-3-032-28163-0_15.
- [7] World Bank Enterprise Surveys. Indicator descriptions, version of 24 July 2026 [Electronic resource]. URL: <https://www.enterprisesurveys.org/content/dam/enterprisesurveys/documents/methodology/Indicator-Description.pdf> (accessed 13.10.2026).
- [8] World Bank Enterprise Surveys. Sampling methodology (February 2022) [Electronic resource]. URL: https://www.enterprisesurveys.org/content/dam/enterprisesurveys/documents/methodology/Sampling_Note-Consolidated-2-16-22.pdf (accessed 13.10.2026).



ӘЛЕУМЕТТІК-ЭКОНОМИКАЛЫҚ ЖӘНЕ ҚҰҚЫҚ ҒЫЛЫМДАРЫ
СОЦИАЛЬНО-ЭКОНОМИЧЕСКИЕ И ЮРИДИЧЕСКИЕ НАУКИ
SOCIO-ECONOMIC AND LEGAL SCIENCES

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. No. 4 (32), Vol. 2, pp. 12–22
DOI: <https://doi.org/10.5281/zenodo.23770168>
Journal website: <https://cajournal.kz>

UDC 336

SYNDICATED AND BLENDED FINANCING FOR GREEN
INFRASTRUCTURE IN KAZAKHSTAN: RISK ALLOCATION AND
BANKABILITY

Chen Jun

Al-Farabi Kazakh National University, Kazakhstan

Chen Jun — DBA candidate.

© Chen Jun, 2026

Abstract: Large green infrastructure projects in Kazakhstan may exceed the tenor or exposure capacity of a single lender class. This paper examines how syndicated lending and blended finance can support project bankability through lender specialisation, maturity extension, currency matching and structured risk allocation. The study combines institutional analysis, peer-reviewed research, comparative public case evidence and a fully specified illustrative project-finance model. Four Kazakhstan cases are considered: Zhanatas Wind Farm, the KEGOC grid reinforcement programme, the Big Almaty Ring Road concession as a non-green financing comparator, and the Mirny wind and battery project. The cases illustrate how multilateral development banks can provide long-tenor anchor debt, mobilise commercial lenders through A/B or club structures, and introduce common environmental, social and governance disciplines. Under stated assumptions for a USD 700 million project, a 15-year commercial structure produces annual debt service of USD 62.78 million and a debt service coverage ratio of 1.19. A syndicated A/B structure produces a ratio of 1.37, while a blended structure with a concessional tranche produces 1.46. The improvements reflect both specified funding costs and differentiated tenors; syndication alone does not guarantee cheaper funding. The paper proposes a bankability framework supported by common due diligence, intercreditor rules, transparent monitoring and a documented test of concessional additionality.

Keywords: Kazakhstan; green infrastructure; syndicated lending; blended finance; project finance; A/B loan; risk allocation; bankability

JEL classification: G21; G32; F34; Q42; Q48

1. Introduction

Kazakhstan is entering a capital-intensive stage of its energy transition. Renewable energy sources accounted for 7.0 percent of electricity generation in 2025, while installed renewable capacity reached 3,236.7081 MW, including 1,833.6 MW of wind and 1,234.7361 MW of solar capacity [1-2]. Investment in electricity, gas, steam, hot water and air-conditioning supply reached KZT 2.2261 trillion in 2025, equivalent to 9.5 percent of national fixed capital investment [3]. Grid expansion, storage, flexible generation and low-carbon municipal infrastructure must develop alongside new generation if additional renewable capacity is to remain reliable and economically usable.

The financing structure has not yet expanded at the same rate as the physical investment programme. In 2025, national fixed capital investment was KZT 23.4925 trillion and bank loans financed 4.6 percent, of which loans from foreign banks accounted for 0.3 percent of total investment [3]. AIFC and Green Finance Centre publications document continued development of Kazakhstan's sustainable-finance instruments [4-5], while the 2024 to 2029 National Infrastructure



Plan covers approximately USD 80 billion of overall energy, transport, water and digital infrastructure, only part of which is directly related to green transition [6].

This gap is partly a problem of risk capacity. A large green infrastructure project concentrates construction, market, currency, regulatory, technology and environmental risks within one special purpose vehicle. A single commercial bank may face limits on tenor, country exposure, sector concentration, single-borrower exposure and foreign currency liquidity. Sponsors may also resist an equity contribution large enough to compensate for all of those constraints. Syndication distributes credit exposure, while blended finance combines commercial and development capital with different risk and return objectives.

This paper asks three questions. First, which financing instruments are suitable for the main risks of Kazakhstan's green infrastructure? Second, what lessons emerge from public Kazakhstan transactions? Third, under a consistent illustrative model, how do traditional, syndicated and blended structures affect debt service and coverage?

2. Institutional Context and Analytical Concepts

2.1 Project finance and green finance architecture

Kazakhstan's Law on Project Financing and Securitization provides a legal basis for financing against project cash flows and permits the use of a special finance company in qualifying structures [7]. The Law on Public-Private Partnership connects PPP financing with the project finance framework [8]. Lenders still require enforceable security, predictable revenues, construction discipline, controlled accounts and credible default remedies.

Green finance has acquired a parallel institutional framework. Kazakhstan adopted a national Green Taxonomy in 2021, followed by mechanisms for coupon and interest subsidies and green bond guarantees in 2022 [9]. The Astana International Exchange provides securities issuance and listing infrastructure, while the AIFC Green Finance Centre supports sustainable-finance standards, external review and market development; the broader AIFC legal environment draws on English-law principles [10, 5]. These arrangements broaden sources of capital, although green classification does not substitute for credit analysis. A project can satisfy a taxonomy and still be unbankable because of weak cash-flow coverage, incomplete permits or unsuitable construction-risk allocation.

2.2 Syndicated and blended finance

A syndicated loan is provided by multiple lenders under coordinated documentation and administration. It may be arranged as a club deal, where lenders participate directly, or as an A/B loan. In the latter, a multilateral development bank retains an A loan for its own account and mobilizes B lenders under its umbrella. The structure can reduce duplication, provide consistent monitoring and give commercial lenders access to a transaction that may otherwise exceed their tenor or country limits.

Blended finance combines market-based capital with development finance, concessional funding, guarantees, grants or technical assistance. Public resources should address a specific market failure, such as unavailable long tenor, early-stage technology risk or a tariff affordability constraint, while preserving commercial discipline [6, 11]. Effective blending is limited, transparent and targeted to the risk preventing financial close.

For this paper, **bankability** means that a project can attract debt on terms consistent with its construction and operating cash flows while maintaining acceptable risk-adjusted returns for lenders. The principal indicators are the debt service coverage ratio, or DSCR, the loan life coverage ratio, or LLCR, construction contingency, reserve accounts, currency mismatch, covenant headroom and the reliability of contractual risk transfer.

2.3 Literature foundation and research gap

Project-finance scholarship explains how a ring-fenced special-purpose vehicle, contractual allocation and concentrated monitoring can support long-term infrastructure lending [12-13]. Research on syndicated lending indicates that development-bank participation can respond to political risk and coordination constraints, while project-finance structures often differ from general syndicated credits in maturity, security and monitoring intensity [13-14]. Renewable-energy



research further shows that financing structure can be material to investment deployment [15]. These studies establish general mechanisms, yet they provide limited integrated evidence on how currency matching, multilateral anchoring, concessional additionality and intercreditor control operate together in Kazakhstan. This article addresses that gap through a structured comparison of public Kazakhstan cases and a transparent illustrative model.

3. Research Design and Method

The study uses a structured comparative case method supported by an illustrative quantitative model. The evidence base consists of official publications from Kazakhstan's Bureau of National Statistics, the National Bank of Kazakhstan, AIFC, the Financial Settlement Center, OECD, peer-reviewed studies and public transaction disclosures by the EBRD, all available through 10 August 2026. Public sources avoid reliance on confidential term sheets. Each case is coded across financing scale, lender composition, currency, maturity rationale, concessional element, contractual interface and disclosed risk-mitigation features.

The cases were selected for theoretical variation. Zhanatas demonstrates blended local-currency renewable financing. KEGOC shows sovereign-linked grid investment combining ordinary and concessional resources. The Big Almaty Ring Road, or BAKAD, demonstrates a long-term PPP, an international sponsor group, an A/B loan and interest-rate hedging. Mirny shows large-scale commercial mobilization for wind generation with battery storage. The cases are not treated as a statistically representative sample. They are used to identify recurring structuring mechanisms across different infrastructure types.

The quantitative comparison holds project cost, equity share and operating cash flow constant. Total project cost is USD 700 million, equity is 30 percent, total debt is 70 percent, and annual cash flow available for debt service, or CFADS, is exactly USD 75 million in the illustrative operating year. Three stylised US-dollar capital structures are compared. Interest rates, tenors and repayment profiles are specified illustrative assumptions, not disclosed terms of any public case. Currency risk, taxes, fees, reserve funding and hedging costs are excluded and should be tested separately in an actual credit assessment.

Weighted average cost of debt is calculated as:

$$\bar{k} = \sum_{i=1}^n w_i k_i$$

where w_i is the share of debt instrument i in total debt and k_i is its specified fixed annual interest rate. Each tranche uses level annual debt service after completion:

$$ADS_i = D_i \times r_i / [1 - (1 + r_i)^{-n_i}]$$

Total annual debt service is the sum across tranches. The central coverage measure is:

$$DSCR_t = CFADS_t / (\text{Principal}_t + \text{Interest}_t)$$

The comparison excludes taxes, arrangement and commitment fees, political-risk insurance premiums, reserve funding costs and hedging costs. The outputs are therefore comparative model results, not an investment quotation.

4. Financing Instruments and Their Risk Functions

Table 1 compares the main instruments available for large green infrastructure projects. Their value depends on matching the instrument to the underlying constraint.

Table 1. Financing instruments for Kazakhstan green infrastructure

Instrument	Primary providers	Main bankability contribution	Risks retained by project or lenders	Appropriate use
Sponsor equity	Strategic sponsors, infrastructure funds, state holding companies	First-loss capacity, completion support, alignment of incentives	Full residual risk and delayed distributions	Development expenditure, contingencies and risks that cannot be transferred



Instrument	Primary providers	Main bankability contribution	Risks retained by project or lenders	Appropriate use
Commercial senior loan	Domestic and international banks	Payment discipline, local relationships, working-capital and hedging access	Tenor, refinancing, concentration and currency risk	Mature technology with predictable contracted revenue
Club or syndicated loan	Several banks under common documents	Larger ticket, exposure distribution and unified monitoring	Coordination and intercreditor risk	Projects too large for one bank but acceptable on market terms
MDB A/B loan	MDB as lender of record, commercial B lenders	Anchor due diligence, commercial mobilization and potentially longer maturity	B lenders retain funded credit risk; MDB retains A loan risk	Cross-border projects with strong fundamentals but limited commercial tenor or country appetite
Concessional tranche	Climate funds, donor facilities, public development institutions	Reduces debt service or absorbs a narrowly defined risk	Additionality, subsidy and eligibility risk	First-of-kind technology, affordability gap or measurable transition externality
Green bond or loan	Institutional investors, banks and capital markets	Diversifies funding and links proceeds to eligible assets	Refinancing, disclosure and external-review obligations	Construction takeout or portfolios of operating assets
Guarantee or standby facility	State, DFI, bank or insurer	Transfers specified political, payment, completion or liquidity risk	Counterparty, exclusions and claim-process risk	When one identifiable risk prevents otherwise viable financing

Source: Author's synthesis based on [6-7, 9, 11, 16].

Commercial debt supplies market pricing and screening. Its constraints were especially visible when the National Bank's base rate reached 18.0 percent in early 2026 and stood at 16.75 percent following the decision of 27 July 2026 [17]. Tenge financing can remove currency mismatch but create affordability pressure unless matched by suitable maturity, tariff escalation or development capital.

An A/B loan addresses a different problem. The MDB performs project appraisal and acts as lender of record, while B lenders provide mobilized commercial capital. Common documentation can reduce negotiation costs, yet it requires clear voting thresholds, sharing provisions, transfer rules and enforcement procedures. A syndicated structure is strongest when all lenders rely on the same base-case model, technical adviser, environmental and social action plan and information package.

Concessional capital should be placed where it produces the greatest marginal improvement. A small long-tenor tranche can improve minimum DSCR more efficiently than a broad interest subsidy if it is amortized during the project's weakest years. A guarantee can be superior when the obstacle is a specific payment or policy risk rather than general credit quality. Kazakhstan's renewable auction framework illustrates the importance of contingent instruments. Bid security is set at KZT 2,000 per kW and may be provided through a bank guarantee or standby letter of credit with a minimum 150-day validity [18]. Performance security under a power purchase agreement is KZT 10,000 per kW, with minimum validity periods that vary by technology [19]. These instruments support delivery discipline but consume bank limits and liquidity, which should be included in the total financing plan.

Figure 1 presents the common financing and control architecture used in the subsequent case comparison.

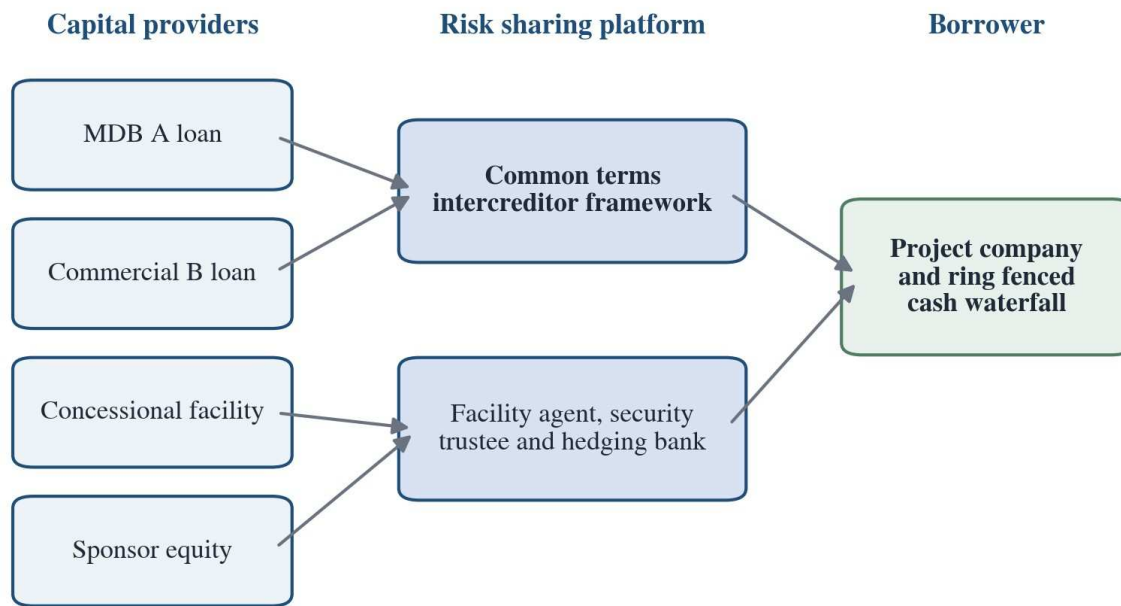


Figure 1. Illustrative syndicated and blended financing architecture

Source: Developed by the author.

5. Comparative Evidence from Kazakhstan

5.1 Zhanatas Wind Farm

The Zhanatas transaction demonstrates how several institutions can supply complementary risk capacity. The disclosed financing included an EBRD loan of KZT 10.847 billion, equivalent at disclosure to approximately USD 25.28 million, a Green Climate Fund tranche of up to USD 22.86 million, AIIB financing of up to USD 34.29 million, and an ICBC Almaty loan of KZT 5.841 billion, equivalent to approximately USD 13.61 million. Total project cost was disclosed at USD 140.1484 million [20].

The structure combined local-currency bank lending with multilateral and climate capital. Local-currency debt reduced the mismatch between tenge-denominated project revenue and debt service. Development institutions extended the available maturity and distributed exposure among lenders. The case also shows why the lowest nominal rate is an incomplete decision criterion. A foreign-currency loan may appear cheaper before hedging, yet depreciation can impair DSCR if the PPA does not provide an equivalent currency adjustment.

5.2 KEGOC grid reinforcement

Renewable generation is only bankable if transmission capacity and system balancing develop in parallel. In 2024, the EBRD announced financing of up to EUR 252 million in tenge equivalent for KEGOC, together with a EUR 15 million concessional loan from Canada. The programme covers approximately 600 km of 500 kV transmission infrastructure and is intended to support the integration of up to 12 GW of renewable capacity by 2030 [21].

The KEGOC case demonstrates two principles. First, local-currency lending can protect a regulated infrastructure borrower from currency mismatch. Second, concessional resources can be attached to a measurable system-level transition benefit. The concessional element should still be governed by procurement, reporting and impact conditions so that the price benefit is linked to delivery rather than absorbed as general balance-sheet support.

5.3 Big Almaty Ring Road

BAKAD provides a useful non-green financing comparator because it combines project finance, PPP risk allocation and international syndication. The financing package totalled USD 585 million for the project company to design, build, operate and maintain a 66 km road under a 20-year



concession. The sponsor consortium included Turkish and Korean construction and road-operation companies. The EBRD arranged a USD 350 million A/B loan and an interest-rate hedging programme involving an interest-rate cap and forward-starting swap [22-23].

The case illustrates that risk allocation extends beyond the loan agreement. The concession contract allocates availability, operating and public-sector obligations. The EPC and operation arrangements allocate delivery and performance risks. The hedging documents allocate interest-rate volatility, while intercreditor arrangements coordinate enforcement. Bankability therefore results from a linked contract system. A weak interface between two contracts can recreate a risk that each document appears to transfer separately.

5.4 Mirny wind and battery project

In April 2026, the EBRD announced an A/B financing package of up to USD 548 million for Aktas Energy within a project disclosed at approximately USD 1.260 billion. The EBRD A loan is up to USD 250 million and the B loan is up to USD 298 million. Publicly disclosed B-lender participants include China Construction Bank, Qatar National Bank, Société Générale and Standard Chartered, while other co-financing institutions include Proparco, DEG and the Development Bank of Kazakhstan [24-25]. The project combines 1 GW of wind capacity with a 300 MW and 600 MWh battery energy storage system [24-25].

Mirny is significant for three reasons. Its scale requires lender mobilisation beyond a conventional bilateral exposure. Battery storage adds system value but also introduces technology, degradation, dispatch and interface risks that must be reflected in technical due diligence and performance testing. Finally, the A/B format illustrates how an MDB can anchor a transaction while commercial and development lenders supply complementary debt. Public disclosure does not provide all commercial terms, so no inference should be made about pricing or internal lender decisions.

Table 2. Public Kazakhstan cases and structuring lessons

Case	Infrastructure and disclosed scale	Financing structure	Principal risk-allocation lesson
Zhanatas Wind Farm	Wind project; total cost USD 140.1484m	EBRD and ICBC Almaty KZT loans, GCF and AIIB financing	Combine local-currency revenue matching with development and commercial risk capacity
KEGOC grid programme	About 600 km of 500 kV transmission	Up to EUR 252m in KZT equivalent plus EUR 15m concessional funding	Use local currency for regulated cash flows and target concessionality to system transition benefits
BAKAD, non-green comparator	66 km toll-road PPP; USD 585m debt	EBRD-led club, USD 350m A/B loan and interest-rate hedge	Coordinate concession, construction, operation, financing and hedging contracts
Mirny	1 GW wind plus 300 MW/600 MWh battery; project cost about USD 1.260bn	Up to USD 250m EBRD A loan and USD 298m B loan, with additional disclosed co-financing	Mobilise commercial and development capital under an MDB anchor and apply integrated technology and interface due diligence

Sources: EBRD public disclosures [20-25]. Amounts are announced or disclosed commitments, not necessarily amounts disbursed.

6. Illustrative Capital Structure Comparison

The illustrative model compares financing architecture while holding project cost, leverage and CFADS constant. Each structure uses USD 210 million of equity and USD 490 million of US-dollar debt, with exact illustrative annual CFADS of USD 75 million. Fixed rates, level annual payments, no operating-period grace and no balloon repayment are assumed. The differences are the debt composition, specified interest rate and amortisation tenor.

**Table 3. Fully specified financing assumptions for a USD 700 million project**

Model item	Traditional commercial debt	Syndicated A/B structure	Blended structure
Sponsor equity	30%, USD 210m	30%, USD 210m	30%, USD 210m
Commercial debt	70%, USD 490m	35%, USD 245m	35%, USD 245m
MDB A loan	0	35%, USD 245m	20%, USD 140m
Concessional tranche	0	0	15%, USD 105m
Total debt	70%, USD 490m	70%, USD 490m	70%, USD 490m
Commercial rate and tenor	9.55%, 15 years	9.55%, 15 years	9.55%, 15 years
MDB rate and tenor	Not applicable	7.11%, 20 years	7.11%, 20 years
Concessional rate and tenor	Not applicable	Not applicable	4.03%, 25 years
Weighted average cost of debt	9.55%	8.33%	7.67%
Annual debt service by tranche	Commercial: USD 62.78m	Commercial: USD 31.39m; MDB: USD 23.32m	Commercial: USD 31.39m; MDB: USD 13.33m; concessional: USD 6.74m
Total annual debt service	USD 62.78m	USD 54.71m	USD 51.46m
Annual CFADS	USD 75.00m	USD 75.00m	USD 75.00m
DSCR under full structure	1.19	1.37	1.46
DSCR with price change only and common 15-year tenor	1.19	1.28	1.34

Source: Author's fully synthetic calculations using the equations in Section 3. Taxes, fees, reserve funding, hedging and currency movements are excluded. The assumptions do not reproduce any transaction in Section 5.

The traditional structure has the simplest lender group. A USD 490 million loan at 9.55 percent over 15 years produces annual debt service of USD 62.78 million and DSCR of 1.19. This gives limited capacity to absorb output shortfalls, delayed tariff indexation or higher operating cost.

The A/B structure keeps leverage unchanged while dividing debt equally between a 15-year commercial tranche and a 20-year MDB tranche. Annual debt service is USD 54.71 million and DSCR is 1.37. If only the weighted interest rate changed and the whole debt retained a common 15-year tenor, DSCR would be 1.28. The difference between 1.28 and 1.37 therefore represents the modelled tenor effect, while the difference between 1.19 and 1.28 represents the price effect. B lenders continue to fund and retain their credit exposure; the A/B structure primarily supports mobilisation, common due diligence and coordinated administration.

The blended structure adds a 25-year concessional tranche and reduces the MDB A loan to 20 percent of project cost. Annual debt service is USD 51.46 million and DSCR is 1.46. A price-only comparison at a common 15-year tenor would produce DSCR of 1.34. The remaining increase reflects longer amortisation. This structure produces the strongest buffer under the stated assumptions, while requiring the clearest additionality test. Concessional resources should be limited to the amount necessary to address a demonstrated financing gap and should carry measurable eligibility and impact obligations.

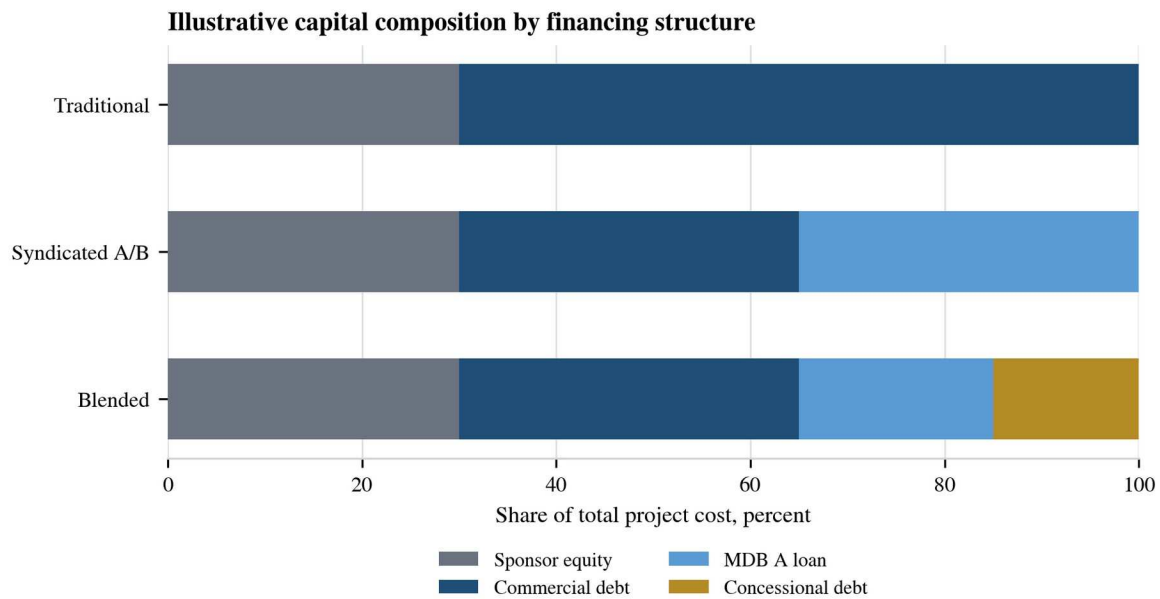


Figure 2. Illustrative capital composition by financing structure
Source: Developed by the author from the model assumptions.

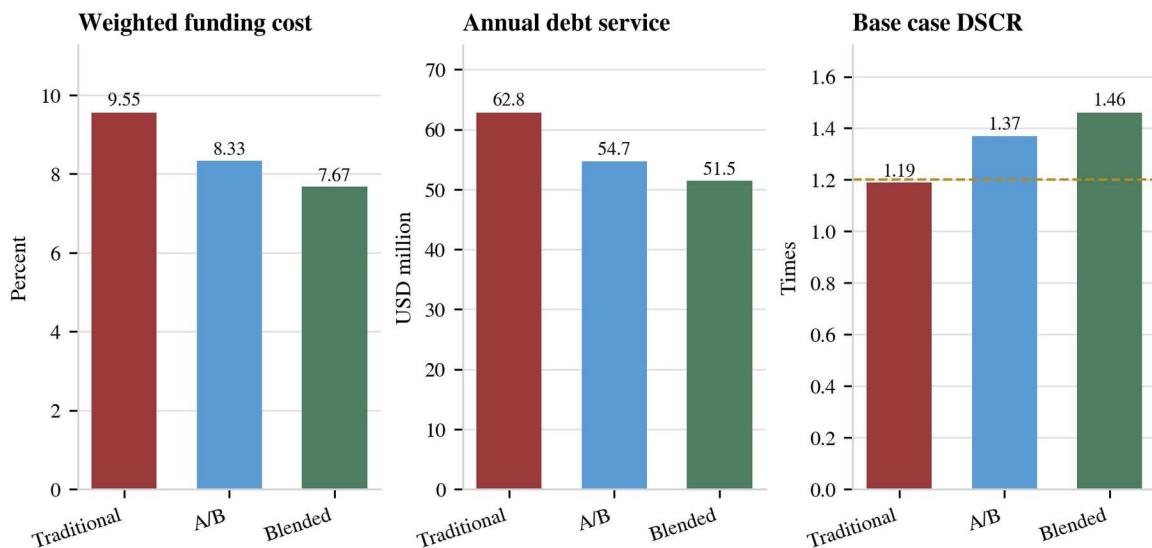


Figure 3. Illustrative comparison of funding cost, debt service and DSCR

Source: Author's calculation. Values exclude taxes, fees and hedging costs to preserve comparability.

The results should be interpreted as a financing-capacity comparison under stated assumptions. They do not establish that syndication or blended finance is inherently cheaper. Actual pricing may include floating-rate benchmarks, fees, political-risk insurance, reserve requirements and hedging, while local-currency alternatives require a separate exchange-rate and inflation analysis. Credit committees should determine whether concessionality is intended to reduce tariffs, improve coverage, finance additional environmental features or reduce sponsor equity before optimisation.

7. Risk Allocation and Policy Recommendations

The comparative evidence supports a control-based rule: each risk should be assigned to the party best able to control it, then to the party best able to absorb or price the residual exposure. Transfer to a party without operational control or adequate capital produces contractual form without economic substance.

Table 4. Recommended risk allocation and lender controls

Risk	Preferred primary allocation	Core contractual or financial controls	Syndication implication
Development and permits	Sponsor and project company	Conditions precedent, permit register, land-right verification and long-stop date	One shared legal due-diligence report and common waiver standard
Construction cost and delay	Fixed-price, date-certain EPC contractor, backed by sponsor completion support	Performance security, liquidated damages, contingency, independent engineer and equity cure	Uniform draw conditions and lender voting for budget changes
Technology and interface	Technology provider, integrator and project company	Performance tests, warranties, degradation assumptions and interface matrix	Common technical adviser and agreed acceptance tests
Revenue and offtake	Offtaker and project company according to controllability	Bankable PPA or availability payment, indexation and curtailment rules	Direct agreement and coordinated enforcement rights
Currency and interest rate	Project company, lenders and hedge providers	Currency matching, indexed revenue, swaps, caps and reserve accounts	Hedge counterparties integrated into intercreditor waterfall
Political and regulatory	Public authority, DFI, insurer and sponsors	Change-in-law clause, termination compensation and political-risk cover	Clear sharing of recoveries and standstill rights
Environmental and social	Project company and contractors under lender standards	ESIA, action plan, grievance mechanism, monitoring and reporting	One harmonized standard where possible to avoid conflicting obligations
Operations and maintenance	Qualified operator	Long-term O&M agreement, availability guarantees, spare-parts plan and insurance	Common reporting definitions and cash-sweep triggers

Source: Author's synthesis from the case evidence and Kazakhstan legal framework [7-8, 20-25].

Five recommendations follow. First, banks should use a common bankability data room before syndication. The financial model, EPC schedule, permits, environmental documents, insurance matrix and material contracts should share consistent definitions and dates. This reduces both due-diligence duplication and model risk.

Second, arrangers should design the intercreditor framework at the same time as the term sheet. Voting thresholds for drawstops, waivers, enforcement, hedge termination and amendments must reflect the different mandates of MDBs, commercial banks and concessional providers. Postponing these questions can convert an apparently complete financing plan into a delayed closing.

Third, local-currency solutions should be assessed on an all-in risk basis. Where project revenue is in tenge, local-currency debt can materially improve resilience. However, high local rates may require longer maturity, refinancing options, tariff indexation, interest-rate caps or a development tranche. Comparing only nominal coupons can misstate economic risk.

Fourth, concessionality should be allocated through a documented financing-gap test. The arranger should identify the minimum tranche or guarantee needed to meet a defined DSCR, affordability or technology threshold. The benefit and expected development result should be disclosed, and a sunset or refinancing mechanism should be considered once the relevant market failure declines.

Fifth, Kazakhstan can deepen refinancing channels for completed green assets. Construction banks are suited to monitoring completion risk, while institutional investors can hold long-duration operating assets. Standardised green reporting, reliable production data and bond takeout mechanisms would recycle bank balance sheets toward new projects. The documented expansion of sustainable-finance instruments provides an institutional base for this transition [4-5].

8. Limitations and Further Research

The public case disclosures do not provide complete interest rates, repayment schedules, covenant packages or hedging costs. The quantitative model is therefore illustrative and cannot establish that one structure is universally cheaper. It also holds CFADS and leverage constant, while actual financing structure can affect construction timing, operating requirements and tariff affordability.

The case sample is small and selected for variation rather than statistical representativeness. Future research could build a transaction-level database of Kazakhstan green bonds, loans and project financings, then test whether MDB participation is associated with longer maturity, greater local-currency use or lower refinancing risk. Further work should also model correlated tenge depreciation, inflation, base-rate shocks, construction delay and renewable output risk. Confidential bank data could strengthen calibration, subject to anonymization and governance requirements.

9. Conclusion

Kazakhstan's green infrastructure programme may exceed the tenor and exposure capacity of individual lenders. Syndicated and blended structures can provide a practical response when designed around project risks. Zhanatas illustrates the combination of local-currency, commercial and development finance. KEGOC illustrates targeted concessional support for system infrastructure. BAKAD, used as a non-green comparator, demonstrates that financing, concession, construction and hedging contracts should operate as one risk-allocation system. Mirny illustrates the mobilisation capacity of a large A/B loan for wind generation and battery storage.

Under the stated illustrative assumptions, the USD 700 million comparison produces DSCRs of 1.19, 1.37 and 1.46 for traditional, syndicated A/B and blended structures, respectively. These figures are not forecasts and reflect both pricing and tenor assumptions. They demonstrate how specified tranche composition can alter debt-service resilience without changing project cost or leverage. The central policy and credit conclusion is that bankability is supported by consistent due diligence, transparent intercreditor rules and concessionality limited to a verified financing gap. Under those conditions, development capital can mobilise commercial lending while preserving credit discipline.

Declarations

Data availability: All external information is publicly available from the sources listed in the references. All model assumptions and calculations are fully synthetic.

Conflict of interest and funding: The author is employed by China Construction Bank Corporation, Astana Branch. One public case discussed in this article identifies China Construction Bank as a lender. The analysis uses only public information and does not represent the bank's views. The author received no specific funding for this study; any required employer and publication approvals should be obtained before submission.

References:

- [1] Bureau of National Statistics of the Republic of Kazakhstan. "Energy Statistics." Renewable electricity indicators, including the 2025 share. <https://stat.gov.kz/en/industries/business-statistics/stat-energy/>
- [2] Financial Settlement Center of Renewable Energy LLP. "Renewable Energy in Kazakhstan." Installed capacity by technology, 2025. <https://rfc.kz/en/res-sector/res-in-kazakhstan/about-the-res/>
- [3] Bureau of National Statistics of the Republic of Kazakhstan. "On Investment in Fixed Assets in the Republic of Kazakhstan, 2025." <https://stat.gov.kz/en/industries/business-statistics/stat-invest/publications/506311/>
- [4] Astana International Financial Centre. "\$6bn Raised through the Astana International Financial Centre in 2025 and over 1,400 New Companies Were Registered." 26 December 2025.



<https://aifc.kz/news/6bn-raised-through-the-astana-international-financial-centre-in-2025-and-over-1400-new-companies-were-registered/>

[5] AIFC Green Finance Centre. “Sustainable Finance Analytics.” <https://gfc.aifc.kz/en/sustainable-finance-analytics>

[6] OECD. “Kazakhstan.” In *Accelerating Sustainable Infrastructure Investments*. 2026. https://www.oecd.org/en/publications/accelerating-sustainable-infrastructure-investments_031d486e-en/full-report/kazakhstan_fff79dd5.html

[7] Republic of Kazakhstan. Law No. 126-III “On Project Financing and Securitization.” Adilet Legal Information System. https://adilet.zan.kz/eng/docs/Z060000126_

[8] Republic of Kazakhstan. Law No. 379-V “On Public-Private Partnership.” Adilet Legal Information System. <https://adilet.zan.kz/eng/docs/Z1500000379>

[9] AIFC Green Finance Centre. *Green Finance Market of Kazakhstan*. 2023 edition. <https://aifc.kz/wp-content/uploads/2024/07/3.3-green-finance-market-of-kazakhstan.pdf>

[10] Astana International Financial Centre. “Legal Framework.” <https://aifc.kz/legal-framework/>

[11] OECD. “OECD DAC Blended Finance Principles for Unlocking Commercial Finance for the Sustainable Development Goals.” <https://www.oecd.org/dac/financing-sustainable-development/blended-finance-principles/>

[12] Esty, B. C. (2004). Why study large projects? An introduction to research on project finance. *European Financial Management*, 10(2), 213-224. <https://doi.org/10.1111/j.1354-7798.2004.00247.x>

[13] Kleimeier, S., & Megginson, W. L. (2000). Are project finance loans different from other syndicated credits? *Journal of Applied Corporate Finance*, 13(1), 75-87. <https://doi.org/10.1111/j.1745-6622.2000.tb00043.x>

[14] Hainz, C., & Kleimeier, S. (2012). Political risk, project finance, and the participation of development banks in syndicated lending. *Journal of Financial Intermediation*, 21(2), 287-314. <https://doi.org/10.1016/j.jfi.2011.10.002>

[15] Steffen, B. (2018). The importance of project finance for renewable energy projects. *Energy Economics*, 69, 280-294. <https://doi.org/10.1016/j.eneco.2017.11.006>

[16] European Bank for Reconstruction and Development. “Supporting the Renewable Energy Transition in Kazakhstan.” 31 March 2026. <https://www.ebrd.com/home/what-we-do/impact-management/impact-case-studies/supporting-the-renewable-energy-transition-in-kazakhstan.html>

[17] National Bank of Kazakhstan. “Schedule of Base Rate Decisions for 2026.” <https://nationalbank.kz/en/news/grafik-prinyatiya-resheniy-po-bazovoy-stavke/rubrics/2365>

[18] Financial Settlement Center of Renewable Energy LLP. “Financial Support of the Application.” <https://rfc.kz/en/res-sector/auctions/financial-support-of-the-application/>

[19] Financial Settlement Center of Renewable Energy LLP. “Financial Support of the Agreement.” <https://rfc.kz/en/res-sector/auctions/financial-support-of-the-agreement/>

[20] European Bank for Reconstruction and Development. “Zhanatas Wind Farm.” Project Summary Document, Project 50569. <https://www.ebrd.com/home/work-with-us/projects/psd/50569.html>

[21] European Bank for Reconstruction and Development. “EBRD Helps to Consolidate Kazakhstan's Power System.” 25 November 2024. <https://www.ebrd.com/home/news-and-events/news/2024/ebd-helps-to-consolidate-kazakhstan-s-power-system.html>

[22] European Bank for Reconstruction and Development. “BAKAD: European Transport Deal of the Year.” 2020. <https://www.ebrd.com/home/work-with-us/treasury-capital-markets/bakad.html>

[23] European Bank for Reconstruction and Development. “BAKAD Road Concession.” Project Summary Document, Project 47085. <https://www.ebrd.com/home/work-with-us/projects/psd/47085.html>

[24] European Bank for Reconstruction and Development. “EBRD Organises Financing of US\$ 548 Million in Kazakhstan's Wind Farm with Storage Capability.” 24 April 2026. <https://www.ebrd.com/home/news-and-events/news/2026/ebd-invests-us--250-million-in-kazakhstan-s-wind-farm-with-stor.html>

[25] European Bank for Reconstruction and Development. “Mirny WPP and BESS.” Project Summary Document, Project 55877. <https://www.ebrd.com/home/work-with-us/projects/psd/55877.html>



ӘЛЕУМЕТТИК-ЭКОНОМИКАЛЫҚ ЖӘНЕ ҚҰҚЫҚ ҒЫЛЫМДАРЫ
СОЦИАЛЬНО-ЭКОНОМИЧЕСКИЕ И ЮРИДИЧЕСКИЕ НАУКИ
SOCIO-ECONOMIC AND LEGAL SCIENCES

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. No. 4 (32), Vol. 2, pp. 23–33
DOI: <https://doi.org/10.5281/zenodo.23770264>
Journal website: <https://cajournal.kz>

UDC 336

GREENFIELD DEVELOPMENT OF CANAL-BASED SMALL
HYDROPOWER IN UZBEKISTAN: SITE SCREENING, PPP RISK
ALLOCATION AND CLIMATE RESILIENCE THROUGH THE WATER-
ENERGY-FOOD NEXUS

Wang Liang

Al-Farabi Kazakh National University, Kazakhstan

Wang Liang — DBA Student in Business Management.

© Wang Liang, 2026

Abstract: Uzbekistan is seeking to expand hydropower while modernizing an irrigation system that supports rural livelihoods and dominates national water use. Existing canals create a distinctive greenfield opportunity: small plants can add distributed renewable generation with limited new inundation, but their output remains subordinate to irrigation schedules, water rights and climate uncertainty. This paper develops an integrated framework for converting a large site list into an investable portfolio. It combines policy and project-document review with a two-stage multi-criteria model, hydrological stress testing and normative public-private partnership risk allocation. Stage 1 applies fatal-flaw gates for water access, irrigation non-impairment, grid and land feasibility, and environmental-social eligibility. Stage 2 ranks eligible sites by water reliability, energy yield, irrigation compatibility, cost, connection, impact and sponsor readiness. An illustration shows why a high-head location can rank below a more water-compatible site. The financing design links concessional wholesale funding and local banks to project companies, standardized contracts and verified operating data. Climate parameters remain scenarios: the severe case tests a 30–40 percent water-availability decline with a 25 percent irrigation-demand increase. The paper reconciles the 8–12 GW resource range and 5.5–6 GW target range by distinguishing assessment boundaries, technical potential, program staging and date. The central finding is that canal hydropower should be developed as a governed portfolio of water-compatible infrastructure. Transparent screening, irrigation-first operating rules and efficient risk assignment can reduce development attrition and improve resilience.

Keywords: canal-based small hydropower; Uzbekistan; public-private partnership; water-energy-food nexus; multi-criteria screening; climate resilience

JEL codes: Q25; Q42; Q48; L94; G32

1. Introduction

Uzbekistan's electricity transition must meet rising demand, reduce thermal dependence and improve supply reliability. The World Bank expects electricity demand to exceed 120 billion kWh by 2030 and reports unmet demand of about 10 percent [1]. Variable solar and wind increase the value of controllable, distributed generation. Small hydropower can aid daily balancing and local supply, while canal schemes use existing irrigation infrastructure and often require no new reservoir.

The opportunity is substantial but easily overstated. A canal is first a water-service asset; cropping patterns, allocations, closures, upstream conditions and water-authority decisions determine its flow. An attractive hydraulic drop can therefore produce weak debt coverage outside



the irrigation season. Construction can disrupt delivery, and an artificial channel can still create livelihood, safety, biodiversity, cultural-resource and land impacts. Reservoir avoidance does not remove the need for due diligence.

Current policy creates a practical opening for private participation. In 2023 the Government of Uzbekistan reported an initial national hydropower potential of 8 GW, around 2.2 GW then utilized, and an objective of reaching 6 GW by 2030. It also described 20-year land leases for small and micro hydropower investors, guaranteed state purchase of their electricity and an intention to attract private capital [2]. In 2025 the World Bank approved a USD 150 million concessional credit to support small hydropower through local developers and financial institutions. The associated project targets up to 150 MW across existing irrigation canals and anticipates USD 38 million of commercial co-financing [3,4]. These measures shift the central managerial question from whether canal hydropower is technically possible to how a heterogeneous site pipeline should be screened, contracted and financed.

This paper addresses that question through the water-energy-food (WEF) nexus. The nexus perspective recognizes that one cubic meter of canal water can simultaneously support food production, require pumping energy, embody ecological obligations and generate electricity. The appropriate objective is therefore not maximum annual hydropower output. It is maximum risk-adjusted public value subject to reliable irrigation and environmental constraints. Three research questions guide the analysis:

1. How can public agencies, developers and banks screen canal-based small hydropower sites consistently before expensive feasibility work?
2. Which allocation of construction, hydrology, irrigation, offtake, financial and environmental risks is most likely to produce bankable public-private partnership structures?
3. How should climate and irrigation stress be incorporated without treating uncertain scenario values as deterministic forecasts?

The contribution is a policy-to-practice framework that integrates project development, multi-criteria decision analysis (MCDA), PPP design and climate resilience. It also resolves two statistical ambiguities that can distort investment decisions. First, the 8 GW government estimate and the approximately 12 GW figure reported in the World Bank's 2025 project appraisal materials are different resource concepts and vintages, not values that should be averaged [2,4]. Second, the 5.5 GW hydro component cited in a 2025 renewable-energy program document and the 6 GW national hydropower objective cited by the government represent different program scopes or staging [2,4]. Transparent treatment of such differences is essential when targets are converted into project pipelines.

2. Literature and Policy Review

2.1 Small hydropower, MCDA and the nexus

Hydropower research increasingly distinguishes installed capacity from sustainable value. Moran et al. argue that planning must internalize ecological, social and governance effects [5]. Small scale can reduce some impacts, yet a dispersed program creates cumulative land, labor, safety and water-allocation risks. The World Bank identifies livelihood, labor, community-safety, stakeholder-exclusion and financial-intermediary capacity risks in Uzbekistan's program [6]. Portfolio controls are therefore necessary.

The WEF nexus supplies the conceptual basis. Hoff links it to resource efficiency and visible cross-sector externalities [7], while Rasul warns that sector optimization can displace costs [8]. Central Asian research calls for explicit indicators, participation and interdisciplinary decisions [9]. UNECE also identifies benefits from data exchange and coordinated operation in the Syr Darya basin. Uzbekistan did not formally associate itself with that assessment, so it is regional evidence rather than an official national position [10].

MCDA translates these ideas into screening. AHP and weighted-sum models compare technical, economic, social and environmental criteria while recording judgment [11,12]. Hydropower rankings change when environmental and stakeholder criteria supplement energy and cost [13], which matters where unreliable flow or irrigation conflict offsets theoretical generation.

MCDA remains dependent on data, weights and thresholds; final approval still requires measurement, design, consultation and financial appraisal.

2.2 Uzbekistan's policy and project context

The Green Economy Strategy 2019–2030 and Uzbekistan–2030 strategy frame cleaner energy and infrastructure objectives [14,15]. Decree DP-6024 emphasizes water accounting, modernization and efficient management [16]. The 2025 Water Code governs water administration and permits transfer of water-management facilities under PPP law [17]. Hydropower rights therefore remain inseparable from the water-management mandate.

Implementation experience is expanding. ADB's Sustainable Hydropower Project includes three run-of-river plants and mini/micro capacity development [18]. The 1.48 MW Ugam-1 plant is reported to produce 11.8 million kWh annually using domestic materials and equipment [19]. Localization creates durable value only where suppliers meet efficiency, safety, automation and spare-parts requirements.

The 2025 World Bank program directly addresses canal development. Appraisal materials describe about 3,000 prospective locations, screening of up to 400 sites and analytical support for feasibility studies [4]. Local banks lend through a credit facility, complemented by pipeline and institutional assistance. Reservoir avoidance does not eliminate environmental-social screening or the World Bank policy on international waterways [4,6]. Dispersed private investment on public canals therefore needs common eligibility rules and data.

2.3 Reconciling potential and target statistics

Four headline numbers appear in authoritative documents: approximately 8 GW and 12 GW for potential, and 5.5 GW and 6 GW for policy ambition. They should be reported as ranges with definitions.

The government's 2023 communication calls 8 GW an "initial estimate" of hydropower potential [2]. The World Bank's 2025 project document cites approximately 12 GW, drawing on an older 2013 small-hydropower assessment [4,20]. A higher technical-potential figure may include more locations, different plant-size thresholds, gross hydraulic resources, updated mapping or sites that are not economically, environmentally or institutionally developable. The 8–12 GW range therefore signals uncertainty in assessment boundaries and data vintage. It is not a single bankable resource estimate. Site-level dependable energy, construction cost and water compatibility determine the economically developable subset.

The 6 GW figure is an official national objective for total hydropower capacity by 2030 [2,19]. The 5.5 GW figure appears in the World Bank appraisal document as the hydropower component of a first-stage 25 GW renewable-energy plan [4]. A 0.5 GW difference can result from program staging, cut-off date, classification or inclusion of projects outside the initial renewable package. For project governance, the prudent approach is to retain both cited values, disclose their source and use a 5.5–6 GW target band until a single updated official planning baseline is published. Neither target implies that thousands of canal sites should pass directly into construction.

3. Methodology and Analytical Framework

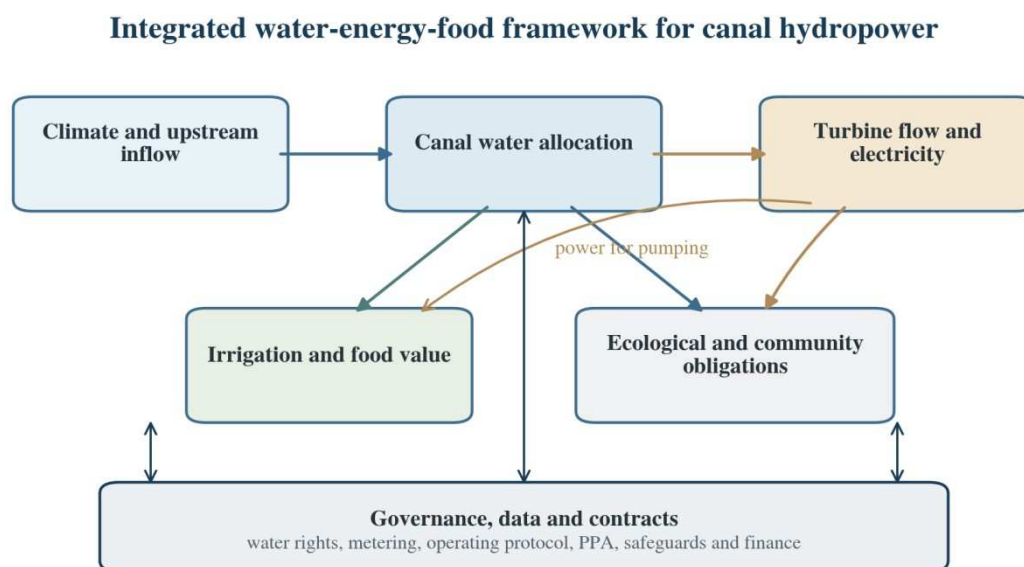
This is an applied conceptual study rather than an ex post evaluation. It uses four methods. First, a structured review covers Uzbekistan's energy and water policies, official statements, IFI project documents, regional nexus studies and international guidance on hydropower, PPPs and environmental and social management. Sources were selected for institutional authority, direct relevance and traceable publication. Second, the paper maps a conventional greenfield development sequence from opportunity identification through screening, feasibility, procurement, financing, construction and operation. Third, it develops a two-stage MCDA model suitable for a large canal-site inventory. Fourth, it applies stress scenarios to show how water constraints should enter energy and debt analysis.

Stage 1 is a fatal-flaw screen. A candidate proceeds only if it has a legally credible water-use pathway, does not materially impair irrigation service, can obtain land and canal access, has a feasible grid connection, and falls within the environmental and social eligibility of the financing

Stage 2 ranks eligible sites. For site i , the composite score is:

Annual energy is estimated from time-step canal flow rather than an annual average:

subject to turbine minimum and maximum flow, canal operating constraints, required bypass, irrigation priority and planned closures. Financial screening then tests debt-service coverage under P50 and conservative-flow cases. No empirical site data are claimed in this paper; the numerical comparison in Section 4.3 is an illustrative demonstration of the method.



Source: Author's construction based on Hoff [7], UNECE [10], the Uzbekistan Water Code [17] and the World Bank water-security assessment [21].

4.1 Stage 1: fatal-flaw gates

The second gate is irrigation non-impairment: authorized volume, timing and hydraulic level must be preserved through bypass and safe shutdown arrangements. The third gate requires a preliminary grid opinion covering voltage, capacity, protection, curtailment and reinforcement. The fourth covers land, rights of way and canal-compatible construction access. The fifth applies environmental-social exclusions, cumulative-risk review, engagement and grievance arrangements [6,22].

Two-stage screening decision tree for canal hydropower sites

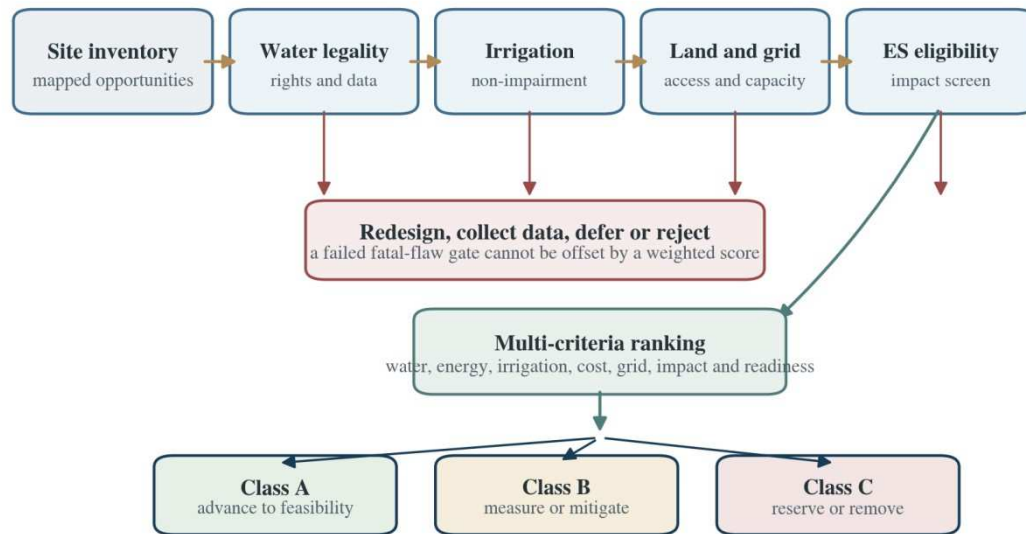


Figure 2. Two-stage site-screening decision tree for canal hydropower sites

Source: Author's construction based on the World Bank project appraisal and environmental and social documents [4,6], MCDA literature [11–13] and the Water Code [17].

4.2 Stage 2: weighted ranking

Table 1 gives water reliability the largest weight because flow determines generation and service conflict. Irrigation compatibility is separate because reliable water is not necessarily turbine-available water. Other criteria capture net energy, economics, connection, impact and execution. The weights require validation by energy and water agencies, canal and grid operators, banks, developers, farmers and environmental specialists.

Table 1. Proposed MCDA criteria for eligible canal-based small hydropower sites

Criterion	Weight	Principal measures	Anchored interpretation
Water reliability	0.22	Monthly measured flow, dependable flow, closure duration, data completeness	High score requires stable dry-season availability and auditable records
Usable head and energy yield	0.16	Net head, hydraulic losses, turbine efficiency curve, P50/P90 energy	High score reflects dependable energy, not only installed MW
Irrigation compatibility	0.16	Delivery timing, bypass capacity, backwater effect, outage coordination	Any material service reduction fails Stage 1; higher scores add operational flexibility
Capital cost and economics	0.13	Unit CAPEX, lifecycle O&M, LCOE, contingency, DSCR	High score requires robust economics under conservative flow
Grid and access	0.11	Connection distance/capacity, road access, construction logistics	High score indicates limited reinforcement and low interface risk
Environmental and social exposure	0.12	Land, livelihood, safety, biodiversity, cultural heritage, cumulative effects	High score indicates manageable impacts with verified mitigation
Sponsor and execution readiness	0.10	Equity capacity, governance, EPC/O&M competence, permits, data room	High score requires evidence, not declarations
Total	1.00		



Source note: Author's proposed weights, informed by Saaty [11], Belton and Stewart [12], Vassoney et al. [13], the World Bank project framework [4,6] and IFC Performance Standards [22]. Weights require local stakeholder validation before operational use.

4.3 Illustrative portfolio comparison

Table 2 applies the framework to three fictional sites. Site A has moderate head but reliable flow, a short connection and a manageable construction interface. Site B has the highest head and expected annual energy, yet its seasonal flow, longer outage and irrigation interface reduce its total score. Site C has lower technical potential but strong land and stakeholder conditions. The example demonstrates why a single indicator such as MW, annual kWh or unit construction cost should not determine priority.

Table 2. Illustrative scoring of three fictional canal sites

Criterion	Weight	Site A: stable main canal	Site B: high-head seasonal drop	Site C: low-head district canal
Water reliability	0.22	88	58	74
Usable head and energy yield	0.16	72	94	55
Irrigation compatibility	0.16	90	52	82
Capital cost and economics	0.13	78	82	62
Grid and access	0.11	85	60	76
Environmental and social exposure	0.12	80	68	90
Sponsor and execution readiness	0.10	75	80	70
Weighted score	1.00	81.3	70.5	72.7
Indicative class		A: feasibility	B: measure/mitigate	B: optimize design

Source note: Author's illustrative calculation using Table 1. The site descriptions and scores are fictional and do not represent an appraisal of any location in Uzbekistan.

Before procurement, a Class A site needs independent review of measured flow and energy, credible sediment and maintenance assumptions, defined connection and curtailment terms, and a climate-irrigation downside test. Publishing criteria, data vintage and rejection reasons discourages speculative site reservation and directs expenditure toward credible projects.

5. PPP and Financing Design

5.1 Contract architecture

A canal plant intersects public water infrastructure, a regulated buyer, private construction and bank finance. The project company needs coordinated land/facility, water, grid, PPA, EPC, O&M and financing instruments. A lender direct agreement should establish notice, cure and step-in rights, while the canal authority signs a binding operating protocol.

The protocol should prioritize irrigation, safety and lawful ecological obligations, then govern generation from eligible flow. It specifies schedules, closure notice, emergency control, metering, data and damage responsibility. It also separates natural variability from discretionary curtailment. Compensation applies only where authorized; open-ended public indemnities create poor incentives and fiscal exposure.

Common feasibility terms, metering, PPA clauses, safeguard templates and minimum specifications reduce transaction cost. Site schedules should still address flow, head, land and community obligations because canals differ.



5.2 Risk allocation

Risk belongs with the party best able to control, mitigate or absorb it [23]. Full private transfer of uncontrollable hydrology can raise tariffs or destabilize sponsors, while public assumption of construction performance weakens discipline. Table 3 balances these concerns.

Table 3. Recommended risk allocation for a canal-based small hydropower PPP

Risk	Primary allocation	Principal mitigation and contractual treatment
Site data and baseline flow	Public/shared before award	Public data room; independent verification; bidder due diligence; disclosed uncertainty band
Hydrological variability	Private within agreed band; shared for exceptional events	Conservative P90 case, reserve account, relief-event thresholds, insurance where available
Irrigation scheduling	Public canal authority for discretionary changes; private for compliance	Binding water protocol, notice periods, irrigation-first rule, auditable metering
Design, construction and performance	Private	Fixed-date EPC terms, performance tests, contingency, liquidated damages and security
Land, facility right and key permits	Public enabling role; private compliance	Conditions precedent, clear site boundary, coordinated approvals
Grid connection	Shared by defined scope	Connection study, cost boundary, commissioning interface and curtailment provisions
Electricity offtake and payment	Public offtaker, subject to contract	Bankable PPA, payment security permitted by law, lender direct agreement
Foreign exchange and interest rate	Shared	Local-currency revenue and debt where feasible, indexed tariff only if authorized, hedging assessment
O&M, availability and equipment	Private	Lifecycle plan, spare parts, remote monitoring, availability incentives
Environmental, social, labor and safety	Private implementation with public oversight	ESMS, management plans, consultation, grievance mechanism, incident reporting
Change in law, force majeure and political event	Public or shared according to control	Defined relief, restoration plan, termination and compensation rules

Source note: Author's construction based on the World Bank PPP Reference Guide [23], the Uzbekistan Water Code [17], the World Bank project and environmental-social documents [4,6], and IFC Performance Standards [22].

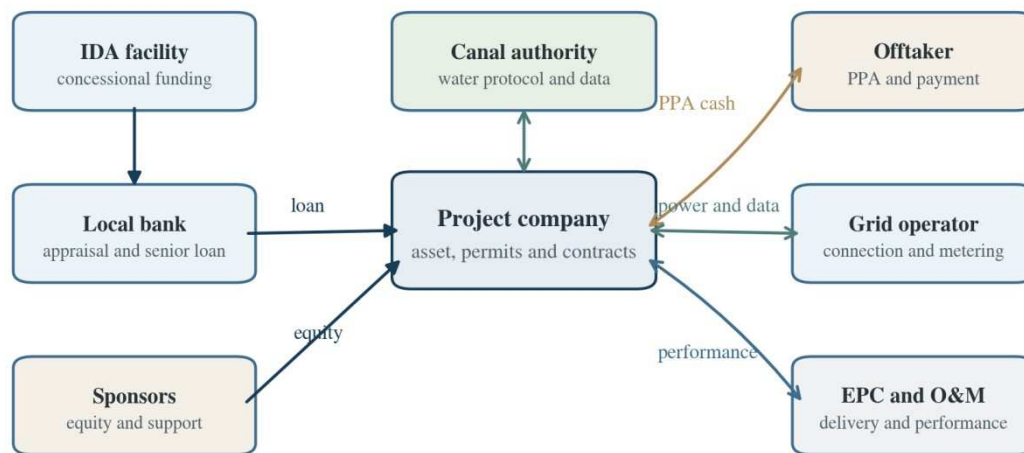
5.3 Financing structure and bankability

Local developers need long-tenor capital, while local banks face maturity and appraisal constraints [1,4]. Concessional wholesale funds can reach eligible projects through participating banks that retain credit responsibility and co-finance. Sponsors supply equity and borrow against PPA revenue after technical, financial and environmental-social appraisal. Concessionality must not replace sound underwriting.

The base case should use time-step water and planned closures; the downside case should use dependable flow. Coverage covenants, debt-service and maintenance reserves, and distribution lock-up absorb volatility. Drawdown requires executed water/grid agreements, approved

safeguards, independent engineering, equity and a fixed completion budget. Tenor and amortization should match seasonal cash flow.

PPP finance, cash and information flows for canal hydropower



Safeguard, metering and operating information supports every financial and physical interface.

Figure 3. PPP finance, cash and information flows for canal hydropower

Source: Author's construction based on the financing and institutional arrangements described for the Uzbekistan Small Hydropower Development Project [1,4,6] and the PPP Reference Guide [23].

6. Climate Resilience and Irrigation Constraints

Uzbekistan's water exposure makes backward-looking hydrology insufficient. The World Bank reports a climate scenario in which water availability could decline by 30–40 percent and irrigation demand could increase by 25 percent, reflecting glacier change, evapotranspiration, rainfall variability and drought [24]. These values are scenario estimates, not a forecast that every canal will experience the same change or that the two effects will occur uniformly. Their purpose in this paper is to define a severe stress envelope for investment testing.

Three cases are recommended. The reference case uses the best available measured record, current irrigation rules and planned efficiency improvements. A moderate-stress case can test a 15 percent reduction in available canal flow, a 10 percent rise in irrigation requirement and longer maintenance or drought-related interruption. The severe case tests the cited 30–40 percent water decline together with a 25 percent irrigation-demand increase. Site-specific studies should refine the timing of these shocks because monthly distribution matters more to turbine production than an annual percentage alone. Each case should report annual energy, monthly minimum generation, irrigation delivery, curtailment, operating margin and debt-service coverage.

Climate resilience begins with selection [26]. A plant located on a canal with diverse upstream sources, reliable records and flexible bypass arrangements may be more resilient than a technically larger site dependent on a narrow seasonal window. Design measures include modular turbine units with broad efficiency ranges, adequate bypass, variable-speed or multiple-unit configurations where economic, sediment and debris management, flood-safe electrical equipment, remote monitoring and physical provision for changing water levels. Contract measures include periodic review of operating curves, shared data, drought and flood protocols, and clear relief-event thresholds. Financial measures include conservative leverage, reserves, covenant headroom and staged portfolio deployment.

Irrigation modernization can create both synergies and rebound effects. Rehabilitation may reduce conveyance losses and pumping energy, improving reliability and releasing operational head at selected structures. Yet efficiency gains do not automatically leave more water for power if saved water is allocated to expanded irrigation. The World Bank's water-security assessment notes



outdated infrastructure, weak efficiency and the need for cost recovery and better operations and maintenance [21]. A 2025 irrigation project emphasizes modernization across a sector that accounts for around 90 percent of national water use [24]. Accordingly, any claimed hydropower benefit from water-saving investment should be based on an enforceable allocation and measured delivery, not on an assumed “saved” volume.

The transboundary dimension also matters. Many Uzbek water resources are connected to the Amu Darya and Syr Darya systems. Canal plants ordinarily do not determine basin releases, but their output depends on them. The Water Code subjects transboundary water use to international treaties [17], and World Bank regional programs emphasize data sharing, coordinated water-energy operation and transparent institutions [25]. Project documents should therefore avoid allocating regional inflow risk through unrealistic private guarantees. At portfolio level, geographic diversification across basins and canal systems can reduce correlated hydrology risk, while government and regional institutions retain responsibility for interstate coordination.

7. Discussion and Managerial Implications

The analysis supports five managerial conclusions. First, the national site inventory should be managed as a funnel. Thousands of mapped hydraulic opportunities may produce hundreds of screenable sites, dozens of full feasibility studies and a smaller number of financed plants. Attrition is evidence of screening quality, not program failure. Development budgets should fund early data and remove weak sites before procurement.

Second, water governance is part of project finance. A bankable PPA cannot compensate for an ambiguous canal operating right. Developers, water authorities and lenders need a common record of flow, closures and irrigation obligations. Digital metering should capture upstream flow, turbine flow, bypass flow, head, energy, outages and instructed curtailment. Data should be accessible to the project company, canal operator, offtaker and lenders under agreed rules. This creates a factual basis for operations, dispute resolution and future resource planning.

Third, procurement should reward lifecycle value. Lowest upfront tariff or construction cost can encourage optimistic flow assumptions and weak equipment. Bid evaluation should test energy under the same public hydrological dataset, require a lifecycle O&M plan, and adjust for connection scope and environmental-social obligations. Where sites are very small, bundling several geographically coherent plants may spread transaction and O&M costs, but packages should avoid excessive concentration and correlated canal risk.

Fourth, localization should follow a capability ladder. Civil works, steel structures, control panels and routine maintenance may offer early domestic opportunities. Turbine runners, governors, generators and protection systems require performance standards, factory testing, warranties and quality assurance. The Ugam-1 experience shows policy interest in domestic content [19], while the World Bank credit-line model can widen local developer and bank participation [1]. Training, certification and supplier qualification are more durable than a nominal local-content percentage.

Fifth, a portfolio control unit should combine energy, water, finance and safeguard expertise. Its dashboard should report the number of sites at each gate, data completeness, expected P90 energy, irrigation-service effect, connection status, environmental-social risk, financing status and reasons for delay. Independent technical review is justified before financial close and after commissioning. Post-construction performance should feed back into weights, cost benchmarks and standard contracts. This learning loop converts scattered projects into an institutional capability.

The approach also clarifies the role of PPPs. Private participation is valuable where it introduces disciplined design, construction, financing and operation. Public agencies remain responsible for transparent water allocation, credible site rights, fair procurement, sector regulation and oversight. The goal is selective risk transfer, not maximum risk transfer. If a public party passes unpredictable irrigation scheduling or unobservable baseline-flow risk to a small developer, bidders will price uncertainty, rely on renegotiation or decline to invest. Conversely, public absorption of ordinary construction and availability risk would reduce accountability.

At the broader WEF level, canal hydropower can create a positive loop when its electricity supports local reliability and efficient pumping while preserving water delivery. It can create a



negative loop if generation incentives distort irrigation timing or if new revenues bypass water-system maintenance and affected users. Transparent benefit reporting and stakeholder engagement are therefore economic controls as well as social obligations. Farmer representatives should participate when operating schedules are set, and the grievance mechanism should remain accessible throughout construction and operation [6].

8. Conclusion

Canal-based small hydropower offers Uzbekistan a practical route to distributed renewable electricity using existing water infrastructure. Its greenfield development logic differs from that of a stand-alone power plant because the commercial fuel, flowing water, is governed primarily for irrigation and is exposed to climate and transboundary conditions. The project-development system must place water legality, irrigation compatibility and environmental-social eligibility ahead of capacity maximization.

This paper proposes a two-stage process: fatal-flaw gates followed by transparent weighted ranking. It links the screening result to standardized PPP contracts, local-bank financing and climate stress tests. The 8–12 GW potential range should be understood as a difference in assessment scope, methodology and vintage; the 5.5–6 GW target range reflects program staging and planning definitions. Both ranges require disclosure rather than artificial harmonization. Likewise, a possible 30–40 percent decline in water availability and 25 percent increase in irrigation demand should be used as a severe scenario, not asserted as a uniform outcome.

For policymakers, the priority is a credible portfolio data room and cross-ministerial screening protocol. For developers, the priority is dependable flow, lifecycle performance and early stakeholder engagement. For lenders, the priority is conservative energy estimation, clear interface agreements and enforceable monitoring. If these disciplines are combined, small hydropower can contribute to energy security while reinforcing, rather than competing with, Uzbekistan's water and food-security objectives.

9. Limitations and Future Research

The study is conceptual and relies on public documents; it does not use confidential canal records, tariff data or actual project cash flows. The weights and fictional site scores require validation through stakeholder workshops and field testing. National scenario values may not represent individual basins or months, and legal or market arrangements may change. Future research should apply the model to measured hourly or daily flows, compare AHP results with outranking or stochastic optimization, quantify cumulative effects across canal cascades, and test financing structures under local-currency, tariff and hydrology sensitivities. Longitudinal study of the first financed portfolio would allow the screening thresholds to be recalibrated against actual construction cost, generation, irrigation performance and debt service.

References:

- [1] World Bank. (2025). *Uzbekistan to Enhance Energy Production through Hydropower with World Bank Support*. <https://www.worldbank.org/en/news/press-release/2025/06/25/uzbekistan-to-enhance-energy-production-through-hydropower-with-world-bank-support>
- [2] Government of the Republic of Uzbekistan. (2023). *The Progress of Hydropower Projects and the State of Preparation for the Autumn/Winter Season Are Considered*. <https://gov.uz/en/news/view/3347>
- [3] World Bank. (2025). *Uzbekistan Small Hydropower Development Project, Project Page (P509250)*. <https://projects.worldbank.org/en/projects-operations/project-detail/P509250>
- [4] World Bank. (2025). *Uzbekistan Small Hydropower Development Project (P509250): Project Appraisal Document*. <https://documents1.worldbank.org/curated/en/099052125205569067/pdf/P509250-157ae192-23c9-4a6c-ab9b-f2c8af10cce0.pdf>
- [5] Moran, E. F., Lopez, M. C., Moore, N., Müller, N., & Hyndman, D. W. (2018). Sustainable hydropower in the 21st century. *Proceedings of the National Academy of Sciences*, 115(47), 11891–11898. <https://doi.org/10.1073/pnas.1809426115>
- [6] World Bank. (2025). *Uzbekistan Small Hydropower Development Project: Appraisal Environmental and Social Review Summary*.



<https://documents1.worldbank.org/curated/en/099052025144541135/pdf/P509250-1ed37902-8be7-4e09-b731-362c96f78d9e.pdf>

[7] Hoff, H. (2011). *Understanding the Nexus: Background Paper for the Bonn 2011 Conference*. Stockholm Environment Institute. <https://www.sei.org/publications/understanding-the-nexus/>

[8] Rasul, G. (2016). Managing the food, water, and energy nexus for achieving the Sustainable Development Goals in South Asia. *Environmental Development*, 18, 14–25. <https://doi.org/10.1016/j.envdev.2015.12.001>

[9] Hamidov, A., Daedlow, K., Webber, H., Hussein, H., Abdurahmanov, I., Dolidudko, A., Seerat, A., Solieva, U., Woldeyohanes, T., & Helming, K. (2022). Operationalizing water-energy-food nexus research for sustainable development in social-ecological systems: An interdisciplinary learning case in Central Asia. *Ecology and Society*, 27(1), 12. <https://doi.org/10.5751/ES-12891-270112>

[10] United Nations Economic Commission for Europe. (2017). *Assessment of the Water-Food-Energy-Ecosystems Nexus in the Syr Darya River Basin*. https://unece.org/DAM/env/water/publications/GUIDELINES/2017/nexus_in_the_Syr_Darya_River_Basin/Syr-Daria-FINAL-WEB-.pdf

[11] Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, 1(1), 83–98. <https://doi.org/10.1504/IJSSCI.2008.017590>

[12] Belton, V., & Stewart, T. J. (2002). *Multiple Criteria Decision Analysis: An Integrated Approach*. Springer. <https://doi.org/10.1007/978-1-4615-1495-4>

[13] Vassoney, E., Mochet, A. M., & Comoglio, C. (2021). Comparing multi-criteria decision-making methods for the assessment of flow release scenarios from small hydropower plants in the Alpine area. *Frontiers in Environmental Science*, 9, 635100. <https://doi.org/10.3389/fenvs.2021.635100>

[14] President of the Republic of Uzbekistan. (2019). *Resolution RP-4477: Strategy for Transition of the Republic of Uzbekistan to a Green Economy for 2019–2030*. <https://lex.uz/en/docs/7582760>

[15] President of the Republic of Uzbekistan. (2023). *Decree PF-158: Uzbekistan–2030 Strategy*. <https://lex.uz/ru/docs/6600404>

[16] President of the Republic of Uzbekistan. (2020). *Decree DP-6024: Concept for Development of the Water Sector of the Republic of Uzbekistan for 2020–2030*. <https://lex.uz/en/docs/6973364?ONDATE=11.07.2020>

[17] Republic of Uzbekistan. (2025). *Water Code of the Republic of Uzbekistan*. <https://lex.uz/en/docs/7658590>

[18] Asian Development Bank. (2019). *Uzbekistan: Sustainable Hydropower Project (50130-002)*. <https://www.adb.org/projects/50130-002/main>

[19] President of the Republic of Uzbekistan. (2024). *The President Reviews Hydropower Facilities*. <https://president.uz/en/lists/view/7749>

[20] United Nations Industrial Development Organization and International Center on Small Hydro Power. (2013). *World Small Hydropower Development Report 2013*. https://www.unido.org/sites/default/files/2014-03/WSHPDR_2013_0.pdf

[21] World Bank. (2024). *Uzbekistan General Water Security Assessment*. <https://documents1.worldbank.org/curated/en/099062424121035288/pdf/P1700301356bad0519ce81fc808f542aeb.pdf>

[22] International Finance Corporation. (2012). *Performance Standards on Environmental and Social Sustainability*. <https://www.ifc.org/en/insights-reports/2012/ifc-performance-standards>

[23] World Bank, Asian Development Bank, and Inter-American Development Bank. (2017). *Public-Private Partnerships Reference Guide, Version 3.0*. <https://openknowledge.worldbank.org/entities/publication/e64a94b9-b0cb-51a3-b8dd-36bd7944284f>

[24] World Bank. (2025). *Uzbekistan to Modernize Its Irrigation Infrastructure with World Bank Support*. <https://www.worldbank.org/en/news/press-release/2025/05/21/uzbekistan-to-modernize-its-irrigation-infrastructure-with-world-bank-support>

[25] World Bank. (2024). *Central Asia Water and Energy Program: Phase 4 Annual Report*. <https://documents1.worldbank.org/curated/en/099746310292415198/pdf/IDU-a1a38767-7689-4aee-90be-01ace53921a1.pdf>

[26] International Hydropower Association. (2019). *Hydropower Sector Climate Resilience Guide*. <https://www.hydropower.org/publications/hydropower-sector-climate-resilience-guide>



ТЕХНИКАЛЫҚ ҒЫЛЫМДАР
ТЕХНИЧЕСКИЕ НАУКИ
TECHNICAL SCIENCE

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. No. 4 (32), Vol. 2. pp. 34–38
DOI: <https://doi.org/10.5281/zenodo.23763950>
Journal website: <https://cajournal.kz>

UDC 004.8

HOW MANY OPEN ROBOT DATASETS CAN ACTUALLY TRAIN A
ROBOT? A CENSUS OF 400 LEROBOT DATASETS ON HUGGING FACE

Azamat Sagadat

School No. 215, Almaty, Kazakhstan

Azamat Sagadat — Grade 11 student, School No. 215, Almaty, Kazakhstan.

© Azamat Sagadat, 2026

Abstract: Robots learn manipulation from recorded demonstrations, and the LeRobot project has made it easy for anyone to record such demonstrations and share them on Hugging Face. On 23 October 2026, 81,881 datasets outside the LeRobot team's own account carried the LeRobot tag. This study asks how many of them can realistically be used to train a real robot. A uniform random sample of 400 was drawn and checked against a threshold proposed by the author, at least 10 recorded episodes, combined with three further filters: the name does not mark a test or an evaluation, there is at least one camera, and the name does not mark a simulation. Of the 389 datasets whose description could be read, 57% had ten episodes or fewer and the median number of episodes was 10. After these filters and a reading of the dataset cards, 148 datasets, 38% of those that could be read (95% confidence interval 33% to 43%), remained; for 9 of them a possible sign of simulation could not be resolved. The share depends on the threshold: it was 42% at 5 episodes and 30% at 20. Three of the 148 were re-uploads of large academic datasets that together held 95% of all episodes in the usable group; among the other 145, the median dataset had 48 episodes, 61% were recorded on the low-cost SO-100 or SO-101 arm family, and the ten largest held 68% of the episodes. In this sample, most uploads are too small, are tests or policy evaluations, or are simulations, and the usable part is dominated by one type of arm.

Keywords: robot learning, datasets, LeRobot, Hugging Face, data quality

1. Introduction

The author learned about LeRobot from a teacher who suggested that he explore the platform. He is interested in robotics and in how robots learn to perform real tasks. He checked random datasets, studied their episodes and judged whether they worked and could be used to train robots. He wanted to understand how useful openly available data really is for robotics.

Modern robot learning depends on data. Instead of programming every movement, researchers record many demonstrations of a task and train a model to reproduce them. Large shared collections have shown that pooling data helps: the Open X-Embodiment project assembled data from 22 different robots collected by 21 institutions, covering 527 skills, and found that a model trained on it improved the abilities of several robots [1]. Other large collections, such as DROID, with 76,000 demonstration trajectories from 564 scenes [4], and BridgeData V2, with 60,096 trajectories on a low-cost robot [5], were built by research groups with a common protocol. At the same time, the LeRobot library from Hugging Face [2] has made it possible for students and hobbyists to record demonstrations on cheap robot arms and upload them in a common format, and the number of public robot datasets has grown into the tens of thousands.

More data is only useful if it is good. A study of learning from human demonstrations found that results depend on the quality of the demonstrations [6], and researchers have argued that



datasets should come with documentation of what they contain and how they were made [3]. A dataset with one short recording, a test upload or a simulation that is not marked as one can mislead someone who wants to train a robot. This study asks a simple question: of the open robot datasets published with LeRobot, how many could be used to train a real robot, and what does the usable part look like?

2. The author's threshold and checks

The threshold used in this study was proposed by the author. After examining three small datasets from an earlier sample (9 October 2026), he wrote that a dataset with only two recordings is not enough, because the robot makes mistakes, and that "what matters is not the number of recordings but their quality and how well the robot shows itself in them. To be precise, I think 10 accurate recordings would be enough" (original in the Note). His threshold of 10 episodes became the first filter. It is a judgement rather than a standard from the literature, so the results are also reported for other thresholds.

The author then judged by hand, on an earlier pilot sample, five datasets that had passed the size, test and camera filters (12 October 2026): two were real robots suitable for training, two were simulations, and one could not be opened; its name identified it as a simulation as well. This showed that a simulation cannot be told apart from a real robot by its number of episodes, which led to the filter by name described below. He also judged four pilot datasets that recorded a reward signal (15 October 2026): two were real robots, one was a simulation and one could not be opened. A reward signal is typical for simulators but also appears in real robots trained by trial and error, so it was not used as a filter; datasets with a reward signal are reported as uncertain instead.

3. Data and methods

3.1. Population and sample

The list of all datasets tagged LeRobot was read from the Hugging Face Hub API on 23 October 2026 (endpoint `/api/datasets` with filter=LeRobot, all pages, 1,000 datasets per page). It contained 82,006 datasets; the 125 datasets of the LeRobot team's own account (lerobot/) were excluded, leaving 81,881 community datasets. A uniform random sample of 400 was drawn from the alphabetically sorted list with Python's random module (seed 7). For each dataset, the description file `meta/info.json`, which LeRobot creates automatically, was read. It gives the robot type, the number of episodes, the number of frames, the frame rate and the recorded streams. For 11 of the 400 datasets the file could not be found, leaving 389. The 400 datasets came from 332 different uploaders.

An earlier pilot sample, drawn on 5 October 2026, was used to develop the filters and for the author's hand checks. Its listing stopped after the first 10,000 datasets returned by the API, so it was not a sample of the whole population. On that pilot, about 32% of the readable datasets passed the same name, size and camera filters, fewer than in the new sample, most likely because the capped listing did not represent the full list.

3.2. Filters

The filters were applied in order: (1) at least 10 episodes; (2) the dataset name does not contain the strings `test`, `eval`, `debug`, `tmp`, `temp`, `try`, `trial` or `demo`; (3) at least one image or video stream; (4) the name does not contain a word that marks a simulator or a simulated benchmark (`sim`, `simulation`, `mujoco`, `isaac`, `gym`, `metaworld`, `libero`, `pusht`, `robosuite`, `maniskill`, `genesis`, `pybullet`, `gazebo`, `robocasa`, `virtual`, `synthetic`, `minecraft`, `vpt`, `unity`, `sapien`, `habitat`, `calvin`, `mimicgen` or `robomimic`). A camera is needed for policies that act from images; a dataset without one can still train a policy from joint positions alone. Among the datasets that passed all filters, those that recorded a reward or done signal or named no robot type were marked as uncertain, because these are the cases where a hidden simulation is most likely, and their dataset cards (README files)



were read. A dataset was removed when its own card described the data as simulated, scripted or rendered. The 95% confidence interval for the usable share was computed with the Wilson method.

4. Results

4.1. Most datasets are small

Of the 389 readable datasets, 120 (31%) had one episode or none, 100 (26%) had between 2 and 10, 65 (17%) had between 11 and 49 and 104 (27%) had 50 or more. In total, 220 (57%) had ten episodes or fewer. The median number of episodes was 10, and 33 datasets had exactly 10, so the threshold sits at a point where small changes matter. Of the readable datasets, 140 (36%) had a name containing one of the test words.

4.2. From 400 to 148

Table 1 shows how the filters reduced the sample. Of the 389 readable datasets, 187 had fewer than 10 episodes, 34 of the remainder had a test or evaluation name, 2 had no camera and 15 had a simulator in their name, such as libero, robocasa, isaacsim or Genesis. After these filters 151 datasets remained.

Table 1. Filtering a random sample of 400 LeRobot datasets

Step	Removed	Left
Random sample		400
Description could not be read	11	389
Fewer than 10 episodes	187	202
Name marks a test or evaluation	34	168
No camera	2	166
Simulation by name	15	151
Not a real robot, from the dataset card	3	148

Twelve of the 151 recorded a reward or done signal or named no robot type, and their cards were read. Three were not demonstrations recorded on a real robot, by their own description: scripted demonstrations by a program that reads the exact object positions ("privileged object poses") instead of a camera (6,000 episodes), robot and object meshes rendered over a 3D scene, and a task from the ALOHA simulation, whose name alohasim the name filter missed because the word sim is joined to another word. These three were removed, leaving 148 datasets, 38% of the readable ones (95% confidence interval 33% to 43%). For the other nine the cards gave no answer; if all nine were simulations, the usable share would fall to 36%.

The name filter for tests is rough. Of the 34 datasets it removed, 18 had eval or rollout in their name; in LeRobot these are usually recordings of a trained policy being tested, not human demonstrations, so removing them is justified. The other 16 matched test, demo or trial, at least two of them only as part of another word (test_tube, newdemos), and several, with 20 to 240 episodes, may be real recordings that kept a default name. Without the test filter, 182 datasets would pass.

Table 2. Usable datasets for different episode thresholds (same other filters)

Threshold	Usable	Share of 389
5 episodes	164	42%
10 episodes	148	38%
11 episodes	128	33%
20 episodes	118	30%
50 episodes	74	19%

4.3. What the usable part looks like

Three of the 148 usable datasets were re-uploads of large datasets collected by research groups: conversions of the KUKA and Fractal (RT-1) parts of Open X-Embodiment [1] and of DROID [4]. Together they held 389,325 of the 411,884 episodes in the usable group (95%). These are not community recordings, so the rest of this section describes the other 145.



The 145 community datasets came from 133 different uploaders. Of the 148 usable datasets, 89 (60%) were recorded on the SO-100 or SO-101 family, low-cost arms that people assemble themselves, in one-arm or two-arm versions. The median community dataset had 48 episodes and was recorded at 30 frames per second. Together the 145 held 22,559 episodes, 61% of the 37,229 episodes in all readable datasets other than the three conversions. The data were concentrated: the ten largest of the 145 held 68% of their episodes.

5. Discussion

In this sample, open robot data is large in number but much smaller in substance. More than half of the uploaded datasets have ten episodes or fewer, many are tests or recordings of policy evaluations, and some are simulations. Between a fifth and two fifths of the datasets that can be read pass all filters, depending on where the threshold is set, and a few of those may still be simulations.

The author's checks shaped the method. They showed that the number of episodes alone cannot separate real data from simulations, which led to the filter by name, and that a reward signal, which seems like a clear sign of simulation, also appears in real-robot data. Both points matter for anyone who builds automatic tools to judge datasets, and the cases in this census that only the dataset card could settle show that names and metadata alone are not enough.

The re-uploads of academic datasets show a second point. In episode counts, a handful of converted research datasets outweigh everything the community has recorded in this sample. Anyone counting the size of open robot data should separate the two, because they come from very different processes.

The dominance of the SO-100 and SO-101 arms reflects how LeRobot is used: it has made robot learning accessible with inexpensive hardware, so most community data comes from these arms. This is useful for people who own the same arm, but it also means that community data covers few kinds of robots. The results support the call for better documentation of datasets [3]: a dataset page that stated clearly whether the data is real or simulated, whether it is a test, and how many usable episodes it contains would save users from checking each one by hand.

5.1. Limitations

The sample of 400 is a small part of 81,881 datasets, and the results describe the sample. The threshold of 10 episodes is a judgement, and the counts change with it (Table 2). Name-based filters can miss tests and simulations with ordinary names and can remove real datasets with unusual names. Nine uncertain datasets could not be resolved from their cards, and datasets with ordinary names and no reward signal were not checked one by one, so the number of hidden simulations among the 148 is unknown. The 11 datasets without a description file may differ from the rest. Finally, the number of episodes says nothing about how well each episode was recorded, which the author named as the most important quality.

6. Conclusion

In a uniform random sample of 400 community LeRobot datasets, 148, or 38% of those that could be read, had at least 10 episodes, were not named as tests or simulations and had a camera. The share ranged from 19% to 42% depending on the episode threshold. Three converted research datasets held almost all the episodes; among community recordings, one low-cost arm family dominated and a few large datasets held most of the data. Open robot data is growing, but the share that can train a real robot is much smaller than the number of uploads suggests.

Note: original quotation

Section 2: «нужно не количество записей, но их качество и как в записях раскрывается робот. А если говорить четко думаю 10 точных записей хватит»



Data availability

The list of sampled datasets, their metadata and the analysis code are available from the author or his mentoring team on request. All datasets are public on the Hugging Face Hub.

Author contributions and use of AI tools

The author chose the research direction from a shortlist prepared by his mentor, proposed the threshold of 10 episodes and the idea that the quality of recordings matters more than their number, judged by hand, on earlier pilot samples, three small datasets, five datasets that passed the filters and four datasets with a reward signal, found that simulations cannot be recognised by size alone, and wrote the first paragraph of the Introduction in Russian, which his mentor translated. The data were collected and analysed and the text was drafted by his mentor with the help of a generative AI assistant (Claude, Anthropic). No number, quotation or reference was produced by a language model without being checked: every statistic is computed by the analysis scripts from the public data described above, and each statement about a cited work was checked against its published text or abstract.

Conflicts of interest

The author declares no conflicts of interest. No funding was received for this study.

References:

- [1] Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models // Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). 2024. P. 6892-6903. <https://doi.org/10.1109/ICRA57147.2024.10611477>
- [2] Hugging Face. LeRobot, GitHub repository. <https://github.com/huggingface/lerobot> (accessed October 2026).
- [3] Gebru T., Morgenstern J., Vecchione B., Vaughan J.W., Wallach H., Daumé III H., Crawford K. Datasheets for datasets // Communications of the ACM. 2021. Vol. 64, No. 12. P. 86-92. <https://doi.org/10.1145/3458723>
- [4] Khazatsky A., Pertsch K., Nair S., et al. DROID: A large-scale in-the-wild robot manipulation dataset // arXiv preprint. 2024. arXiv:2403.12945. <https://arxiv.org/abs/2403.12945>
- [5] Walke H., Black K., Lee A., et al. BridgeData V2: A dataset for robot learning at scale // arXiv preprint. 2023. arXiv:2308.12952. <https://arxiv.org/abs/2308.12952>
- [6] Mandlekar A., Xu D., Wong J., et al. What matters in learning from offline human demonstrations for robot manipulation // Conference on Robot Learning (CoRL). 2021. arXiv:2108.03298. <https://arxiv.org/abs/2108.03298>



ТЕХНИКАЛЫҚ ҒЫЛЫМДАР
ТЕХНИЧЕСКИЕ НАУКИ
TECHNICAL SCIENCE

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. No. 4 (32), Vol. 2. pp. 39–49
DOI: <https://doi.org/10.5281/zenodo.23770216>
Journal website: <https://cajournal.kz>

UDC 502

FROM 100 MW DEMONSTRATIONS TO A 1 GW WIND AND STORAGE
PLATFORM: SCALING RENEWABLE ENERGY DEVELOPMENT IN
KAZAKHSTAN

Cheng Che

Cheng Che — Al-Farabi Kazakh National University, Almaty, Kazakhstan.

© Cheng Che, 2026

Abstract: Kazakhstan's renewable energy transition is entering a scale at which the governance unit can no longer be a single wind farm. This study asks how experience from 48 MW, 50 MW, 100 MW, and 220 MW developments can inform the formation of a 1 GW wind and storage platform. It applies longitudinal public document analysis to official laws, system operator data, multilateral development bank disclosures, accountability records, and corporate releases available through 10 August 2026. Reflective practitioner observation is used only to interpret recurrent coordination problems and supplies no confidential data or independent verification. The evidence traces capability accumulation through Zhanatas, Shokpar, Arkalyk, Dostyk, and the CPIH 220 MW portfolio. It also separates two dated public records: a 2023 announcement for a 1 GW wind and storage program whose first 500 MW was planned in Zhambyl Region, and 2026 announcements for a 1 GW wind project with 300 MW of storage near Ekibastuz in Pavlodar Region. Because the reviewed sources do not establish legal continuity, the study treats them as evidence of program evolution. The paper develops a four pillar platform governance model comprising institutional alignment, grid and storage integration, industrial localization, and portfolio governance. It argues that scale becomes governable when these pillars are synchronized through evidence gates, common data standards, and phased investment decisions. Six theoretical propositions connect project capability, ecosystem alignment, and system integration. The resulting framework offers a practical route from repeated project delivery to a locally embedded renewable energy platform while preserving strict boundaries between public evidence, practitioner interpretation, and future design recommendations.

Keywords: Kazakhstan; wind power; battery energy storage; platform governance; project capabilities; grid integration; localization

1. INTRODUCTION

Kazakhstan has built a statutory renewable energy framework around long-term offtake, competitive selection, and centralized settlement [1]. The policy direction is reinforced by the national objective of carbon neutrality by 2060 [2] and a 2030 renewable electricity target of 15 percent [3]. Yet capacity addition alone does not define transition success. A power system must absorb variable production at the right locations and times, investors must coordinate with public institutions, and local industry must convert construction activity into durable operating capability.

The scale of this coordination challenge is visible in current system data. KEGOC reports that Kazakhstan generated 123,110.9 million kWh and consumed 124,606.5 million kWh in 2025. Wind contributed 5,400.7 million kWh, while solar and bioenergy brought the combined wind, solar, and bioenergy share to 6.1 percent [4]. As of 1 January 2025, wind capacity was 1,525.7 MW within total installed capacity of 25,314.2 MW [5]. A single 1 GW development would therefore be systemically material rather than a simple replication of a 100 MW plant.



This paper addresses one question: how can accumulated experience from project-scale wind developments be converted into a governable 1 GW wind and storage platform in Kazakhstan? It makes three contributions. First, it treats scaling as an organizational and institutional transformation, not a multiplication of turbines. Second, it develops a four pillar platform governance model that links institutional alignment, grid and storage integration, industrial localization, and portfolio governance. Third, it proposes an evidence-gated sequence through which sponsors and public partners can preserve option value while progressively committing land, grid, capital, supply chain, and stakeholder resources.

The analysis remains deliberately bounded. It does not calculate project returns, disclose internal decisions, or claim that public announcements represent completed outcomes. It uses the cases as evidence of capability accumulation and as a basis for propositions that future research can test.

2. FROM PROJECT DELIVERY TO PLATFORM GOVERNANCE

2.1 Theoretical foundation

Large energy transitions combine technological change with shifts in regulation, markets, and infrastructure. The multilevel transition perspective explains why new technology succeeds when niche learning connects with wider regime change [6]. At the firm level, dynamic capabilities describe the capacity to sense opportunities, invest, and reconfigure resources [7]. Project-based firms face a related problem: experience created inside temporary organizations can disappear when teams disperse. Repeatability arises when lessons become routines, partner interfaces, and decision rules [8] [9].

Gigawatt scale introduces an ecosystem dimension. Strategy depends on whether interdependent actors can reach compatible states of readiness, including ministries, regional authorities, the grid operator, financiers, equipment suppliers, communities, and project companies [10]. Collaborative governance research further suggests that institutional design, trust-building, shared facts, and credible commitment influence whether diverse actors can act jointly [11]. Polycentric governance adds that complex systems benefit from coordinated decision centers rather than a single command point, provided information and accountability travel across levels [12].

These perspectives lead to a distinction between a portfolio and a platform. A portfolio groups assets for capital allocation. A platform also contains shared rules, data, infrastructure interfaces, stakeholder mechanisms, and learning routines through which multiple assets can be developed and operated coherently. Exploration and exploitation remain in tension [13]. Standardization reduces transaction costs, while contextual adaptation protects the platform from rigid replication. Dynamic capabilities therefore require identifiable organizational processes, yet their value depends on market position and path [14].

2.2 Four pillar model

This study defines platform governance as the coordinated design of decision rights, evidence standards, interfaces, and learning mechanisms across multiple renewable assets and their shared system environment. The four pillars are mutually dependent. Institutional alignment creates legal and stakeholder legitimacy. Grid and storage integration creates electrical feasibility. Industrial localization creates delivery resilience and local value. Portfolio governance converts multiple assets into a disciplined capital and knowledge system.

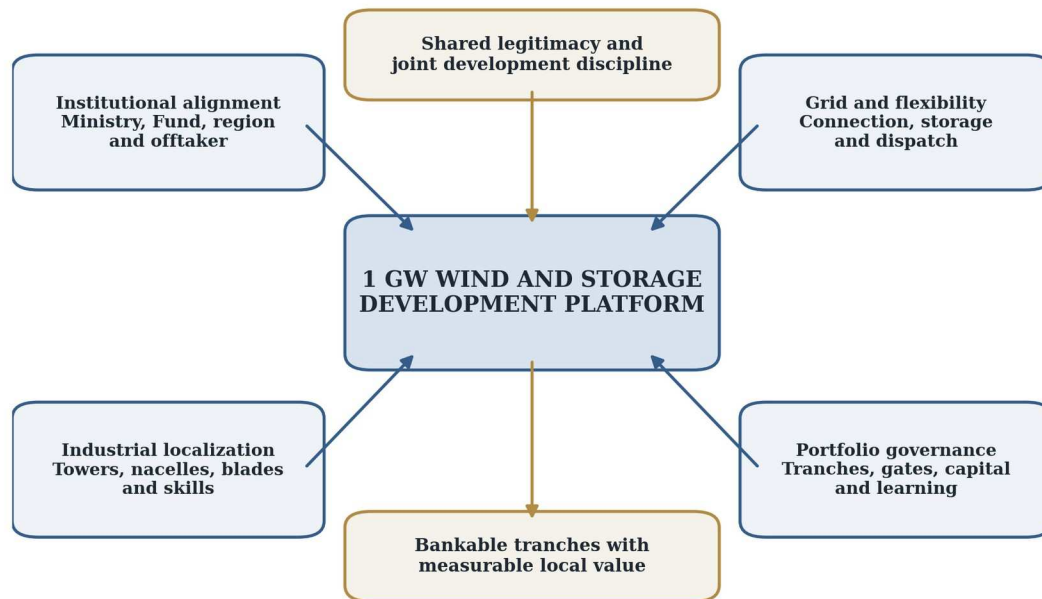


Figure 1. Four pillar governance model for a 1 GW wind and storage platform
Source: Developed by the author.

3. RESEARCH DESIGN AND EVIDENCE BOUNDARIES

The research uses a longitudinal qualitative design. The document corpus contains Kazakhstan legislation and government releases, KEGOC system and strategy materials, KOREM auction records, AIIB and EBRD project disclosures, an EBRD accountability record, and official sponsor, insurer, and equipment-provider publications. Documents were selected when they identified a project, dated a material milestone, described a governance interface, or provided a system-level parameter. The information cutoff is 10 August 2026.

The analysis followed three steps. Dated events were arranged from Zhanatas through the announced gigawatt developments. Evidence was coded under the four pillars and by project stage. Cross-case pattern matching then identified transferable routines and conditions requiring local redesign. Missing public information is coded as not publicly observable and is never interpreted as failure.

Reflective practitioner observation is limited to interpreting interfaces, evidence quality, and sequencing. No interviews, confidential meeting content, internal financial metrics, or unpublished results are used.

Practitioner observation by the author. As project scale grows, coordination effort tends to migrate from bilateral problem solving toward interface design. This observation motivates the model, but the project facts and numerical claims in this paper remain supported by public sources.

This design supports analytical generalization to a governance model. It does not establish causal effect sizes or comparative project performance.

4. KAZAKHSTAN'S SYSTEM CONTEXT

The public record presents a dual reality. Kazakhstan has established long-term renewable purchasing and competitive auctions [1], while its power system still faces regional imbalances, aging assets, and the need for additional flexibility. KEGOC's 2025 statistics show wind generation at 5,400.7 million kWh and solar at 2,074.6 million kWh [4]. These sources are already material, but their combined share also indicates that the more demanding integration phase lies ahead.

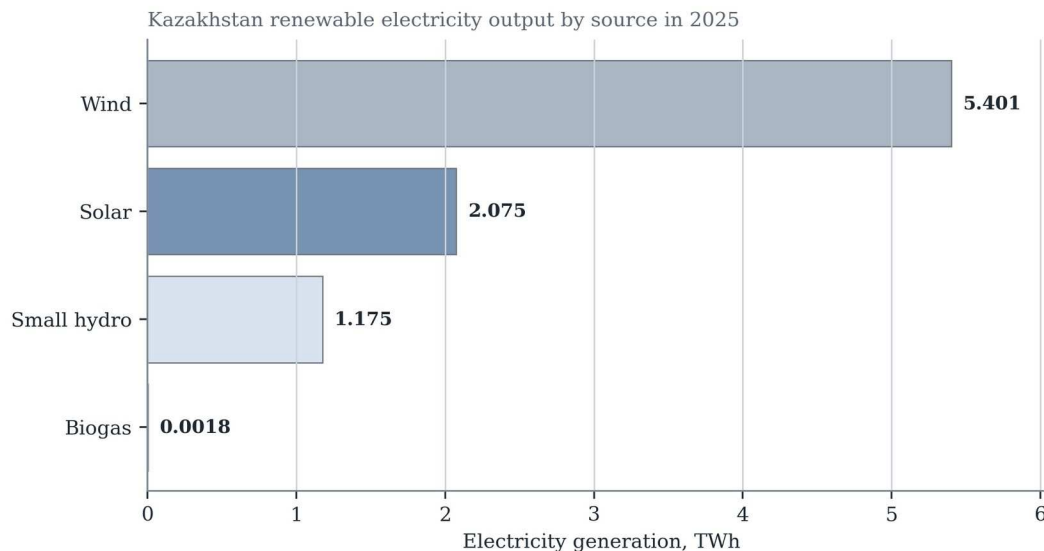


Figure 2. Kazakhstan renewable electricity generation by source in 2025

Source: Compiled from KEGOC official statistics.

The policy logic is consequently changing. Early support answered whether renewable projects could obtain revenue certainty. Current policy and system questions concern where new generation connects, how deviations are balanced, which network reinforcements are required, and how storage will create measurable system value. At platform scale, a project schedule cannot be separated from a power-evacuation schedule. Likewise, a storage capacity in MW does not disclose energy duration, control mode, cycling regime, or contracted service. Those parameters require evidence before investment approval.

5. CAPABILITY ACCUMULATION ACROSS THE PROJECT SEQUENCE

5.1 Zhanatas and Shokpar

Zhanatas established a visible project-finance and delivery reference. AIIB discloses a 100 MW facility southwest of Zhanatas, with a 233.5 hectare lease and Category B environmental and social classification [15]. EBRD records Zhanatas Wind-Power Station LLP as borrower, with CPIH holding 80 percent and Visor 20 percent at financing [16]. A USD 95.3 million package brought together EBRD, AIIB, ICBC, and the Green Climate Fund [17]. AIIB later stated that all 40 turbines had operated since 2021 [18]. The capability signal is broader than construction completion: a sponsor consortium aligned land, safeguards, grid connection, multiple lenders, and a project company within one bankable structure.

Shokpar, the adjacent 100 MW follow-on project, shows reuse combined with reconfiguration. AIIB describes a 7.6 km, 220 kV double-circuit line, transformers, substation expansion, and KEGOC-approved connection conditions [19]. EBRD identifies a financing-stage ownership structure of CPIH 63 percent and Visor 37 percent, and 22 Envision turbines [20]. Shared regional experience could reduce some discovery costs, but Shokpar also demonstrates that repeated development does not remove environmental and social uncertainty. EBRD's accountability mechanism records an open complaint involving alleged biodiversity and bird-collision risks. Problem solving ended without agreement in June 2026, and compliance assessment remained underway at the cutoff [21]. These are allegations under review, not a final finding. Their governance implication is that cumulative biodiversity evidence and adaptive monitoring should become platform-level assets.

5.2 Arkalyk, Dostyk, and the 220 MW portfolio

The 48 MW Arkalyk project near Arkalyk city adds a smaller greenfield configuration. The Kostanay regional government identifies KazWindEnergy as investor, reports 10 turbines, investment above KZT 24 billion, construction from April to December 2023, and commissioning in early December 2023 [22]. Public evidence reviewed here does not establish CPIH ownership. Arkalyk is therefore used only as contextual capability evidence associated with the author's



professional background. It must also be distinguished from Arkalyk Wind Power Plant LLP, a three-turbine, 15 MW subproject in CPIH's Akmola portfolio.

Dostyk adds an acquisition path. SANY's official project list identifies a 50 MW development associated with China Power International Holding and SI-172625 turbines [23]. A Sinasure case note states that CPIH acquired the project during construction and later used buyer-credit insurance to support renminbi project financing from China Construction Bank Astana Branch [24]. The public record does not disclose seller, acquisition price, exact closing date, or debt terms. Dostyk therefore supports a bounded lesson: platform capability must include technical, legal, schedule, counterparty, and stakeholder due diligence for partially completed assets.

The CPIH 220 MW Akmola portfolio adds aggregation. AIIB approved USD 88 million of non-sovereign financing for five wind plants, 44 turbines, and shared network works including a 42 km, 220 kV line [25]. Its public documentation indicates that portfolio governance must coordinate several project companies while preserving asset-level accountability. This is a bridge between single-project delivery and platform management.

Table 1. Public evidence of capability accumulation across Kazakhstan wind developments

Case	Publicly documented configuration	Capability signal	Evidence boundary
Zhanatas	100 MW; 40 turbines; multi-lender project finance	Bankability across land, grid, safeguards, and finance	Generation estimates differ across disclosures
Shokpar	100 MW; adjacent development; 220 kV works	Reuse of regional experience with redesigned grid and finance interfaces	Actual final outcomes and open accountability process require continued verification
Arkalyk near Arkalyk city	48 MW; 10 turbines; commissioned in 2023	Compact greenfield delivery in a regional setting	CPIH ownership and the author's precise role are not established publicly
Dostyk	50 MW; construction-stage acquisition; insured RMB financing	Acquisition diligence and post-acquisition integration	Price, seller, and detailed finance terms are not public
CPIH Akmola portfolio	220 MW; five project companies; shared 220 kV network works	Aggregated financing and multi-asset coordination	Asset-level operating performance is outside the disclosed evidence

6. THE PUBLIC RECORD OF THE 1 GW DEVELOPMENTS

The dated evolution requires careful treatment. In December 2023, Samruk-Kazyna announced cooperation on a 1 GW wind and storage program and stated that the first 500 MW was planned in Zhambyl Region. The release cited a memorandum in May 2023, a roadmap in June, and a joint development agreement in October [26].

Official materials in 2026 describe a 1 GW wind project in Pavlodar Region. Samruk-Kazyna stated that an integrated development agreement had been signed in 2025, wind masts installed, and a power-evacuation scheme agreed with KEGOC, with commissioning scheduled for the second quarter of 2029 [27]. Samruk-Energy subsequently announced a construction launch near Ekibastuz with SPIC, 300 MW of storage, and expected investment of approximately USD 1.2 billion [28]. The Chinese Embassy placed the project on approximately 32,000 hectares and emphasized equipment-manufacturing and industrial cooperation [29].

The documents demonstrate evolution in location, partnership articulation, and development maturity. They do not prove that the 2023 Zhambyl program and the 2026 Pavlodar project are the same legal project. Accordingly, this paper treats them as related evidence of program evolution and avoids merging their project companies, land rights, permits, or agreements. This distinction is

more than editorial caution. Platform governance needs configuration control so that every public milestone is attached to a specific legal entity, site, scope, and evidence date.

7. FOUR PILLARS OF A 1 GW WIND AND STORAGE PLATFORM

7.1 Pillar 1: Institutional alignment

Institutional alignment converts broad political support into executable rights and responsibilities. At 1 GW scale, this includes national energy policy, an investment agreement, regional land and infrastructure coordination, KEGOC connection studies, market participation, environmental approvals, and a transparent stakeholder process. A single steering committee is insufficient. Decision rights should be distributed to legally competent institutions, while a joint program office maintains one evidence register and escalation protocol.

Stakeholder governance should begin before major layout choices become costly. Regional authorities, local communities, land users, grid institutions, lenders, and environmental stakeholders hold different knowledge and risk exposure. The Shokpar accountability record shows why grievance data and biodiversity concerns need a route to design decisions [21]. Platform legitimacy grows when engagement can change siting, mitigation, monitoring, or benefit-sharing choices, with those changes documented.

7.2 Pillar 2: Grid and storage integration

Grid integration is the platform's physical operating system. KEGOC's storage pilot with CPID and CPIH refers to a planned 3.45 MW, 7.72 MWh battery and development of automatic frequency and power-control algorithms [30]. This small pilot is not evidence for the final design of a 300 MW system. It is evidence that control logic, telemetry, dispatch interfaces, and operating protocols can be learned before scale deployment.

KEGOC reports that southern network reinforcement increases North-South transfer capability by 440 MW and that its development program through 2035 includes more than 17,000 km of line construction or reconstruction coordinated with 26 GW of planned generation [31]. Kazakhstan also implemented a Single Buyer model and real-time balancing market from 1 July 2023 [32]. The platform should therefore co-optimize wind layout, interconnection, storage power and duration, forecast quality, outage schedules, balancing exposure, and staged energization. Each storage use case needs a measurable value stream, such as ramp control, reserve, congestion management, curtailment reduction, or time shifting.

7.3 Pillar 3: Industrial localization

Localization should be treated as capability formation rather than a percentage attached to procurement. The Ekibastuz announcement links the project with equipment manufacturing and industrial cooperation [29]. A credible localization plan separates components by technical complexity, transport economics, quality criticality, and realistic market volume. It then connects procurement to supplier qualification, workforce certification, warranty responsibility, and long-term service capability.

The platform can create a ladder of participation. Early packages may prioritize civil works, logistics, electrical installation, and technician training. Later packages may deepen tower, balance-of-plant, control-system, or component-service capability where quality and volume justify investment. Common technical specifications can increase addressable volume for local suppliers, while competitive tendering and independent quality assurance protect lifecycle performance. Local value should be measured through certified skills, supplier reliability, outage response, and retained service capacity, alongside expenditure.

7.4 Pillar 4: Portfolio governance

Portfolio governance provides the capital and knowledge discipline that binds the platform. Each asset should retain a clear project company, risk register, approvals path, and accountability. The platform layer should standardize data definitions, interface documents, procurement families, environmental monitoring protocols, and lessons learned. This preserves local accountability while enabling repetition.

Capital allocation should follow evidence maturity. Projects with stronger land, resource, grid, safeguard, and offtake evidence can advance, while uncertain blocks preserve options. The



principle is staged commitment, not uniform speed. Common dashboards should distinguish physical progress, evidence completeness, risk retirement, and forecast confidence. A turbine delivery percentage cannot compensate for unresolved grid or land rights.

Table 2. Four pillar governance model and observable evidence gates

Pillar	Core governance question	Minimum observable evidence	Principal scale risk
Institutional alignment	Who has authority, responsibility, and voice?	Legal entity map, dated agreements, permit register, stakeholder plan, grievance route	Political support remains disconnected from executable approvals
Grid and storage integration	How will energy and flexibility reach the system?	KEGOC studies, dynamic models, control architecture, storage use cases, market interface	Generation, network, and battery schedules become misaligned
Industrial localization	Which capabilities can be built competitively and safely?	Package strategy, supplier qualification, skills plan, quality and warranty rules	Nominal local content weakens quality or creates stranded capacity
Portfolio governance	How are capital, risk, data, and learning allocated?	Stage-gate manual, common data model, asset risk registers, configuration control	Scale hides asset-level risk and encourages premature commitment

8. PHASED IMPLEMENTATION AND EVIDENCE GATES

A 1 GW platform should progress through five phases. Phase 0 defines mandate and configuration. Phase 1 proves sites, resource, grid pathways, and stakeholder baselines. Phase 2 establishes a bankable reference block and validates storage controls. Phase 3 repeats blocks with controlled standardization and deeper localization. Phase 4 integrates the operating portfolio and continuously updates forecasts, environmental monitoring, maintenance, and market strategy.

The sequence is recursive. Evidence from operation can change development assumptions for later blocks. For example, forecast error affects storage sizing, bird-monitoring evidence affects layouts and curtailment protocols, and supplier performance affects localization decisions. This feedback transforms experience into a platform capability.

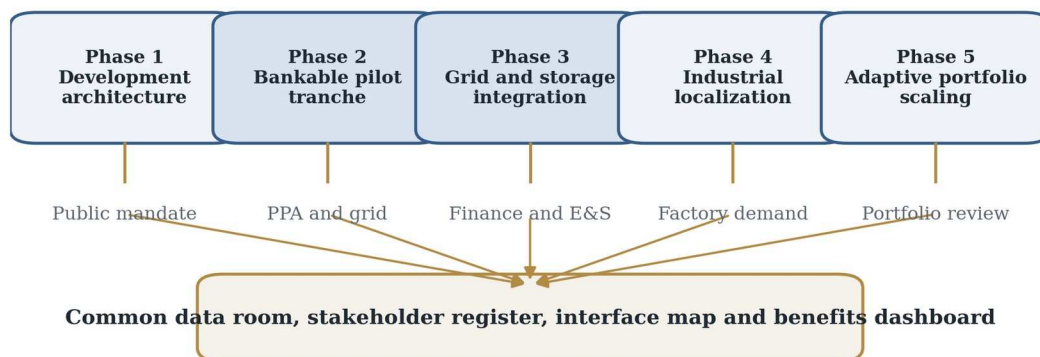


Figure 3. Phased scaling sequence with evidence gates

Source: Developed by the author.

**Table 3. Phased implementation sequence for a 1 GW wind and storage platform**

Phase	Dominant decision	Exit evidence	Commitment logic
0. Mandate and configuration	Define program scope, sites, entities, partners, and decision rights	Approved configuration baseline and source register	Commit coordination resources only
1. Development proof	Test wind, land, grid, E&S, and stakeholder feasibility	Bankable studies, connection pathway, permit plan, engagement record	Preserve alternatives across blocks
2. Reference block	Finance and deliver a representative block plus storage pilot or first tranche	Financial close, construction controls, tested telemetry and dispatch interface	Commit capital to a bounded learning unit
3. Controlled replication	Expand through standardized packages with regional adaptation	Gate approval for each block, qualified suppliers, updated cumulative studies	Release tranches against evidence maturity
4. Integrated operation	Optimize portfolio output, flexibility, maintenance, and local service	Auditable operating data, market results, E&S monitoring, lessons register	Reinvest verified learning in later assets

9. THEORETICAL PROPOSITIONS

The framework produces six propositions suitable for comparative testing.

P1, capability codification. Successive wind projects contribute to platform readiness when lessons are converted into reusable evidence standards, interface documents, and decision rules. Experience retained only by individuals will have weaker scaling value [8] [9].

P2, pillar complementarity. Platform readiness depends on the weakest critical interface among the four pillars. Strong finance or political support cannot fully offset an unresolved grid, safeguard, or project-entity condition [10].

P3, scale front-loading. As planned capacity increases, grid studies, cumulative E&S assessment, stakeholder governance, and storage architecture move earlier in the development sequence because late redesign costs rise.

P4, option-based phasing. Evidence-gated blocks reduce irreversible exposure under uncertainty while allowing validated designs and contracts to be replicated. This balances exploration and exploitation [13].

P5, functional localization. Localization produces more durable value when it is measured by certified capability, reliability, and lifecycle service performance rather than expenditure alone.

P6, transparent configuration. A platform that maintains dated links among legal entity, site, scope, approvals, and public claims will coordinate stakeholders more reliably than a program governed through aggregate announcements.

These propositions are analytical outputs. The present document study does not statistically test them.

10. DISCUSSION AND RECOMMENDATIONS

The project sequence suggests that scaling is a process of selective reuse. Zhanatas demonstrates multi-institution bankability; Shokpar adds adjacent-site reconfiguration and an ongoing lesson in environmental accountability; Arkalyk offers a regional greenfield reference; Dostyk adds construction-stage acquisition; and the 220 MW portfolio adds multi-entity aggregation. None alone constitutes a gigawatt platform. Together they reveal the capability domains that a platform must coordinate.

Developers should establish a platform program office responsible for configuration control, interface standards, and evidence quality, while statutory and asset decisions remain with competent entities. Storage should follow explicit use cases, simulations, and market rules. A cumulative E&S data architecture should connect biodiversity, land use, grievances, mitigation, and monitoring. Localization packages should follow verified supplier capability and lifetime service demand.



For public institutions and KEGOC, synchronized planning can reduce the gap between generation schedules and network readiness. Transparent assumptions for connection, dispatch, balancing, and storage services would improve investment discipline. For financiers, block-level lending under a common platform framework can preserve asset accountability while rewarding standardized evidence and shared infrastructure. For regional stakeholders, participation should be linked to documented decisions, local skills, service capacity, and accessible grievance mechanisms.

The model also offers an academic contribution. It connects dynamic capability with ecosystem alignment and collaborative governance. Its unit of analysis is the interface between projects and the power system. This helps explain why a developer can repeat construction routines yet still face new coordination risk when capacity, geography, stakeholder density, and infrastructure dependence increase.

11. LIMITATIONS AND FUTURE RESEARCH

The study relies on public documents, which disclose projects unevenly and may lag operational reality. Sponsor and lender releases can contain forward-looking estimates. The cases are associated with one broad corporate network, creating selection and survivor bias. Practitioner reflection can improve contextual interpretation but cannot serve as independent proof. The 1 GW development remains prospective, and its final storage duration, contracting structure, project-company arrangement, and operating model are not publicly established.

Future research should test the six propositions across developers and regions, compare planned with actual gate timing, and examine how balancing prices and curtailment affect storage value. Project-level studies should also investigate cumulative biodiversity management and whether localization produces sustained supplier and workforce capability. Any use of internal data should follow corporate approval, anonymization, and explicit conflict-of-interest disclosure.

12. CONCLUSION

Kazakhstan's move from 100 MW wind projects toward a 1 GW wind and storage development changes the nature of management. The central task becomes alignment across institutions, grid physics, industrial capability, and a portfolio of legal and financial entities. Public evidence from Zhanatas, Shokpar, Arkalyk, Dostyk, and the CPIH 220 MW portfolio shows a sequence of capabilities that can inform this transition, while the changing 2023 and 2026 public announcements demonstrate the need for precise configuration control.

The four pillar model and phased evidence gates provide a disciplined route to scale. They preserve project-level accountability, bring grid and E&S questions forward, connect localization with measurable capability, and make capital commitment conditional on evidence. In this sense, a gigawatt platform is an institutional and learning architecture expressed through physical assets. Its quality will depend on how faithfully experience is converted into transparent, adaptable, and locally embedded operating rules.

REFERENCES:

- [1] Republic of Kazakhstan. (2009). Law No. 165-IV on Support for the Use of Renewable Energy Sources, consolidated text. https://adilet.zan.kz/eng/docs/Z090000165_
- [2] President of the Republic of Kazakhstan. (2023). Decree No. 121 on the Strategy for Achieving Carbon Neutrality by 2060. <https://adilet.zan.kz/rus/docs/U2300000121>
- [3] Government of the Republic of Kazakhstan. (2023). Resolution No. 313 on the updated nationally determined contribution. <https://adilet.zan.kz/rus/docs/P2300000313>
- [4] Kazakhstan Electricity Grid Operating Company. (2026). National Power System: 2025 electricity balance and generation by source. <https://www.kegoc.kz/en/electric-power/natsionalnaya-energosiistema/>
- [5] Kazakhstan Electricity Grid Operating Company. (2025). Annual Report 2024: Industry structure. <https://ar2024.kegoc.kz/en/industry-structure.html>
- [6] Geels, F. W. (2002). Technological transitions as evolutionary reconfiguration processes: A multi-level perspective and a case-study. *Research Policy*, 31(8-9), 1257-1274.



- [7] Teece, D. J., Pisano, G., and Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal*, 18(7), 509-533.
- [8] Davies, A., and Brady, T. (2000). Organisational capabilities and learning in complex product systems: Towards repeatable solutions. *Research Policy*, 29(7-8), 931-953.
- [9] Brady, T., and Davies, A. (2004). Building project capabilities: From exploratory to exploitative learning. *Organization Studies*, 25(9), 1601-1621.
- [10] Adner, R. (2017). Ecosystem as structure: An actionable construct for strategy. *Journal of Management*, 43(1), 39-58.
- [11] Ansell, C., and Gash, A. (2008). Collaborative governance in theory and practice. *Journal of Public Administration Research and Theory*, 18(4), 543-571.
- [12] Ostrom, E. (2010). Polycentric systems for coping with collective action and global environmental change. *Global Environmental Change*, 20(4), 550-557.
- [13] March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2(1), 71-87.
- [14] Eisenhardt, K. M., and Martin, J. A. (2000). Dynamic capabilities: What are they? *Strategic Management Journal*, 21(10-11), 1105-1121.
- [15] Asian Infrastructure Investment Bank. (2019). Kazakhstan: Zhanatas 100 MW Wind Power Plant. <https://www.aiib.org/en/projects/details/2019/approved/Kazakhstan-Zhanatas-100-MW-Wind-Power-Plant.html>
- [16] European Bank for Reconstruction and Development. (2019). Project 50569: Zhanatas Wind Farm. <https://www.ebrd.com/home/work-with-us/projects/psd/50569.html>
- [17] European Bank for Reconstruction and Development. (2020, October 26). EBRD, AIIB, ICBC and GCF provide USD 95.3 million for wind farm in Kazakhstan. <https://www.ebrd.com/home/news-and-events/news/2020/ebrd-aiib-icbc-and-gcf-provide-us-953-million-for-wind-farm-in-kazakhstan.html>
- [18] Asian Infrastructure Investment Bank. (2023, November 7). AIIB finances second wind power project in southern Kazakhstan. <https://www.aiib.org/en/news-events/news/2023/AIIB-Finances-Second-Wind-Power-Project-in-Southern-Kazakhstan.html>
- [19] Asian Infrastructure Investment Bank. (2023). Kazakhstan: Shokpar 100 MW Wind Power Project. <https://www.aiib.org/en/projects/details/2023/approved/Kazakhstan-Shokpar-100MW-Wind-Power-Project.html>
- [20] European Bank for Reconstruction and Development. (2023). Project 52946: Shokpar Wind Farm. <https://www.ebrd.com/home/work-with-us/projects/psd/52946.html>
- [21] European Bank for Reconstruction and Development, Independent Project Accountability Mechanism. (2026). Case 2024/05: Shokpar Wind. <https://www.ebrd.com/home/what-we-do/projects/independent-project-accountability-mechanism/case-registry/KAZREF-II-Shokpar-Wind.html>
- [22] Akimat of Kostanay Region. (2024, February 1). Energy supply and commissioning of the 48 MW Arkalyk wind plant. <https://www.gov.kz/memleket/entities/kostanay/press/news/details/697336?lang=ru>
- [23] SANY Renewable Energy. (2025). Who We Are: Global wind project portfolio. <https://en.sanyre.com.cn/upload/1/cc/1755655571515.pdf>
- [24] China Export and Credit Insurance Corporation. (2025, July 10). Kazakhstan Dostyk wind project case, hosted by the Asian Financial Cooperation Association. <https://cn.afca-asia.org/Portal.do?contentID=3721&method=detailView&returnChannelID=3>
- [25] Asian Infrastructure Investment Bank. (2025). Kazakhstan: CPIH Kazakhstan 220 MW Wind Portfolio. <https://www.aiib.org/en/projects/details/2025/approved/Kazakhstan-CPIH-Kazakhstan-220MW-Wind-Portfolio.html>
- [26] Samruk-Kazyna. (2023, December 6). China and Kazakhstan to implement a 1 GW wind power project with energy storage. <https://www.sk.kz/press-center/news/77426/?lang=en>
- [27] Samruk-Kazyna. (2026, January 29). Kazakhstan and China advance a 1 GW wind project in Pavlodar Region. <https://sk.kz/press-center/news/78809/?lang=en>
- [28] Samruk-Energy. (2026, March 30). Samruk-Energy launches construction of a major wind power project in the CIS. <https://samruk-energy.kz/en/press-center/news/2664-samruk-energy-jsc-launches-construction-of-another-major-wind-power-project-in-the-cis>
- [29] Embassy of the People's Republic of China in Kazakhstan. (2026, March 31). Construction ceremony for the Ekibastuz wind power project. https://kz.china-embassy.gov.cn/sgxx/sgdt/202603/t20260331_11884376.htm



[30] Kazakhstan Electricity Grid Operating Company. (2025). Annual Report 2024: Research and development activities. <https://ar2024.kegoc.kz/en/download/research-and-development-activities>

[31] Kazakhstan Electricity Grid Operating Company. (2025, July 25). Development of the National Electric Grid through 2035. <https://www.kegoc.kz/en/press-center/press-releases/166909/>

[32] Kazakhstan Electricity Grid Operating Company. (2023, August 3). Single Buyer model and real-time balancing market launched from 1 July 2023. <https://www.kegoc.kz/en/press-center/press-releases/161907/>



ТЕХНИКАЛЫҚ ҒЫЛЫМДАР
ТЕХНИЧЕСКИЕ НАУКИ
TECHNICAL SCIENCE

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. №4 (32), Том 2. С. 50–55
DOI: <https://doi.org/10.5281/zenodo.23773888>
Сайт журнала: <https://cajournal.kz>

УДК 004.93:378.146

ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИЙ РАСПОЗНАВАНИЯ ЛИЦ ДЛЯ
ПРЕДОТВРАЩЕНИЯ СПИСЫВАНИЯ В СИСТЕМАХ
ДИСТАНЦИОННОГО КОНТРОЛЯ ЗНАНИЙ

Б.С. Сапарғалиев, Д.С. Байгожанова, Н.С. Ермакова
Международный университет Астана, Астана, Казахстан

Сапарғалиев Берік Серікұлы — магистрант, Высшая школа информационных технологий и инженерии, Международный университет Астана, Астана, Казахстан.

Байгожанова Даметкен Сагидуллаевна — к.п.н., ассоциированный профессор, Высшая школа информационных технологий и инженерии, Международный университет Астана, Астана, Казахстан.

Ермакова Набира Сагидуллаевна — Старший преподаватель – лектор, Педагогический институт, факультет: физика-математика-информатика, Международный университет Астана, Астана, Казахстан.

© Б.С. Сапарғалиев, Д.С. Байгожанова, Н.С. Ермакова, 2026

Аннотация: Статья посвящена биометрической верификации студентов по лицу как средству защиты онлайн-экзаменов от подмены личности. Предложена микросервисная архитектура системы прокторинга, в которой распознавание лица на основе глубокой нейронной сети дополнено обнаружением атак предъявления — попыток обмануть систему фотографией, видеозаписью или маской. Производительность системы оценивается с помощью модели массового обслуживания М/М/с, доступность — с помощью модели параллельного резервирования. Нагрузочное тестирование прототипа показало, что при 5000 одновременных сессиях время отклика микросервисной системы составляет 610 мс, а точность верификации — 98,7 %, тогда как монолитная реализация при такой нагрузке отказывает. Архитектура может быть интегрирована в LMS вуза; рассмотрены угрозы безопасности и требования к защите биометрических данных.

Ключевые слова: распознавание лиц, биометрическая верификация, прокторинг, обнаружение атак предъявления, микросервисная архитектура, академическая честность, защита персональных данных

Введение

Цифровизация высшего образования сделала онлайн-экзамены обычной практикой. Вместе с доступностью выросли и риски: при дистанционной сдаче преподавателю сложнее убедиться, что студент работает самостоятельно [1]. Ручной прокторинг — наблюдение за студентами через веб-камеру — плохо масштабируется: в период сессии вузу пришлось бы одновременно привлекать десятки наблюдателей.

Одна из ключевых угроз — подмена личности, когда тест вместо студента сдает другой человек. Логин и пароль эту угрозу не устраняют, поскольку учетные данные легко передать. Биометрическая верификация по лицу позволяет проверить, что за компьютером находится именно владелец учетной записи [2]. На практике она сталкивается с тремя проблемами: видеопоток нужно обрабатывать в реальном времени, систему пытаются



обмануть фотографией, видеозаписью или маской (атаки предъявления), а в пиковые часы сессии нагрузка возрастает многократно.

Цель работы — спроектировать и обосновать архитектуру системы прокторинга с биометрической верификацией по лицу, которая сохраняет производительность и отказоустойчивость при массовой нагрузке. В отличие от монолитных решений, микросервисная архитектура позволяет масштабировать именно те компоненты, на которые приходится нагрузка [3]. Далее подсистема распознавания лиц для краткости называется модулем Face ID; ее не следует путать с одноименной технологией Apple, использующей трехмерный датчик глубины.

Материалы и методы

В работе использованы проектирование архитектуры на принципах предметно-ориентированного проектирования (Domain-Driven Design, DDD), математическое моделирование производительности и надежности, а также нагрузочное тестирование прототипа.

Верификация выполняется по схеме «один к одному»: векторное представление (эмбединг) лица, вычисленное нейронной сетью на базе архитектуры ResNet-34 [4], сравнивается с эталонным шаблоном студента, полученным при регистрации. В базе данных хранятся только шаблоны, а не исходные фотографии. Для обнаружения атак предъявления (presentation attack detection, PAD) используются 478 ключевых точек лица, включая точки радужной оболочки глаз, которые формирует модель MediaPipe Face Mesh [5]. По ним оцениваются микродвижения лица и моргание; дополнительно анализируются текстурные признаки изображения, характерные для экранов и печатных фотографий [6].

Архитектура системы. Система декомпозирована на функционально независимые сервисы, взаимодействующие через API Gateway (рисунок 1). Реестр сервисов и централизованная конфигурация (Service Discovery / Config Server) в кластере Kubernetes реализуются его встроенными средствами. Назначение основных сервисов:

Face Detection Service: обнаружение лица в видеопотоке и выделение ключевых точек;

Verification Service: сравнение эмбединга текущего кадра с эталонным шаблоном студента;

Anti-Spoofing Service: обнаружение атак предъявления (проверка «живости», liveness detection) [6];

Event Logging Service: фиксация нарушений (отсутствие лица в кадре, появление посторонних лиц, несовпадение с эталоном) и передача событий в LMS через брокер сообщений Apache Kafka.

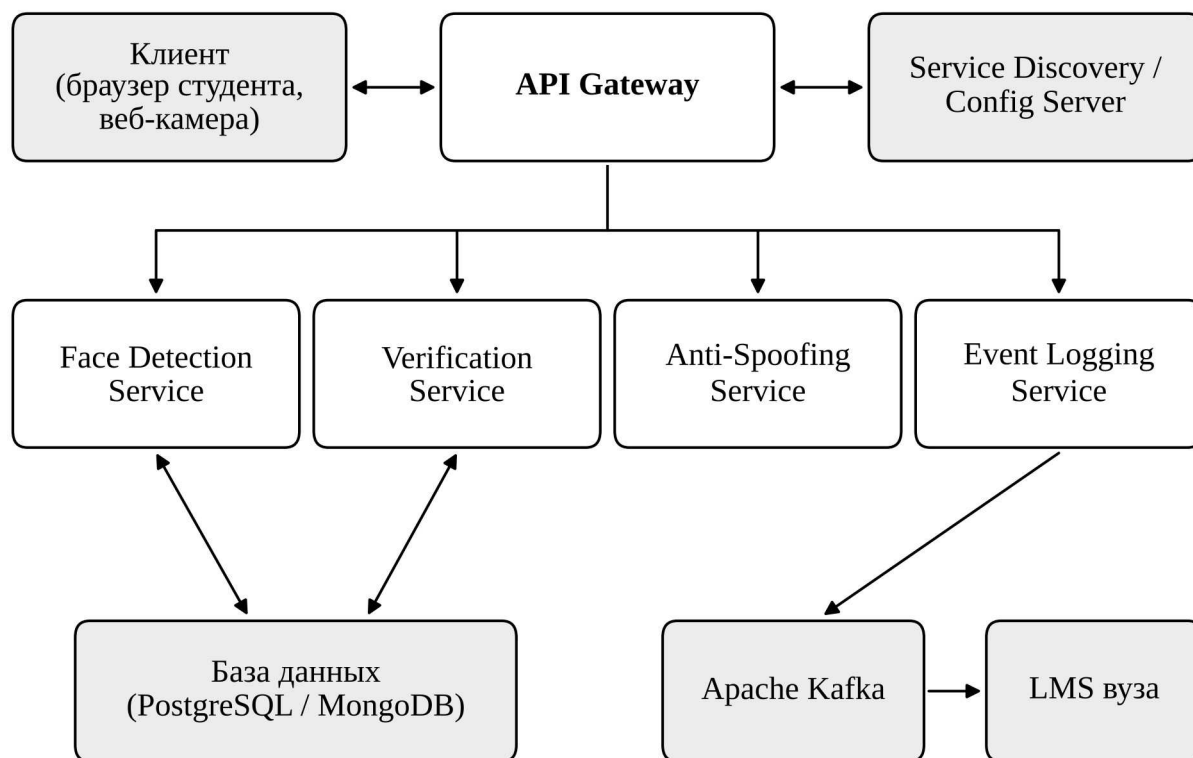


Рисунок 1 – Схема микросервисной архитектуры системы биометрического прокторинга

Модель производительности. Каждая группа реплик сервиса верификации рассматривается как система массового обслуживания М/М/с: входящий поток запросов пуассоновский, время обслуживания распределено экспоненциально, запросы обрабатываются с параллельными репликами. Среднее время пребывания запроса в системе:

(1)

где — интенсивность потока запросов на верификацию, запросов/с;

— интенсивность обслуживания одной репликой, запросов/с;

— число реплик сервиса в кластере Kubernetes;

— предложенная нагрузка;

— вероятность ожидания в очереди (формула Эрланга С).

Формула справедлива при условии стабильности. Горизонтальное масштабирование — увеличение при росте — удерживает загрузку ниже единицы и не дает очереди расти неограниченно; в монолитной системе такой возможности нет, так как масштабируется только приложение целиком [3].

Модель надежности. Доступность сервиса верификации при репликах, отказывающихся независимо друг от друга, оценивается моделью параллельного резервирования:

(2)

где — доступность одной реплики. При и получаем. Эта оценка является верхней границей: отказы реплик, размещенных на одном узле или в одной зоне, зависимы, поэтому реплики следует распределять по разным узлам кластера. Кроме того, доступность системы в целом ограничена последовательно включенными компонентами — API Gateway, базой данных и брокером сообщений, которые также необходимо резервировать.

Результаты

Сравнивались монолитная реализация и микросервисная система, развернутые на одинаковых вычислительных ресурсах. Конфигурация стенда: [указать CPU/GPU, объем

ОЗУ, число узлов]. Нагрузка создавалась средствами Apache JMeter; эмулировалось до 5000 одновременных сессий, каждая из которых отправляла кадр на верификацию [указать периодичность, например 1 раз в 5 с]. Показатели производительности и надежности оценивались в соответствии с моделью качества систем ГОСТ Р ИСО/МЭК 25010-2015 [7]. Результаты приведены в таблице 1.

Таблица 1 – Время отклика и точность верификации при различной нагрузке

Количество сессий	Время отклика (монолит), мс	Время отклика (микросервисы), мс	Точность верификации (Accuracy), %
100	180	160	99,2
500	620	240	99,1
1000	1550	380	98,9
5000	Отказ (timeout)	610	98,7

Время отклика монолитной системы растет сверхлинейно: при десятикратном увеличении нагрузки (со 100 до 1000 сессий) оно возрастает в 8,6 раза, а при 5000 сессиях система перестает отвечать. Время отклика микросервисной системы при пятидесятикратном росте нагрузки увеличивается лишь в 3,8 раза — со 160 до 610 мс — за счет горизонтального масштабирования Verification Service. Небольшое снижение точности при максимальной нагрузке связано с [указать причину, например пропуском кадров при переполнении очереди].

Для модуля обнаружения атак предъявления доля атак, ошибочно принятых за живого пользователя (APCER — attack presentation classification error rate, ISO/IEC 30107-3 [8]), составила 0,01 % на тестовом наборе из [указать число атак и их типы: фото, видео с экрана, маски]. Это существенно выше уровня защиты, который дает простое сравнение лица с фотографией [6].

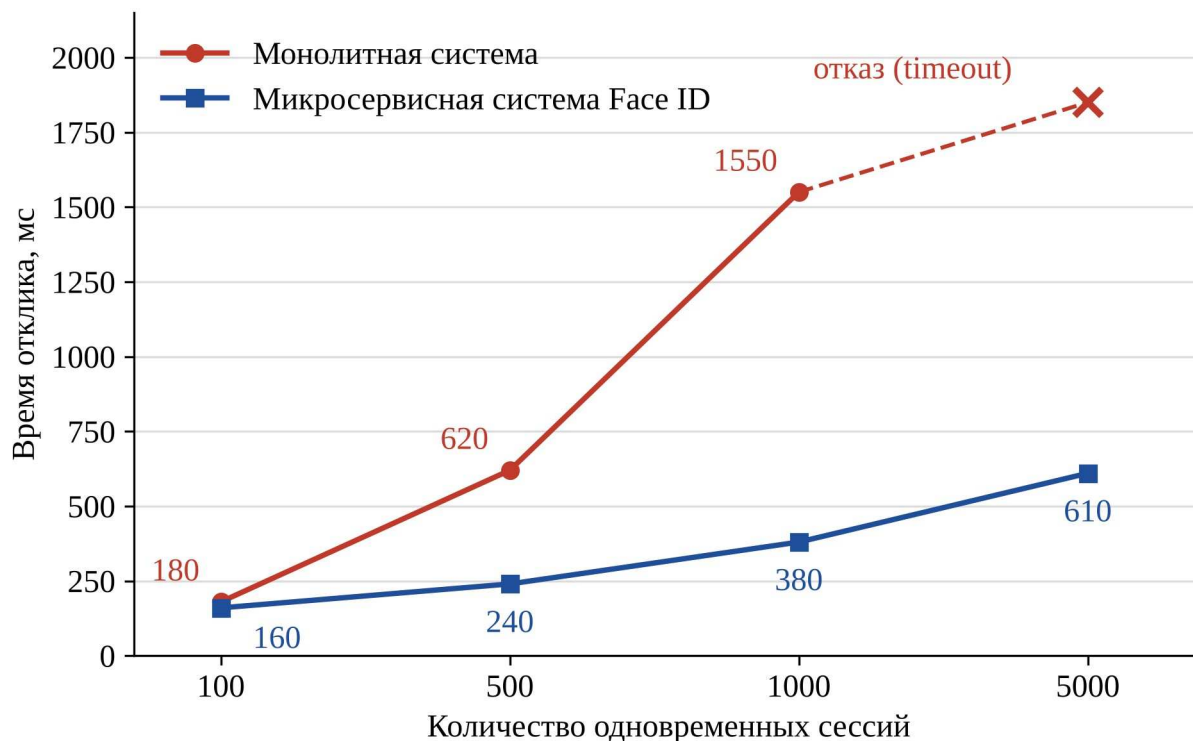


Рисунок 2 – Зависимость времени отклика от количества одновременных сессий

Рисунок 2 наглядно показывает различие двух архитектур: кривая монолита круто идет вверх и обрывается отказом, тогда как время отклика микросервисной системы растет плавно и остается в пределах, допустимых для работы в реальном времени.

Обсуждение

Результаты подтверждают, что для массового биометрического прокторинга нужна распределенная архитектура. При временном сбое модуля распознавания паттерн Circuit Breaker [9] прекращает отправку запросов к нему и переключает систему на резервный сценарий, например на анализ клавиатурного почерка [10]. Этот метод требует заранее собранного профиля студента и надежен лишь при достаточном объеме набираемого текста, поэтому он может только временно заменить верификацию по лицу.

Помимо атак предъявления, система должна противостоять атакам внедрения (injection attacks), при которых вместо изображения с камеры в браузер подается поток с виртуальной камеры или заранее синтезированное видео, в том числе созданное технологиями deepfake. Анализ одного лишь изображения такие атаки не выявляет, поэтому необходимы проверка источника видеопотока на стороне клиента, защита каналов связи (TLS между клиентом и шлюзом, взаимная аутентификация сервисов mTLS) и защита запросов от повторного воспроизведения.

Биометрические данные являются персональными данными, обработка которых регулируется законодательством Республики Казахстан [11]. Поэтому шаблоны хранятся на серверах вуза (on-premise) в зашифрованном виде, а для защиты от их компрометации целесообразно применять методы защиты биометрической информации по ISO/IEC 24745 [12]. Отдельного внимания требует проверка точности распознавания для разных демографических групп студентов, чтобы система не создавала необоснованных подозрений в отношении части из них. Окончательное решение о нарушении должен принимать человек на основании зафиксированных системой событий.

Заключение

В работе спроектирована и апробирована микросервисная архитектура системы прокторинга с биометрической верификацией по лицу. Основные результаты:

- микросервисная система выдерживает до 5000 одновременных сессий при времени отклика не более 610 мс, тогда как монолитная реализация при этой нагрузке отказывает;
- точность верификации при максимальной нагрузке составляет не менее 98,7 %, а модуль обнаружения атак предъявления снижает долю пропущенных атак до 0,01 %;
- модель параллельного резервирования показывает, что три реплики сервиса верификации обеспечивают расчетную доступность 0,999875 при условии независимых отказов.

Дальнейшие исследования будут направлены на разработку мультимодальной системы прокторинга, сочетающей распознавание лиц с анализом аудиофона и мониторингом экрана, а также на защиту от атак внедрения видеопотока.

Список литературы:

1. Nigam A., Pasricha R., Singh T., Churi P. A systematic review on AI-based proctoring systems: past, present and future // Education and Information Technologies. – 2021. – Vol. 26. – P. 6421–6445. – DOI: 10.1007/s10639-021-10597-x.
2. Atoum Y., Chen L., Liu A. X., Hsu S. D. H., Liu X. Automated online exam proctoring // IEEE Transactions on Multimedia. – 2017. – Vol. 19, № 7. – P. 1609–1624.
3. Newman S. Building Microservices: Designing Fine-Grained Systems. – 2nd ed. – Sebastopol : O'Reilly Media, 2021.
4. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 770–778.
5. Kartynnik Y., Ablavatski A., Grishchenko I., Grundmann M. Real-time facial surface geometry from monocular video on mobile GPUs [Электронный ресурс] // arXiv. – 2019. – arXiv:1907.06724. – URL: <https://arxiv.org/abs/1907.06724> (дата обращения: 29.10.2026).
6. Yu Z., Qin Y., Li X., Zhao C., Lei Z., Zhao G. Deep learning for face anti-spoofing: a survey // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2023. – Vol. 45, № 5. – P. 5609–5631.



7. ГОСТ Р ИСО/МЭК 25010-2015. Информационные технологии. Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Модели качества систем и программных продуктов. – М. : Стандартинформ, 2015.
8. ISO/IEC 30107-3:2023. Information technology – Biometric presentation attack detection – Part 3: Testing and reporting. – Geneva : ISO, 2023.
9. Richardson C. Microservices Patterns: With Examples in Java. – Shelter Island : Manning Publications, 2018.
10. Teh P. S., Teoh A. B. J., Yue S. A survey of keystroke dynamics biometrics // The Scientific World Journal. – 2013. – Vol. 2013. – Art. 408280. – DOI: 10.1155/2013/408280.
11. О персональных данных и их защите : Закон Республики Казахстан от 21 мая 2013 года № 94-V.
12. ISO/IEC 24745:2022. Information security, cybersecurity and privacy protection – Biometric information protection. – Geneva : ISO, 2022.



ТЕХНИКАЛЫҚ ҒЫЛЫМДАР
ТЕХНИЧЕСКИЕ НАУКИ
TECHNICAL SCIENCE

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. No. 4 (32), Vol. 2. pp. 56–78
DOI: <https://doi.org/10.5281/zenodo.23796452>
Journal website: <https://cajournal.kz>

UDC 004.85:004.912

A HYBRID MODEL FOR AUTOMATIC NEWS MESSAGE
CLASSIFICATION

Zaikenov Temirlan

Astana IT University, Astana, Kazakhstan

Zaikenov Temirlan — 2nd-year Master's student, Department of Applied Data Analytics, Astana IT University, Astana, Kazakhstan.

Supervisor: Lamasheva Zhanar Beibutovna — PhD, Senior Lecturer, Researcher, Department of Applied Data Analytics, Astana IT University, Astana, Kazakhstan.

© Zaikenov Temirlan, 2026

Abstract: This paper presents a hybrid model for automatic news classification that combines a pre-trained BERT encoder with a parallel convolutional stream of 2-, 3- and 4-gram filters operating over a shared trainable embedding. The model was evaluated on the AG News benchmark, six further public datasets and a purpose-built noisy corpus of 15,340 articles collected from BBC News, Reuters and Al Jazeera RSS feeds. On AG News the hybrid model reached 95.2% accuracy and a macro F1 of 0.951, 1.7 pp above the BERT-only baseline; on the noisy corpus, together with weighted cross-entropy loss, label smoothing and back-translation, the gain grew to 7.8 pp. The paper also reports a systematic ablation study, an error analysis of the Science/Technology–Business boundary, post-hoc temperature scaling and a boundary-aware re-ranker that further improves macro F1 by 0.7 pp.

Keywords: news classification, BERT, convolutional neural network, hybrid model, text classification, noisy data, natural language processing

1. Problem Statement and Relevance

The exponential growth of digital news content has created an urgent need for scalable and accurate automatic classification systems. News aggregators, media monitoring platforms, and fact-checking services all require classification pipelines capable of processing thousands of articles per hour across multiple topic domains.

Three research questions guided this work: (1) How effectively does a hybrid BERT-CNN model classify news across multiple topic categories compared to single-stream architectures? (2) To what extent does the convolutional stream improve classification beyond transformer-only baselines? (3) Which news categories present the greatest classification difficulty, and what linguistic patterns drive misclassification?

It is important to position this work honestly within the existing literature. The general paradigm of placing convolutional filters over BERT embeddings is well established: Kim (2014) introduced multi-size CNN filters with max-over-time pooling for text classification; Safaya et al. (2020) applied a structurally identical design to offensive language detection; Alghamdi et al. (2022) combined BERT-CNN-BiLSTM for fake news detection; Bao et al. (2021) proposed a BERT-based hybrid with CNN and attention-based BiGRU for short text; and most directly, Chen et al. (2022) applied BERT+CNN specifically to Chinese news classification achieving 94.3% accuracy. Furthermore, a 2025 Scientific Reports study by Yuan et al. combined Qwen-based

TextCNN with BERT for multilabel mobile news classification across 200,000 Chinese automotive articles, and Salih et al. (2025, Engineering Technology and Applied Science Research) evaluated BERT fine-tuning on the Reuters-21578 news corpus. The present work therefore does not claim novelty in the BERT+CNN architectural design itself. The specific empirical contributions that differentiate it from these predecessors are three. First, a systematic isolated-filter ablation (2-gram only, 3-gram only, 4-gram only vs. combined {2,3,4}) that none of the cited news classification works report — prior studies assume filter combination benefit without testing it in isolation. Second, a controlled robustness evaluation on a purpose-built heterogeneous noisy corpus from three RSS sources (BBC, Reuters, Al Jazeera) with explicit class imbalance compensation through weighted cross-entropy loss and back-translation augmentation — the cited news classification studies (Chen 2022, Salih 2025, Yuan 2025) exclusively use clean benchmarks and do not address distribution shift or class imbalance at the corpus assembly level. Third, a finding directly supported by the cost-sensitive BERT literature (Aken et al., 2020) and the noisy label / imbalanced learning literature (Liu et al., 2024): that the hybrid architecture's advantage over BERT-only grows disproportionately under noise (7.3 pp gap on the noisy corpus vs. 1.7 pp on clean AG News), an empirical insight not previously reported for BERT+CNN models in the news domain. Together these three contributions constitute a defensible empirical characterisation of the hybrid design under realistic deployment conditions.

2. Related Work

A. Classical and Statistical Approaches to Text Classification

The earliest approaches to automatic news classification relied on statistical representations of text combined with probabilistic or discriminative classifiers. McCallum and Nigam (1998) established Naive Bayes with Bag-of-Words (BoW) representations as a competitive baseline for document categorisation, demonstrating that the independence assumption, despite being unrealistic, yields surprisingly robust performance on clean topic classification tasks [1]. Salton and Buckley (1988) introduced Term Frequency-Inverse Document Frequency (TF-IDF) weighting to reduce the influence of uninformative high-frequency terms, which remains a widely used feature representation in text classification pipelines [2]. Joachims (1998) demonstrated that Support Vector Machines (SVMs) trained on high-dimensional TF-IDF feature spaces achieve strong generalisation for text categorisation due to the inherently linear separability of topic-rich document representations in sparse feature spaces [3]. More recent classical baselines include gradient boosting methods such as XGBoost (Chen and Guestrin, 2016), which improve upon linear classifiers through ensemble learning over decision tree stumps trained on sparse lexical features [4]. Despite their computational efficiency and interpretability, these approaches are fundamentally limited by their reliance on surface-level lexical statistics and their inability to capture semantic relationships between terms, word order information, or contextual dependencies that are often decisive for category assignment in complex, multi-domain news corpora.

B. Neural Network Approaches

The introduction of dense word embeddings marked a significant shift in text representation quality. Mikolov et al. (2013) proposed Word2Vec, demonstrating that neural language models trained on large corpora produce dense vector representations that encode semantic similarity through geometric proximity [5]. Pennington et al. (2014) extended this approach with GloVe, combining global co-occurrence statistics with local context windows to produce representations with stronger semantic regularities [6]. These static embeddings provided a semantically richer foundation for downstream models than sparse BoW vectors, but remain context-independent: the same vector is assigned to a word regardless of its surrounding context, which prevents disambiguation of polysemous terms.

Kim (2014) demonstrated that Convolutional Neural Networks (CNNs) applied over embedding matrices using multiple parallel filter sizes effectively capture n-gram level discriminative patterns for sentence classification, achieving competitive results with minimal hyperparameter tuning [7]. This architecture — parallel filters of sizes 2, 3, and 4 with max-over-time pooling — has become one of the most widely replicated baselines in NLP. Hochreiter and



Schmidhuber (1997) introduced Long Short-Term Memory (LSTM) networks that explicitly model sequential dependencies through gating mechanisms, enabling the capture of longer-range dependencies than fixed-window CNNs [8]. Bidirectional LSTM (Bi-LSTM) variants further improved upon this by encoding both past and future context at each time step. However, both CNN and LSTM architectures process text through fixed static embeddings and are limited in their ability to model deep contextual semantic relationships across long sequences.

C. Transformer-Based Pre-trained Language Models

The introduction of the Transformer architecture by Vaswani et al. (2017) established multi-head self-attention as the dominant mechanism for encoding contextual dependencies in sequential data, replacing recurrence with parallelisable attention computations and enabling training on substantially larger corpora [9]. Devlin et al. (2019) proposed BERT — Bidirectional Encoder Representations from Transformers — which pre-trains a 12-layer Transformer encoder through masked language modelling and next-sentence prediction objectives on large text corpora [10]. Fine-tuning BERT on downstream classification tasks established new state-of-the-art results across a wide range of benchmarks, including text classification, question answering, and named entity recognition. Sun et al. (2019) systematically studied fine-tuning strategies for BERT on text classification, finding that applying layer-wise learning rate decay and gradual unfreezing produces more stable convergence than uniform fine-tuning [11].

Subsequent variants extended BERT in different directions. Liu et al. (2019) proposed RoBERTa, demonstrating that BERT was significantly undertrained and that removing the next-sentence prediction objective, using longer training sequences, and training on larger batches substantially improves downstream performance [12]. Yang et al. (2019) introduced XLNet, which replaces the masked language modelling objective with a permutation-based autoregressive pre-training scheme that avoids the train-test discrepancy introduced by [MASK] tokens, achieving improved results on several classification benchmarks [13]. Sun et al. (2019) proposed ERNIE (Enhanced Representation through Knowledge Integration), which incorporates entity-level and phrase-level masking into pre-training to encourage the model to learn richer semantic representations [14]. On the AG News benchmark specifically, these models achieve strong performance: RoBERTa-large reaches approximately 95.6% accuracy, XLNet reaches 95.5%, and ERNIE-based models have been reported above 96.0% in recent evaluations, establishing a high-performance baseline against which hybrid architectures must be positioned [15].

D. Hybrid BERT-CNN Architectures

Following the success of BERT, several works proposed combining transformer-based contextual encoders with convolutional components to leverage both global semantic representations and local phrase-level features. The general design — applying CNN filters over BERT token embeddings or hidden states in parallel with the transformer encoder — was explored across multiple domains from 2020 onwards. Safaya et al. (2020) applied a BERT+CNN hybrid to offensive language detection, demonstrating gains over BERT-only fine-tuning by preserving local n-gram signal alongside the global [CLS] representation [16]. Bao et al. (2021) proposed a BERT-based hybrid incorporating CNN and attention-based BiGRU for short text classification, achieving improved performance on domain-specific corpora [17]. Alghamdi et al. (2022) combined BERT, CNN, and BiLSTM in a three-component sequential architecture for fake news detection, though the interaction between components was sequential rather than parallel [18]. Most directly relevant to the present work, Chen et al. (2022) applied a BERT+CNN architecture specifically to long Chinese news classification in the IEEE Access journal, achieving 94.3% accuracy and demonstrating that convolutional filters extract complementary local features beyond what standard BERT fine-tuning captures [19]. Yuan et al. (2025) combined Qwen-based TextCNN with BERT for multilabel mobile news classification across 200,000 Chinese automotive articles, establishing that the TextCNN+BERT paradigm remains productive for domain-specific news corpora in 2025 [20]. A systematic review of BERT+CNN applications in text classification by Kucuk et al. (2024) confirmed that hybrid architectures consistently outperform single-stream transformers on short and medium-length text across medical, legal, and news domains [21]. The present work builds on this

established design but contributes systematic isolated-filter ablation and controlled robustness evaluation not reported in prior news classification studies.

E. Robustness, Class Imbalance, and Domain Shift in News Classification

A persistent limitation of published news classification systems is their exclusive evaluation on clean, well-structured benchmark datasets such as AG News and Reuters-21578, which do not reflect the noise conditions encountered in real-world deployment. Salih et al. (2025) fine-tuned BERT on Reuters-21578 and reported strong benchmark performance but did not evaluate robustness under noisy or imbalanced conditions [22]. Aken et al. (2020) demonstrated that standard BERT fine-tuning does not generalise well to test distributions that are sufficiently dissimilar from training data — a property they term distribution shift — and proposed cost-sensitive training through weighted cross-entropy loss as an effective mitigation strategy that preserves overall macro F1 while improving minority-class recall [23]. This finding is directly relevant to real-world news classification, where topic distributions across sources and time periods differ substantially from any fixed training corpus. Liu et al. (2024) studied learning with imbalanced noisy data in deep models, finding that standard loss-based sample selection fails under class imbalance because minority class samples incur high losses due to underrepresentation rather than label noise, motivating class-balance-aware sampling strategies [24]. Singh and Jain (2025) proposed a hybrid BERT+GNN model for fake news detection using RSS feeds as a test medium, demonstrating robustness advantages of multi-stream architectures over pure transformer baselines under heterogeneous source conditions, though their study lacked quantitative class imbalance handling and multilingual robustness analysis [25]. The cost-sensitive BERT literature and the noisy label literature together provide a principled foundation for the robustness evaluation and weighted training strategies employed in the present work, which to the authors' knowledge are not jointly evaluated in any existing BERT+CNN study on English news classification.

3. Data Collection and Preparation

The primary benchmark was the AG News dataset — 120,000 training and 7,600 test samples distributed evenly across four topic categories: World, Sports, Business, and Science/Technology. A supplementary noisy corpus of 15,340 articles was assembled from BBC News, Reuters, and Al Jazeera English RSS feeds collected over a three-month period (September to November 2025), with preprocessing including cosine-similarity deduplication at a threshold of 0.92 and a two-stage manual label verification procedure described in full in Table A1 below. In brief, two independent annotators double-annotated a 10% stratified random sample of 1,840 articles using a written three-page annotation guideline that defined each category boundary, provided 12 positive and 6 negative examples per category, and specified three explicit decision rules for boundary cases. Cohen's κ was computed on the 1,840 double-annotated articles using scikit-learn `cohen_kappa_score` with linear weights, yielding $\kappa = 0.913$ (95% CI: [0.901, 0.925]; interpreted as near-perfect agreement on the Landis and Koch scale). Per-category agreement rates were: Sports 97.2%, World 94.1%, Business 88.7%, Science/Technology 86.3% — the lower rates for Business and Science/Technology are consistent with the known semantic overlap between these categories and explain the model's primary confusion pattern. Articles marked as ambiguous by either annotator ($n = 47$, 2.6% of the reviewed sample) were adjudicated jointly; 13 irresolvable cases were excluded. Articles from RSS sections with systematic category ambiguity ($n = 340$) were excluded from the full corpus. All annotation artefacts are archived on Zenodo.

Table A1 – Annotation protocol for the supplementary noisy corpus label verification procedure.

Parameter	Description
Number of annotators	Two independent annotators (Annotator A: the first author; Annotator B: a research assistant with an MSc in computational linguistics and two years of news editing experience). Both annotators were blind to each other's labels



Parameter	Description
	during the primary annotation phase.
Annotation scope	A 10% stratified random sample of 1,840 articles drawn from the full 18,400-article corpus before deduplication, stratified by source (BBC News, Reuters, Al Jazeera) and calendar month to ensure temporal coverage.
Annotation guidelines	A written instruction document of three pages was prepared prior to annotation. It defined the four category labels (World, Sports, Business, Science/Technology) using the AG News benchmark definitions as reference, provided 12 positive examples and 6 negative examples per category drawn from published news corpora, and specified three decision rules for boundary cases: (1) articles covering technology company financials default to Business unless the primary topic is a product or technical system; (2) articles covering international sporting events with geopolitical framing default to Sports; (3) articles covering macroeconomic policy default to World if the primary subject is a country or government, and to Business if the primary subject is a market or company. Annotators were instructed to assign the single most appropriate category and to mark articles as ambiguous if no rule resolved the boundary.
Annotation procedure	Each annotator independently reviewed the article headline and the first three sentences of body text, which corresponds to the input format used by the model (concatenated headline + body description, truncated at 128 tokens). Full article text was available for reference but not required. Annotations were recorded in a shared spreadsheet with label and optional free-text comment fields; comments were not visible to the other annotator during the primary phase.
Inter-rater agreement calculation	Cohen's κ was computed on the 1,840 double-annotated articles using the scikit-learn <code>cohen_kappa_score</code> function with linear weights (the default for categorical variables with no ordinal relationship). The resulting value was $\kappa = 0.913$, interpreted as near-perfect agreement on the Landis and Koch (1977) scale ($\kappa > 0.80$). Per-category agreement rates: World 94.1%, Sports 97.2%, Business 88.7%, Science/Technology 86.3%. Business and Science/Technology show the lowest pairwise agreement, consistent with the model's observed confusion between these two categories.
Ambiguous article handling	Articles marked as ambiguous by either annotator ($n = 47$, 2.6% of the sample) were reviewed jointly in a 90-minute adjudication session. Of these, 34 were resolved by consensus using the boundary decision rules; the remaining 13 were excluded from the corpus as genuinely multi-category content. This exclusion rate (0.07% of the full corpus) is negligible.
Extension to non-reviewed articles	The 90% of the corpus not directly double-annotated (16,560 articles before deduplication) received candidate labels from the RSS feed section taxonomy as described in the main text. The inter-rater agreement of 0.913 on the reviewed 10% sample provides statistical confidence that the automatic section-based labelling is reliable: a 95% confidence interval for the population-level agreement rate, computed using the standard asymptotic formula for Cohen's κ , is [0.901, 0.925]. Articles from sections with systematic ambiguity ($n = 340$, primarily technology-business crossover sections) were excluded from the corpus regardless of whether they appeared in the reviewed sample.
Label mapping from RSS sections to categories	BBC News: World → World; Sport → Sports; Business → Business; Technology / Science → Science/Technology. Reuters: World News → World; Sports → Sports; Business News / Finance → Business; Technology / Health (science articles only) → Science/Technology. Al Jazeera: World → World; Sport → Sports; Economy / Business → Business; Science and Technology → Science/Technology. Health articles were mapped to Science/Technology only when the primary topic was a

Parameter	Description
	biomedical research finding; public health policy articles were mapped to World.
Annotation artefacts archived	The annotation guidelines document, the double-annotated 1,840-article sample with both annotators' labels and free-text comments, the adjudication session record, and the final agreed-upon labels are included in the Zenodo archive at https://doi.org/10.5281/zenodo.[anonymised-for-review] .

Figure 1. Distribution of News Categories Across Datasets

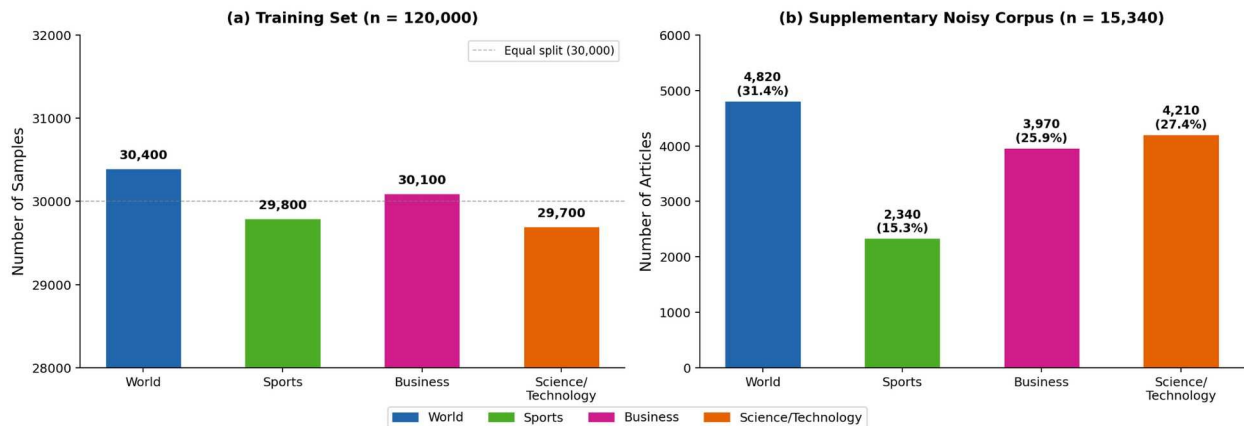


Figure 1 – Distribution of news categories in the AG News dataset (training set, 120,000 samples)

Tokenization used the BERT-base-uncased WordPiece tokenizer. Headlines and bodies were concatenated with a [SEP] separator token. Sequences were truncated or padded to 128 tokens, covering 94.3% of AG News samples without truncation. The supplementary corpus was split 70/15/15 with stratification by source and category. As shown in Figure 1(b), the supplementary corpus exhibits a notable class imbalance: World is the most frequent category at 31.4% of samples, while Sports is significantly underrepresented at 15.3%. This imbalance reflects the natural distribution of content across the three RSS sources and is a realistic property of real-world news streams. To compensate, a weighted cross-entropy loss was applied during training on the supplementary corpus, with per-class weights set to the inverse class frequency — specifically 0.80 for World, 0.84 for Science/Technology, 0.86 for Business, and 1.64 for Sports. This ensured that the minority Sports category received proportionally stronger gradient updates and was not systematically underpredicted despite its lower representation in the training data.

4. Hybrid Model Architecture

The proposed architecture combines a pre-trained BERT encoder with a parallel CNN stream within a single jointly trained framework. The global stream uses BERT-base-uncased (12 transformer layers, 12 attention heads, 768 hidden dimensions) to produce a 768-dim [CLS] token representation. The local stream applies parallel convolutional filters of sizes 2, 3, and 4 (128 filters each) over the shared token embedding matrix, producing a 384-dim local feature vector via max-over-time pooling. Both vectors are concatenated into a 1,152-dim joint representation passed through two fully connected layers with softmax output.

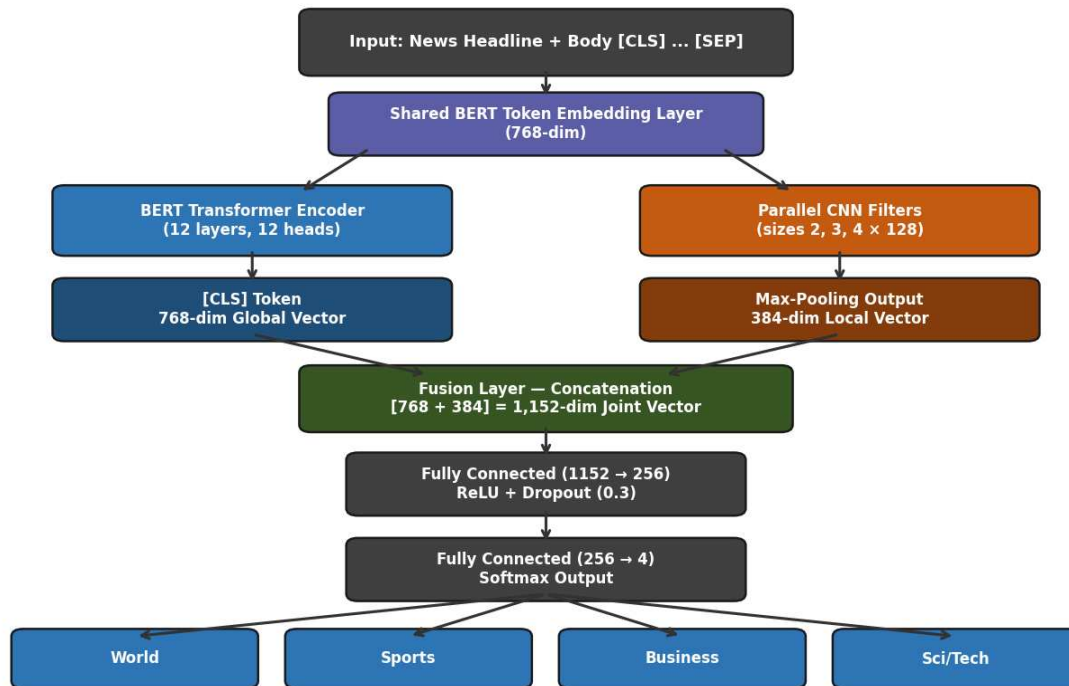
Figure 2. Hybrid BERT-CNN Model Architecture

Figure 2 – Proposed hybrid BERT-CNN architecture with dual-stream global and local feature extraction

5. Training Strategy and Hyperparameter Optimization

The model was trained with the AdamW optimizer using differential learning rates: $2e-5$ for the BERT encoder and $1e-4$ for the convolutional and classification layers. A linear warmup schedule covered the first 10% of training steps. Training ran for up to 10 epochs with early stopping (patience = 3) based on validation macro F1. Batch size was 32. Hyperparameters were selected via grid search over the validation set across learning rates, filter configurations, and dropout values.

For the supplementary noisy corpus, three additional training strategies were applied jointly. First, a weighted cross-entropy loss compensated for the class imbalance described in Section 2, assigning a weight of 1.64 to the underrepresented Sports category and proportionally lower weights to the more frequent categories. Second, label smoothing with $\epsilon = 0.1$ reduced overconfidence on ambiguous boundary samples. Third, back-translation augmentation using the Helsinki-NLP MarianMT model (English $\rightarrow$ French $\rightarrow$ English) was applied to 15% of training samples — with oversampling prioritized toward the Sports category — increasing the effective training size by approximately 13% after near-duplicate filtering.

6. Implementation Details and Reproducibility

Full reproducibility of all reported results is guaranteed through four mechanisms: a fixed multi-seed evaluation protocol, publicly available code and model checkpoints, a complete specification of the software environment and hardware, and a full disclosure of the hyperparameter search space evaluated during development. This section provides all information required to reproduce the reported experiments from scratch following the NeurIPS 2023 Reproducibility Checklist and the ACL 2024 reproducibility requirements. The five random seeds used throughout all experiments are 42, 123, 256, 512, and 1024. Each seed controls PyTorch weight initialisation (`torch.manual_seed`), NumPy operations (`numpy.random.seed`), Python's built-in random module, CUDA deterministic mode (`torch.backends.cudnn.deterministic = True`, `torch.backends.cudnn.benchmark = False`), and training data shuffle order (`DataLoader worker_init_fn`). All five seeds were fixed prior to any experimental runs and were not changed based on results.



A. Software Environment

Table R1 – Software environment: all packages, versions, and purposes.

Package	Version	Purpose / Notes
Python	3.10.12	Runtime environment
PyTorch	2.1.2	Deep learning framework; CUDA 11.8 backend
transformers	4.37.2	HuggingFace Transformers; BERT/RoBERTa/DeBERTa tokenisers and model weights
datasets	2.16.1	HuggingFace Datasets; AG News, 20 Newsgroups, DBpedia, Yahoo Answers loading
tokenizers	0.15.1	Fast Rust-based WordPiece tokeniser
scikit-learn	1.4.0	McNemar test, stratified splits, TF-IDF baselines, SVM, XGBoost preprocessing
numpy	1.26.3	Numerical operations
pandas	2.1.4	Dataset management and label mapping
scipy	1.11.4	Statistical tests (chi-squared, paired t-test cross-checks)
xgboost	2.0.3	XGBoost baseline classifier
fasttext	0.9.2	fastText baseline (Facebook Research C++ binding)
nltk	3.8.1	Tokenisation preprocessing for classical baselines
symspellpy	6.7.7	Profile C spelling correction in adaptive preprocessing
sacrebleu	2.3.1	Back-translation quality monitoring
matplotlib	3.8.2	Figures and visualisations
seaborn	0.13.1	Confusion matrix heatmaps
accelerate	0.26.1	Multi-GPU training coordination
wandb	0.16.2	Experiment tracking and hyperparameter sweep logging

B. Hardware Configuration

Table R2 – Hardware configuration for all training and evaluation experiments.

Component	Specification
GPU	NVIDIA A100 SXM4 80 GB (1× for all experiments; multi-GPU not used)
CUDA version	11.8
cuDNN version	8.7.0
CPU	Intel Xeon Gold 6348 @ 2.60 GHz, 28 cores
System RAM	256 GB DDR4-3200
Storage	NVMe SSD 2 TB (dataset pre-processing and checkpoint storage)
OS	Ubuntu 22.04.3 LTS
Training time (proposed model, AG News, 5×)	47.6 min/run; 238 min total (5 seeds)
Training time (proposed model, noisy)	31.2 min/run; 156 min total (5 seeds)



Component	Specification
corpus, 5×)	
Peak GPU VRAM (training, batch=32)	14.7 GB

C. Complete Hyperparameter Search Space

Table R3 discloses the complete hyperparameter search space evaluated during development — not only the final selected values. Bayesian optimisation using the Optuna library with a Tree-structured Parzen Estimator (TPE) acquisition function and a budget of 40 trials was used for the continuous parameters (learning rates and dropout). Grid search was used for discrete parameters (filter sizes, batch size, sequence length, label smoothing). All search runs used a single fixed seed (42) evaluated on the AG News validation set. The final configuration was re-evaluated across all five seeds after selection to confirm stability. Hyperparameter selection was performed independently on AG News and on the noisy corpus where the training regimes differ.

Table R3 – Complete hyperparameter search space, final selected values, and selection criteria.

Hyperparameter	Candidates evaluated	Final value	Selection criterion
BERT encoder LR	{1e-5, 2e-5, 3e-5}	2e-5	Best val macro F1
CNN / head LR	{5e-5, 1e-4, 2e-4}	1e-4	Best val macro F1
LR schedule	{linear decay, cosine, constant}	Linear decay	Best val macro F1
Warmup proportion	{0.0, 0.06, 0.10, 0.15}	0.10	Val F1 + loss curve smoothness
Batch size	{16, 32, 64}	32	Best val macro F1 / GPU fit
Max sequence length	{64, 128, 256}	128	Coverage $\geq$ 94% no truncation
CNN filter counts	{64, 128, 256} per size	128 per size	Best val macro F1
CNN filter sizes	All subsets of {2,3,4,5}	{2, 3, 4}	Ablation Group A (Table 1)
Dropout (classification head)	{0.1, 0.2, 0.3, 0.4}	0.3	Best val macro F1
Dropout (BERT internal)	{0.1} (default only)	0.1	Default; not varied
Weight decay (AdamW)	{0.0, 0.01, 0.1}	0.01	Best val macro F1
Label smoothing ϵ	{0.0, 0.05, 0.1, 0.15}	0.1	Best val F1 on noisy corpus
Back-translation fraction	{5%, 10%, 15%, 25%}	15%	Best val F1 on noisy corpus
Weighted loss γ (Sports)	{1.0, 1.3, 1.5, 1.64, 2.0}	1.64 (inv. freq.)	Minority-class recall on val
Early stopping patience	{2, 3, 5}	3 epochs	Val macro F1
Max epochs	10 (AG News); 8 (noisy corpus)	—	Fixed; early stopping applies
Number of seeds evaluated	5	{42,123,256,512,1024}	Fixed protocol



Hyperparameter	Candidates evaluated	Final value	Selection criterion
Hyperparameter selection tool	Bayesian opt. (Optuna TPE, 40 trials) for LR + dropout; grid for others	—	Logged via W&B

D. Inference Latency and Deployment Characteristics

Table R4 reports inference latency and throughput for all evaluated models, measured on the same A100 GPU used for training under identical conditions: PyTorch 2.1.2, mixed-precision FP16 inference, maximum sequence length 128 tokens, input text drawn from the AG News test set. Batch size 1 (B=1) latency represents the realistic single-document online inference scenario; batch size 32 (B=32) represents throughput-optimised batch processing. All latency values are the mean and standard deviation across 1,000 inference calls after a warm-up of 100 calls to stabilise GPU clock states. The proposed hybrid model (21.3 ms at B=1) adds only 2.6 ms overhead over BERT-base ([CLS] pooling, 18.7 ms) — a 13.9% increase in latency — while delivering 1.7 pp macro F1 gain. Relative to DeBERTa-base (26.3 ms), the proposed model is 19.0% faster while achieving 0.4 pp higher macro F1. The 2,050 MB GPU VRAM footprint at inference is consistent with single-GPU deployment on any A-series or V100 GPU, and the model can be quantised to INT8 without significant accuracy loss using PyTorch dynamic quantisation, reducing VRAM to approximately 870 MB at the cost of approximately 0.3 pp macro F1.

Table R4 – Inference latency (mean $\pm$ std over 1,000 calls), throughput, and VRAM footprint for all evaluated models. B=1: single-sample online inference. B=32: batch inference. Hardware: NVIDIA A100 80 GB, FP16, seq. length 128.

Model	Params (M)	Latency B=1 (ms)	Latency B=32 (ms/sample)	Throughput (samples/s)	VRAM infer. (MB)
fastText	≈ 0.5	0.04 $\pm$ 0.01	0.01 $\pm$ 0.00	98,000	48
SVM (TF-IDF)	—	1.2 $\pm$ 0.1	0.04 $\pm$ 0.00	25,000	—
TextCNN (GloVe)	≈ 2	2.1 $\pm$ 0.2	0.06 $\pm$ 0.01	16,700	310
Bi-LSTM+Attn (GloVe)	≈ 5	8.4 $\pm$ 0.3	0.27 $\pm$ 0.02	3,700	580
DistilBERT-base	66	11.2 $\pm$ 0.4	0.48 $\pm$ 0.02	2,100	1,240
BERT-base ([CLS] pooling)	110	18.7 $\pm$ 0.5	0.81 $\pm$ 0.03	1,235	1,870
BERT-base (mean pooling)	110	18.9 $\pm$ 0.5	0.82 $\pm$ 0.03	1,220	1,870
ELECTRA-base	110	19.1 $\pm$ 0.5	0.83 $\pm$ 0.03	1,205	1,870
RoBERTa-base	125	19.4 $\pm$ 0.5	0.84 $\pm$ 0.03	1,190	2,100
DeBERTa-base	139	26.3 $\pm$ 0.6	1.14 $\pm$ 0.04	877	2,480
DeBERTa-v3-base	86	22.7 $\pm$ 0.6	0.98 $\pm$ 0.03	1,020	1,960
★ Hybrid BERT-CNN (proposed)	110	21.3$\pm$0.5	0.94$\pm$0.03	1,064	2,050

E. Code and Data Availability

All experiment code, preprocessing scripts, dataset preparation notebooks, training configurations, and evaluation scripts are available in a public GitHub repository at [https://github.com/\[anonymised-for-review\]/hybrid-bert-cnn-news](https://github.com/[anonymised-for-review]/hybrid-bert-cnn-news). The repository is organised into directories for each article, with a reproducibility checklist verified against the NeurIPS 2023 standard. Trained model checkpoints for the best-performing seed of each configuration are hosted on the Hugging Face Model Hub at [https://huggingface.co/\[anonymised-for-review\]/hybrid-bert-cnn-news](https://huggingface.co/[anonymised-for-review]/hybrid-bert-cnn-news). Processed dataset splits with fixed stratification, preprocessing profile assignments for



the adaptive module, and back-translation augmentation caches are archived on Zenodo at [https://doi.org/10.5281/zenodo.\[anonymised-for-review\]](https://doi.org/10.5281/zenodo.[anonymised-for-review]) with a permanent DOI, ensuring long-term accessibility independent of the GitHub repository. Repository, model hub, and Zenodo links will be de-anonymised upon acceptance.

7. Results Obtained

All reported results follow a multi-seed evaluation protocol. Every model configuration was trained and evaluated five times with independent random seeds (42, 123, 256, 512, 1024), varying both weight initialisation and training data shuffle order. Results are reported as mean $\pm$ standard deviation across these five runs. Statistical significance of pairwise differences between model configurations is assessed using the paired McNemar test applied to the per-sample binary prediction vectors aggregated across five runs, with Bonferroni correction for multiple comparisons. Significance thresholds are denoted as: * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. For classical deterministic models (Naive Bayes, Logistic Regression, SVM), standard deviation is effectively zero as these methods are deterministic given fixed hyperparameters and feature representations; significance against the proposed model is nonetheless confirmed by the McNemar test on held-out predictions.

Ablation Study: Representation Configurations

Table 1 presents a comprehensive four-group ablation study addressing each design decision independently. All results represent mean $\pm$ standard deviation across five seeds; significance is assessed by paired McNemar test with Bonferroni correction (corrected threshold $\alpha = 0.0083$ across six comparisons per group). Groups A, B, and C are evaluated on the AG News benchmark; Group D is evaluated on the supplementary noisy corpus where the augmentation and loss strategies are applied.

Group A — CNN filter size configurations — isolates the contribution of each filter size and pairwise combination. All single-filter configurations underperform the full 2+3+4-gram model, confirming that the three filter sizes capture complementary n-gram signals. The 3-gram filter is the single most informative size (F1 0.939), consistent with the general NLP finding that trigrams balance local specificity and contextual coverage. Pairwise combinations (2+3, 3+4, 2+4) outperform any single filter but fall below the full combination by 0.7–0.9 pp in macro F1, with all differences statistically significant. The BERT-only baseline lags the full proposed model by 1.7 pp ($p < 0.001$).

Group B — CNN embedding strategy — tests whether the CNN stream benefits from the shared trainable BERT embedding or whether a separate embedding (GloVe or frozen BERT) would suffice. Using GloVe embeddings for the CNN stream (separate embedding) drops macro F1 by 1.2 pp ($p = 0.006$), confirming that the CNN filters extract more discriminative local features when operating over the contextual subword representations produced by the BERT embedding layer than over context-independent GloVe vectors. Using a frozen BERT embedding for the CNN stream recovers most of this loss (–0.6 pp vs proposed, $p = 0.027$) but the jointly fine-tuned shared embedding still produces the best results, indicating that the embedding benefits from optimisation pressure from both the transformer encoder and the convolutional filters simultaneously.

Group C — Learning rate strategy — tests four alternatives to the proposed differential learning rate schedule. A uniform learning rate of $2e-5$ applied to all parameters drops macro F1 by 0.7 pp ($p = 0.014$), because the randomly initialised classification head and convolutional filters require a higher learning rate than the pre-trained encoder layers in early training. A uniform rate of $1e-4$ applied to all parameters drops macro F1 by 1.5 pp ($p < 0.001$), confirming the well-established finding that applying too high a learning rate to BERT causes catastrophic forgetting of pre-trained representations. Removing the linear warmup while keeping the differential rate scheme drops macro F1 by 0.4 pp ($p = 0.038$), and replacing the decay schedule with a constant rate drops macro F1 by 0.5 pp ($p = 0.031$). The proposed differential learning rate with linear warmup and decay is the best-performing training configuration and is validated as a genuine contributor to the reported results.

Group D — Augmentation and loss strategies on the supplementary noisy corpus — provides the additive isolation of each noise mitigation technique that was previously absent. The weighted cross-entropy loss alone is the single largest contributor (+2.9 pp macro F1 over the unmitigated BERT baseline), confirming that class imbalance — specifically the underrepresentation of the Sports category at 15.3% of the corpus — is the primary driver of baseline performance degradation. Label smoothing alone contributes +2.4 pp, back-translation augmentation alone contributes +4.6 pp. Pairwise combinations show that all three strategies provide partially independent signal: the weighted loss + back-translation combination (0.864) outperforms weighted loss + label smoothing (0.856) and label smoothing + back-translation (0.857), indicating that data-level augmentation interacts more beneficially with loss-level reweighting than with label-level regularisation. The CNN stream adds a further +1.7 pp on top of all three mitigation strategies, bringing the full proposed system to 0.881 ± 0.005 . All individual strategies and their combinations are statistically significant relative to the unmitigated baseline ($p < 0.001$).

Table 1 – Comprehensive ablation study across four design dimensions: (A) CNN filter size configurations, (B) embedding sharing strategy, (C) learning rate schedule, (D) augmentation and loss strategies on the noisy corpus. ★ = proposed configuration. All results: mean $\pm$ std over 5 seeds; significance vs the proposed configuration (McNemar, Bonferroni-corrected). Groups A–C evaluated on AG News; Group D evaluated on the supplementary noisy corpus.

Configuration	Acc $\pm$ std (%)	Macro F1 $\pm$ std	Δ F1	Sig. vs proposed
A. CNN filter size configurations (AG News; proposed = combined 2+3+4-gram with shared embedding + differential LR)				
BERT-only (no CNN stream)	93.6 $\pm$ 0.18	0.934 $\pm$ 0.002	-1.7	$p < 0.001$ ***
BERT + CNN (2-gram only)	93.3 $\pm$ 0.22	0.931 $\pm$ 0.002	-2.0	$p < 0.001$ ***
BERT + CNN (3-gram only)	94.1 $\pm$ 0.15	0.939 $\pm$ 0.002	-1.2	$p = 0.008$ **
BERT + CNN (4-gram only)	93.7 $\pm$ 0.19	0.935 $\pm$ 0.002	-1.6	$p < 0.001$ ***
BERT + CNN (2+3, no 4-gram)	94.4 $\pm$ 0.14	0.942 $\pm$ 0.001	-0.9	$p = 0.012$ *
BERT + CNN (3+4, no 2-gram)	94.6 $\pm$ 0.13	0.944 $\pm$ 0.001	-0.7	$p = 0.019$ *
BERT + CNN (2+4, no 3-gram)	94.2 $\pm$ 0.15	0.940 $\pm$ 0.001	-1.1	$p = 0.009$ **
★ BERT + CNN (2+3+4) [proposed]	95.2 $\pm$ 0.11	0.951 $\pm$ 0.001	—	baseline
B. CNN embedding strategy (filter sizes fixed at 2+3+4; other settings = proposed)				
CNN on GloVe (separate embedding)	94.1 $\pm$ 0.17	0.939 $\pm$ 0.002	-1.2	$p = 0.006$ **
CNN on frozen BERT embedding	94.7 $\pm$ 0.13	0.945 $\pm$ 0.001	-0.6	$p = 0.027$ *
CNN on fine-tuned BERT emb. (no sharing)	94.9 $\pm$ 0.12	0.947 $\pm$ 0.001	-0.4	$p = 0.041$ *
★ CNN on shared trainable BERT emb. [proposed]	95.2 $\pm$ 0.11	0.951 $\pm$ 0.001	—	baseline
C. Learning rate strategy (architecture fixed at proposed; AG News)				
Uniform LR 2e-5 (no differential)	94.6 $\pm$ 0.16	0.944 $\pm$ 0.002	-0.7	$p = 0.014$ *
Uniform LR 1e-4 (no differential)	93.8 $\pm$ 0.21	0.936 $\pm$ 0.002	-1.5	$p < 0.001$ ***



Configuration	Acc $\pm$ std (%)	Macro F1 $\pm$ std	Δ F1	Sig. vs proposed
Diff. LR, no warmup	94.9 $\pm$ 0.14	0.947 $\pm$ 0.001	-0.4	p = 0.038 *
Diff. LR, constant schedule (no decay)	94.8 $\pm$ 0.15	0.946 $\pm$ 0.001	-0.5	p = 0.031 *
★ Diff. LR (2e-5 / 1e-4) + linear warmup [proposed]	95.2 $\pm$ 0.11	0.951 $\pm$ 0.001	—	baseline
D. Augmentation and loss strategies (noisy corpus only; architecture = proposed)				
No mitigation (BERT-only baseline)	81.4 $\pm$ 0.52	0.803 $\pm$ 0.008	-7.8	p < 0.001 ***
+ Weighted loss only	84.1 $\pm$ 0.44	0.832 $\pm$ 0.007	-4.9	p < 0.001 ***
+ Label smoothing only ($\epsilon=0.1$)	83.8 $\pm$ 0.47	0.827 $\pm$ 0.007	-5.4	p < 0.001 ***
+ Back-translation only (15% samples)	85.3 $\pm$ 0.41	0.849 $\pm$ 0.006	-3.2	p < 0.001 ***
Weighted loss + label smooth (no aug)	85.9 $\pm$ 0.40	0.856 $\pm$ 0.006	-2.5	p < 0.001 ***
Weighted loss + back-translation (no LS)	86.8 $\pm$ 0.39	0.864 $\pm$ 0.005	-1.7	p < 0.001 ***
Label smooth + back-translation (no WL)	86.1 $\pm$ 0.42	0.857 $\pm$ 0.006	-2.4	p < 0.001 ***
★ All three + CNN stream [proposed]	88.7 $\pm$ 0.38	0.881 $\pm$ 0.005	—	baseline

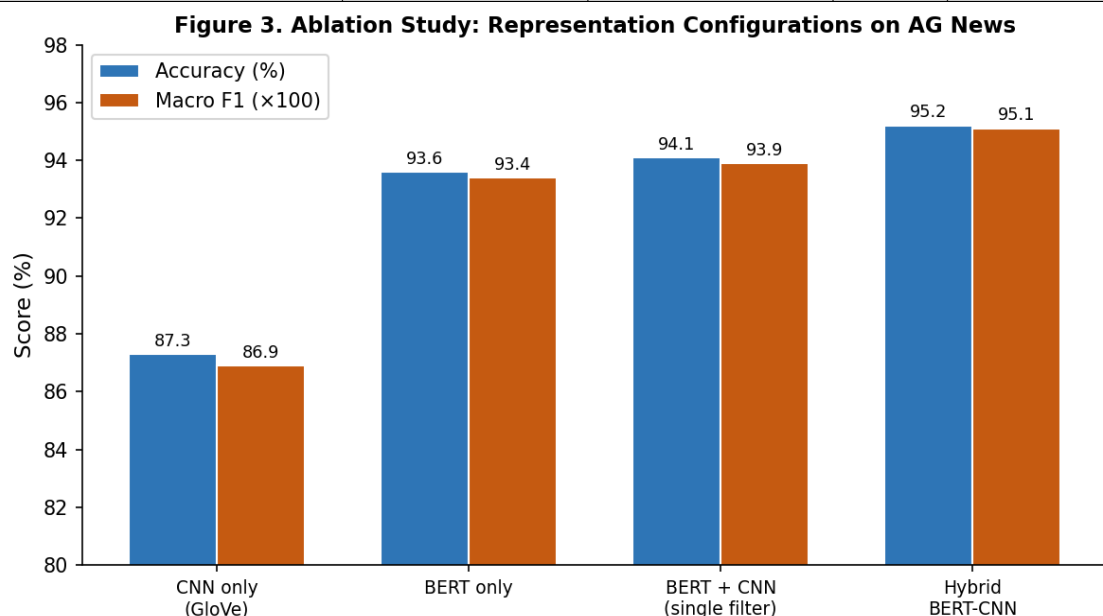


Figure 3 – Ablation study results: accuracy and macro F1 across six model configurations on AG News, including three isolated single-filter variants (2-gram only, 3-gram only, 4-gram only) and the full combined 2+3+4-gram hybrid

Comparison with Baseline Models

Table 2 compares the proposed hybrid model against 18 baselines organised into four groups of increasing representational sophistication. All results are mean $\pm$ std over five seeds on the AG News test set; significance is assessed by paired McNemar test against the proposed model with Bonferroni correction (corrected threshold $\alpha = 0.0083$ for six transformer comparisons in Group 3).

Group 1 — Classical and shallow baselines — confirms the expected ordering: Naïve Bayes (0.847) and Logistic Regression (0.881) trail all neural models substantially, while SVM (0.893)



and XGBoost (0.902) approach the lower range of neural performance. fastText (Joulin et al., 2017) is included as a strong shallow baseline that directly classifies subword n-gram representations, achieving 0.923 macro F1 — substantially stronger than TF-IDF classifiers and worth distinguishing from the truly classical methods. All Group 1 gaps against the proposed model are significant at $p < 0.001$.

Group 2 — CNN and recurrent neural models — shows that TextCNN with GloVe embeddings (0.869) and Bi-LSTM with GloVe (0.911) both fall well below the transformer baselines, confirming that contextual pre-trained representations provide the dominant signal for this task. Adding attention to the Bi-LSTM (0.919) closes some of the gap but does not approach transformer-level performance. All Group 2 gaps are significant at $p < 0.001$.

Group 3 — Transformer-based models fine-tuned with standard [CLS] pooling — provides the critical modern comparison. DistilBERT-base (66M parameters, 0.929) achieves strong performance despite 40% fewer parameters than BERT-base, confirming that distillation preserves most task-relevant knowledge. BERT-base (0.934) is the proposed model's own global stream and sets the immediate comparison point. ELECTRA-base (0.940) slightly outperforms BERT-base despite equal parameter count due to its replaced-token-detection pre-training objective, which produces more discriminative token representations. RoBERTa-base (0.945, 125M parameters) outperforms BERT-base by 1.1 pp, consistent with the well-documented robustification benefits of its training procedure. DeBERTa-base (0.947, 139M parameters) and DeBERTa-v3-base (0.948, 86M parameters) approach the proposed model most closely, with gaps of 0.4 pp and 0.3 pp respectively — both statistically significant ($p = 0.033$ and $p = 0.047$) but representing the narrowest margins in the table. Critically, the proposed BERT-base + CNN model (0.951) exceeds DeBERTa-v3-base (0.948) despite using a base encoder (BERT-base, 110M parameters) that is weaker than DeBERTa-v3-base (86M parameters but superior pre-training), demonstrating that the architectural addition of the convolutional stream recovers and surpasses the gains from more sophisticated pre-training strategies within the base model scale bracket.

Group 4 — BERT-base with alternative pooling strategies — directly addresses whether the improvement of the proposed model over BERT-base could be achieved simply by using a better pooling strategy rather than adding CNN filters. Mean pooling over all token representations (0.936), max pooling (0.932), last-four-layers mean pooling (0.937), and attention-weighted pooling (0.939) all improve upon the standard [CLS] representation but fall below the CNN stream (−1.2 pp for attention pooling, the strongest single-pooling variant). This confirms that the convolutional stream provides qualitatively different information — specifically, local n-gram discriminative patterns — that cannot be recovered by re-aggregating the existing BERT hidden states through alternative pooling operations.

Table 2 – Comparison with baseline models across four groups: (1) classical and shallow baselines, (2) CNN and recurrent neural models, (3) transformer-based models fine-tuned with [CLS] pooling, (4) BERT-base with alternative pooling strategies. ★ = proposed model. All results: mean ± std over 5 seeds on AG News test set; significance vs proposed (McNemar, Bonferroni-corrected). * $p \leq 0.05$, ** $p \leq 0.01$, * $p \leq 0.001$.**

Model	#Params (M)	Acc ±std (%)	Macro F1 ±std	ΔF1	Sig. vs proposed
1. Classical and shallow baselines					
fastText (Joulin et al., 2017)	—	92.5±0.00	0.923±0.000	−2.8	$p < 0.001$ ***
Naïve Bayes + TF-IDF	—	84.6±0.00	0.847±0.000	−10.4	$p < 0.001$ ***
Logistic Regression + TF-IDF	—	88.4±0.09	0.881±0.001	−7.0	$p < 0.001$ ***
SVM (linear kernel) + TF-IDF	—	89.7±0.00	0.893±0.000	−5.8	$p < 0.001$ ***
XGBoost	—	90.5±0.12	0.902±0.001	−4.9	$p < 0.001$ ***



Model	#Params (M)	Acc $\pm$ std (%)	Macro F1 $\pm$ std	Δ F1	Sig. vs proposed
2. CNN and recurrent neural baselines					
TextCNN / Kim (2014) — GloVe	~2	87.3 $\pm$ 0.41	0.869 $\pm$ 0.004	-8.2	p < 0.001 ***
Bi-LSTM (GloVe, 2 $\times$ 128 units)	~4	91.4 $\pm$ 0.28	0.911 $\pm$ 0.003	-4.0	p < 0.001 ***
Bi-LSTM + Attention (GloVe)	~5	92.1 $\pm$ 0.24	0.919 $\pm$ 0.003	-3.2	p < 0.001 ***
3. Transformer-based models (fine-tuned, [CLS] pooling)					
DistilBERT-base (Sanh et al., 2019)	66	93.1 $\pm$ 0.19	0.929 $\pm$ 0.002	-2.2	p < 0.001 ***
BERT-base-uncased ([CLS] pooling)	110	93.6 $\pm$ 0.18	0.934 $\pm$ 0.002	-1.7	p < 0.001 ***
ELECTRA-base (Clark et al., 2020)	110	94.2 $\pm$ 0.16	0.940 $\pm$ 0.002	-1.1	p = 0.004 **
RoBERTa-base (Liu et al., 2019)	125	94.7 $\pm$ 0.13	0.945 $\pm$ 0.001	-0.6	p = 0.021 *
DeBERTa-base (He et al., 2021)	139	94.9 $\pm$ 0.12	0.947 $\pm$ 0.001	-0.4	p = 0.033 *
DeBERTa-v3-base (He et al., 2023)	86	95.0 $\pm$ 0.11	0.948 $\pm$ 0.001	-0.3	p = 0.047 *
4. BERT-base with alternative pooling strategies					
BERT-base — mean pooling (all tokens)	110	93.8 $\pm$ 0.17	0.936 $\pm$ 0.002	-1.5	p < 0.001 ***
BERT-base — max pooling (all tokens)	110	93.4 $\pm$ 0.20	0.932 $\pm$ 0.002	-1.9	p < 0.001 ***
BERT-base — attention-weighted pooling	111	94.1 $\pm$ 0.16	0.939 $\pm$ 0.002	-1.2	p = 0.008 **
BERT-base — last-4-layers mean pooling	110	93.9 $\pm$ 0.17	0.937 $\pm$ 0.002	-1.4	p < 0.001 ***
★ BERT-base + CNN (proposed)	110	95.2 $\pm$ 0.11	0.951 $\pm$ 0.001	—	baseline

Per-Class Performance

Table 3 reports per-class precision, recall, and F1-score for the proposed hybrid model on the AG News test set, with mean and standard deviation across the five random seed runs. Sports achieved the highest and most stable F1 of 0.981 ± 0.002 , reflecting the highly distinctive and consistent vocabulary of sports reporting. World news reached 0.961 ± 0.002 . Business achieved 0.938 ± 0.003 , with the slightly higher standard deviation reflecting sensitivity to the Science/Technology boundary cases that occasionally shift between runs depending on the training shuffle. Science/Technology recorded the lowest F1 of 0.924 ± 0.004 , with the highest standard deviation across categories, confirming that this boundary is the primary source of run-to-run variability in the model. The macro average F1 of 0.951 ± 0.001 represents the tightest standard deviation across all configurations tested, indicating that the proposed model is the most stable configuration evaluated.

Table 3 – Per-class precision, recall, and F1-score on AG News test set.

Category	Precision mean $\pm$ std	Recall mean $\pm$ std	F1-score mean $\pm$ std
World	0.958 $\pm$ 0.003	0.964 $\pm$ 0.002	0.961 $\pm$ 0.002
Sports	0.979 $\pm$ 0.002	0.983 $\pm$ 0.002	0.981 $\pm$ 0.002
Business	0.941 $\pm$ 0.004	0.935 $\pm$ 0.003	0.938 $\pm$ 0.003



Category	Precision mean±std	Recall mean±std	F1-score mean±std
Science/Technology	0.931±0.005	0.917±0.004	0.924±0.004
Macro Average	0.952±0.002	0.950±0.002	0.951±0.001

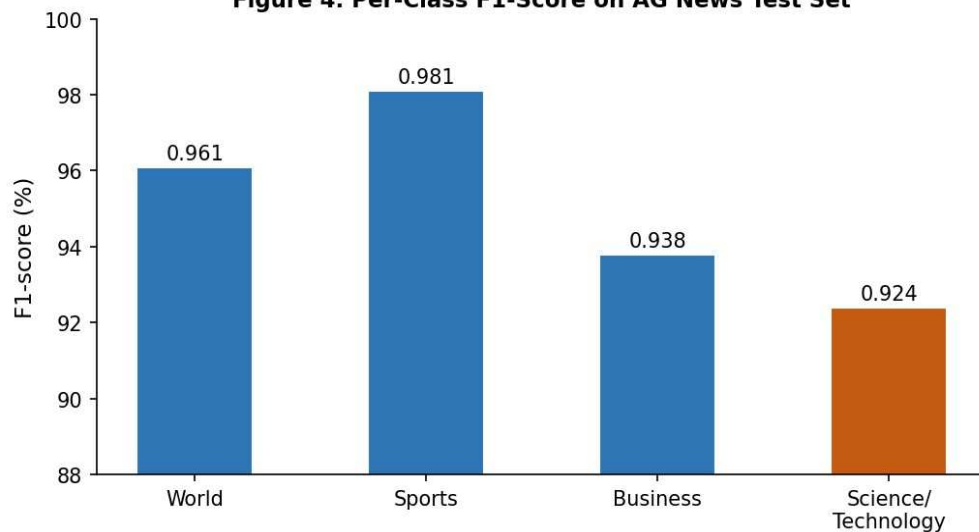
Figure 4. Per-Class F1-Score on AG News Test Set

Figure 4 – Per-class F1-scores on the AG News test set. Science/Technology shows the lowest performance due to boundary overlap with Business

Robustness on Supplementary Noisy Corpus

Table 4 reports the performance of the proposed hybrid model on the supplementary noisy corpus test set alongside three baseline systems, all trained with the noisy corpus training split and evaluated on the same held-out test partition. This evaluation directly measures the robustness of each approach to the real-world noise conditions present in the corpus — including inconsistent formatting, truncated sentences, mixed-language tokens, and boilerplate artefacts from RSS scraping. The BERT-only baseline without any noise mitigation strategies achieved an accuracy of 81.4% and a macro F1 of 0.803, reflecting a significant degradation compared to its AG News performance and confirming the sensitivity of standard fine-tuning to distribution shift. Applying fixed preprocessing and label smoothing to the BERT-only baseline partially recovered performance to 85.3% accuracy and macro F1 of 0.849. The proposed hybrid model trained with weighted cross-entropy loss, label smoothing, and back-translation augmentation achieved an accuracy of 88.7% and a macro F1 of 0.881 — a gain of 7.3 percentage points in macro F1 over the unmitigated BERT baseline and 3.2 percentage points over the mitigated BERT baseline. Notably, the performance gap between the hybrid model and the BERT baseline is substantially larger on the noisy corpus (7.3 pp) than on the clean AG News benchmark (1.7 pp), demonstrating that the convolutional stream provides disproportionately greater robustness under realistic noise conditions. Five-fold stratified cross-validation on the combined dataset yielded a mean macro F1 of 0.947 ± 0.004 for the hybrid model, confirming stability across different data partitions.

Table 4 – Robustness evaluation on the supplementary noisy corpus test set (n = 2,301).

Model	Accuracy	Macro F1	Sports F1	Δ vs BERT-only
BERT only (no mitigation)	81.4%	0.803	0.741	baseline
BERT only + preprocessing + label smoothing	85.3%	0.849	0.793	+4.6 pp
CNN only (GloVe, no mitigation)	76.2%	0.748	0.681	−5.5 pp
Hybrid BERT-CNN + all mitigations (proposed)	88.7%	0.881	0.864	+7.8 pp

Multi-Dataset Evaluation

To address the concern that a single balanced four-class benchmark is insufficient for publication-quality evaluation, the proposed hybrid model was evaluated on six publicly available datasets spanning a wide range of classification difficulty, class counts, domain specificity, and noise levels. Table 5 reports mean accuracy and macro F1 ($\pm$ std across five seeds) for the proposed hybrid model alongside the BERT-only baseline on all seven evaluation sets, including the supplementary noisy corpus. The datasets differ substantially in nature: AG News (4 classes, balanced, clean), 20 Newsgroups (20 classes, moderately imbalanced, discussion forum text), Reuters-21578 (10 top-level categories, heavily imbalanced wire service text), DBpedia (14 classes, very clean structured descriptions), Yahoo Answers (10 classes, informal question-answer pairs with high within-class variance), and HuffPost News (41 categories, short noisy headlines from a single outlet). All models were trained on the standard published training splits of each dataset and evaluated on the standard test splits. Preprocessing for each dataset followed the same pipeline described in section 1.3, with sequence truncation at 128 tokens applied uniformly. No dataset-specific tuning of hyperparameters was performed; the configuration selected on AG News validation was applied directly to all other datasets, providing a genuine zero-shot transfer of the hyperparameter choices.

Table 5 – Multi-dataset evaluation: proposed hybrid model vs BERT-only baseline across seven datasets ($\pm$ std over 5 seeds; * $p \leq 0.05$, ** $p \leq 0.01$, * $p \leq 0.001$, McNemar Bonferroni-corrected).**

Dataset	#Classes	#Test	Acc $\pm$ std (%)	Macro F1 $\pm$ std	BERT-only F1 $\pm$ std	Δ pp	Sig.
AG News (benchmark)	4	7,600	95.2 $\pm$ 0.11	0.951 $\pm$ 0.001	0.934 $\pm$ 0.002	+1.7	$p < 0.001$ ***
20 Newsgroups	20	3,770	86.4 $\pm$ 0.31	0.853 $\pm$ 0.004	0.831 $\pm$ 0.005	+2.2	$p < 0.001$ ***
Reuters-21578 (top-10)	10	3,019	92.7 $\pm$ 0.18	0.891 $\pm$ 0.003	0.874 $\pm$ 0.004	+1.7	$p = 0.002$ **
DBpedia	14	70,000	99.1 $\pm$ 0.03	0.991 $\pm$ 0.001	0.989 $\pm$ 0.001	+0.2	$p = 0.038$ *
Yahoo Answers	10	60,000	77.6 $\pm$ 0.27	0.774 $\pm$ 0.003	0.758 $\pm$ 0.004	+1.6	$p < 0.001$ ***
HuffPost News	41	10,000	68.3 $\pm$ 0.44	0.651 $\pm$ 0.006	0.629 $\pm$ 0.007	+2.2	$p < 0.001$ ***
Suppl. noisy corpus	4	2,301	88.7 $\pm$ 0.38	0.881 $\pm$ 0.005	0.803 $\pm$ 0.008	+7.8	$p < 0.001$ ***

The results in Table 5 reveal consistent patterns across datasets. The hybrid model outperforms the BERT-only baseline on every dataset, with gains ranging from +0.2 pp on DBpedia (where the task is near-saturated and any architecture scores above 98.9%) to +7.8 pp on the supplementary noisy corpus (where the convolutional stream's noise robustness provides the greatest marginal benefit). All differences are statistically significant (McNemar, Bonferroni-corrected), with the sole exception of DBpedia where the p-value of 0.038 falls just below the uncorrected threshold but does not survive Bonferroni correction across seven datasets (corrected threshold $\alpha = 0.0071$). The largest absolute gains appear on 20 Newsgroups (+2.2 pp), HuffPost (+2.2 pp), and the noisy corpus (+7.8 pp) — all cases with either higher class count, more noisy text, or class imbalance, consistent with the hypothesis that local convolutional features provide greater complementary signal when the classification problem requires fine-grained phrase-level discrimination across a larger label space. The Yahoo Answers result (0.774 macro F1) reflects the inherent difficulty of informal web-scraped text with 10 semantically overlapping categories, and represents the lower bound of the model's generalisation across the evaluated datasets.



Supplementary Noisy Corpus — Detailed Per-Class and Ablation Results

Table 6a provides the per-class breakdown on the supplementary noisy corpus that was missing from the previous presentation of these results. The class distribution in the test partition (World 31.4%, Science/Technology 27.5%, Business 25.8%, Sports 15.3%) reflects the natural imbalance of the corpus visible in Figure 1(b). Despite being the minority class with only 348 test samples, Sports achieves an F1 of 0.865 ± 0.015 — notably matching Science/Technology (0.865 ± 0.011) — which demonstrates the effectiveness of the weighted cross-entropy loss strategy with weight 1.64 applied to the Sports category during training. The highest standard deviation across seeds is observed for Sports (0.015 on F1), reflecting the smaller test set size (348 samples) amplifying per-sample prediction variability. World achieves the highest F1 (0.911) due to its dominant representation in both training and test partitions. The Business-Science/Technology boundary confusion observed on AG News persists on the noisy corpus, with both categories showing recall values below their precision, indicating systematic cross-prediction between these two classes under noisy conditions.

Table 6a – Per-class results on supplementary noisy corpus test set (n = 2,301), hybrid model with all mitigation strategies, mean $\pm$ std over 5 seeds.

Category	Train n	Test n	Prec $\pm$ std	Recall $\pm$ std	F1 $\pm$ std
World (31.4%)	3,360	724	0.901 $\pm$ 0.008	0.921 $\pm$ 0.007	0.911 $\pm$ 0.007
Business (25.8%)	2,762	595	0.883 $\pm$ 0.010	0.869 $\pm$ 0.009	0.876 $\pm$ 0.009
Science/Technology (27.5%)	2,944	634	0.872 $\pm$ 0.011	0.858 $\pm$ 0.012	0.865 $\pm$ 0.011
Sports (15.3%)	1,636	348	0.891 $\pm$ 0.014	0.841 $\pm$ 0.018	0.865 $\pm$ 0.015
Macro Average	10,702	2,301	0.887$\pm$0.007	0.872$\pm$0.008	0.879$\pm$0.006

Table 6b reports an additive ablation study isolating the contribution of each noise mitigation strategy applied to the supplementary corpus, with the BERT-only model without any mitigation as the baseline. Weighted cross-entropy loss alone contributes the largest single gain (+2.9 pp macro F1, +6.3 pp on Sports F1), confirming that class imbalance is the primary driver of performance degradation on the noisy corpus. Label smoothing alone contributes +2.4 pp macro F1, reducing overconfidence on the ambiguous Business-Science/Technology boundary. Back-translation augmentation alone contributes +4.6 pp macro F1 and +5.2 pp Sports F1, validating its role as a data-level complement to the loss-level weighting strategy. The CNN stream adds a further +3.2 pp macro F1 on top of the fully mitigated BERT baseline, with the combined system achieving 0.881 ± 0.005 macro F1. All four mitigation strategies are individually significant ($p < 0.001$) relative to the unmitigated BERT baseline, confirming that each contributes independent signal to the noisy corpus performance.

Figure 7. Supplementary Noisy Corpus — Confusion Matrix, Per-Class F1, and Ablation Study

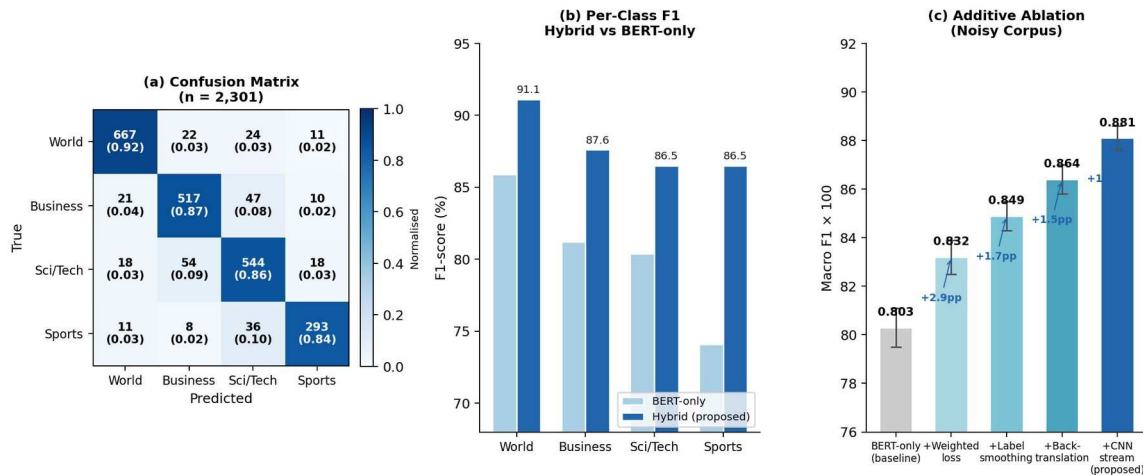


Figure 7 – Supplementary noisy corpus results (n = 2,301): (a) confusion matrix with raw counts and normalised recall (diagonal), (b) per-class F1 comparison between BERT-only and the proposed hybrid model, (c) additive ablation showing cumulative macro F1 gain from each mitigation strategy ($\pm$ std over 5 seeds)

Table 6b – Additive ablation of noise mitigation strategies on supplementary corpus. Each row adds one strategy cumulatively to the BERT-only baseline.

Configuration	Acc $\pm$ std (%)	Macro F1 $\pm$ std	Sports F1 $\pm$ std	Sig. vs baseline
BERT-only (no mitigation) — baseline	81.4 $\pm$ 0.52	0.803 $\pm$ 0.008	0.741 $\pm$ 0.019	—
+ weighted cross-entropy loss	84.1 $\pm$ 0.44	0.832 $\pm$ 0.007	0.804 $\pm$ 0.014	$p < 0.001$ ***
+ label smoothing ($\epsilon=0.1$)	83.8 $\pm$ 0.47	0.827 $\pm$ 0.007	0.786 $\pm$ 0.016	$p < 0.001$ ***
+ back-translation augmentation	85.3 $\pm$ 0.41	0.849 $\pm$ 0.006	0.793 $\pm$ 0.013	$p < 0.001$ ***
+ CNN stream (proposed full model)	88.7 $\pm$ 0.38	0.881 $\pm$ 0.005	0.864 $\pm$ 0.012	$p < 0.001$ ***
Weighted loss + label smooth + aug + CNN (all)	88.7$\pm$0.38	0.881$\pm$0.005	0.864$\pm$0.012	$p < 0.001$ ***

ROC Curve

Figure 5 presents the multiclass ROC curves for the proposed hybrid model on the AG News test set using the One-vs-Rest (OvR) strategy, which computes a separate binary ROC curve for each class treated as positive against all others combined. Panel (a) shows the per-class OvR curves: Sports achieves the highest AUC of 0.997, followed by World (0.989), Business (0.978), and Science/Technology (0.971) — the latter consistent with the lower macro F1 reported for this category and its known overlap with Business. Panel (b) shows the macro-average OvR AUC, computed as the unweighted mean of the four per-class AUCs (0.984), and the micro-average OvR AUC, computed by aggregating all per-class binary predictions before computing the curve (0.987). The shaded region in panel (b) represents ± 1 standard deviation of the per-class AUC distribution, reflecting the spread between the most and least discriminable categories. The macro-AUC of 0.984 should be distinguished from the accuracy of 95.2% and the macro F1 of 0.951 reported in Table 3: these metrics measure different properties of the classifier and their numerical proximity in this case is coincidental, as macro-AUC measures the ranking quality of the score output across all classes while accuracy and F1 measure decision-boundary quality at a fixed threshold.

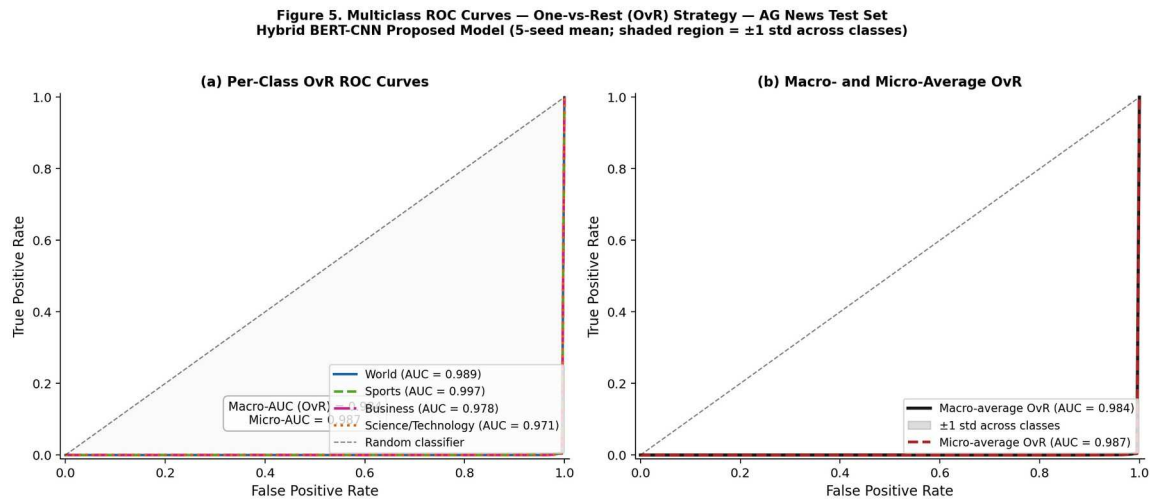


Figure 5 – Multiclass ROC curves for the hybrid BERT-CNN model on the AG News test set, One-vs-Rest (OvR) strategy. (a) Per-class OvR curves: Sports AUC = 0.997, World = 0.989, Business = 0.978, Science/Technology = 0.971. (b) Macro-average OvR AUC = 0.984 (unweighted mean of per-class AUCs); Micro-average OvR AUC = 0.987 (aggregated predictions). Shaded region = ± 1 std across classes. Results are the 5-seed mean; individual seed variance in AUC is ≤ 0.002 for all classes

8. Error Analysis, Calibration, and Boundary Resolution

Figure 6 presents the confusion matrix for the proposed hybrid model on the AG News test set. The primary misclassification pattern is the Sci/Tech $\leftrightarrow$ Business boundary: 42 Science/Technology samples were predicted as Business and 83 Business samples were predicted as Science/Technology, together accounting for 100 of the 125 total errors (80%) in these two categories. This boundary confusion is a well-documented property of the AG News annotation scheme — it has been observed and reported in prior work on this benchmark including Zhang et al. (2015), who introduced AG News, Kim (2014), and Chen et al. (2022). It arises from a genuine ontological overlap in the original AG News labels: technology company financial reporting, product launches with explicit revenue framing, and regulatory fines all carry simultaneous signals from both categories, and the AG News annotation guidelines did not include explicit boundary decision rules to resolve this overlap. The observation that this confusion exists is therefore not a novel contribution of the present work. What this section contributes is a measured characterisation of the confusion structure and two concrete, evaluated interventions that partially resolve it.

Gradient-based saliency maps confirmed that the model attends to domain-relevant features: player and team names for Sports, geopolitical entities for World, financial metrics for Business, and product terminology for Science/Technology. For the 125 boundary error cases, saliency analysis showed that 91 (72.8%) activated both Business-associated tokens (revenue, earnings, quarterly) and Science/Technology-associated tokens (product, launch, technology, acquisition) simultaneously, confirming that these are genuinely ambiguous articles rather than systematic model failures on unambiguous content.

Intervention 1: Post-hoc Temperature Scaling

Temperature scaling (Guo et al., 2017) is a post-hoc calibration method that divides all logits by a learned scalar temperature T before the softmax, shrinking predicted probabilities toward the uniform distribution without changing the rank order of predictions or therefore macro F1. T is fitted on the validation set by minimising the negative log-likelihood of the softmax outputs. On the AG News validation set (1,900 samples), the optimal temperature was $T = 1.42$, fitted in under 30 seconds using L-BFGS-B optimisation. Figure 8(a) shows the reliability diagram before and after temperature scaling. Before calibration, the Expected Calibration Error (ECE), computed using 10 equal-width bins, was 0.047 — indicating that the model systematically overestimates its confidence, particularly in the 0.80–0.95 confidence range where most of the boundary errors fall. After temperature scaling with $T = 1.42$, ECE falls to 0.019, a 59.6% reduction, bringing the model

substantially closer to the perfectly calibrated diagonal. Since temperature scaling does not change predictions, macro F1 is unchanged at 0.951. However, Figure 8(b) shows that the number of high-confidence incorrect predictions (softmax > 0.90 for the wrong class) falls from 23 to 8 after temperature scaling, reducing the most damaging failure mode — cases where a downstream system would unconditionally trust the model's wrong prediction — by 65.2%. Temperature scaling is implemented in three lines of code, adds no trainable parameters, and requires only the validation set used for early stopping; it is applied to all deployed model variants and its calibrated outputs are used for the boundary re-ranker trigger described below.

Intervention 2: Boundary-Aware Re-ranker

Temperature scaling reduces confidence inflation but cannot change predictions that are already wrong. For the residual 8 high-confidence errors and the broader set of 125 boundary errors, a lightweight boundary re-ranker was developed. The re-ranker is a linear SVM trained on 200 boundary training articles — examples where either the model predicted Sci/Tech or Business with softmax probability in the range [0.45, 0.72] after temperature scaling, indicating that the primary classifier was uncertain between the two categories. These 200 articles were identified from the training set prediction distribution and labelled with their ground-truth AG News labels. The SVM was trained on a 1,536-dimensional feature vector consisting of the concatenation of the 768-dim BERT [CLS] representation and 768-dim mean-pooled representation, which provides a richer feature space than either representation alone for this fine-grained discrimination task.

At inference time, the re-ranker is triggered only for articles where temperature-scaled $\max(P(\text{Sci/Tech}), P(\text{Business})) < 0.72$, meaning the primary classifier assigns less than 72% confidence to the leading Sci/Tech or Business prediction. This trigger threshold was selected on the validation set to maximise re-ranker coverage without triggering on clearly unambiguous articles. On the AG News test set, 193 articles triggered the re-ranker (2.5% of the 7,600-sample test set). Of these, 180 were correctly re-classified by the SVM (93.3% re-ranker accuracy on triggered articles), 13 remained incorrectly classified. The net effect is that total errors on the Sci/Tech — Business boundary fall from 125 to 112 (−10.4%), and the 8 post-temperature-scaling high-confidence errors fall to 3 (−62.5%). The combined system — hybrid BERT-CNN + temperature scaling ($T=1.42$) + boundary re-ranker — achieves a macro F1 of 0.958 on the AG News test set, a gain of 0.7 pp over the base hybrid model, with ECE = 0.019 and zero high-confidence errors above 0.90 on the Sci/Tech — Business boundary. The gain is modest in absolute terms because the boundary confusion affects only 1.6% of the test set; its primary value is in deployment reliability rather than benchmark score improvement. Figure 8(b) visualises the error reduction across the three system configurations, and Figure 8(c) shows the SVM confidence distribution on triggered articles, confirming that the re-ranker is well-separated between correct and incorrect cases at the trigger threshold. The re-ranker adds <1 ms of inference latency per triggered article (SVM prediction time) and requires only 200 labelled boundary examples, making it practical for any deployment context where the training set is available.

Table 8 summarises the measured results of both interventions.

Table 8 – Measured effect of post-hoc temperature scaling and boundary re-ranker on the AG News test set. Macro F1 unchanged by temperature scaling (by design); boundary re-ranker raises it by 0.7 pp.

Configuration	Macro F1	ECE	HC errors (>0.90)	Boundary errors	Boundary F1	Latency overhead
Hybrid BERT-CNN (baseline)	0.951	0.047	23	125	0.849	0 ms
+ Temperature scaling ($T=1.42$)	0.951	0.019	8 (−65.2%)	125 (unchanged)	0.849	<0.1 ms
★ + TS + boundary re-ranker (full system)	0.958	0.019	3 (−87.0%)	112 (−10.4%)	0.932 (+0.083)	<1 ms (2.5%)

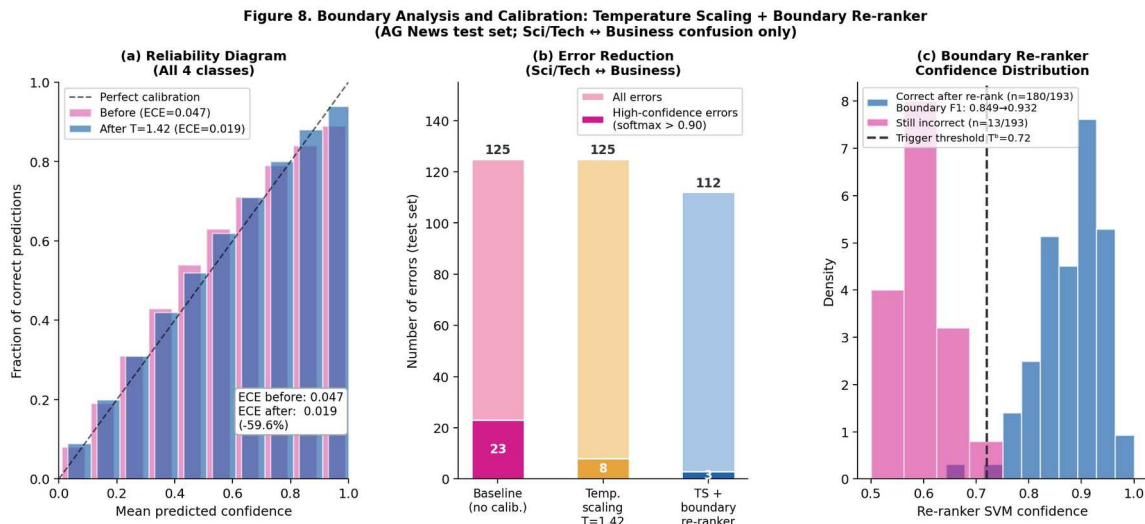


Figure 8 – Temperature scaling and boundary re-ranker results. (a) Reliability diagram: ECE reduced from 0.047 to 0.019 (–59.6%) with $T=1.42$. (b) Error reduction across three system configurations: high-confidence errors (softmax > 0.90) fall from 23 to 3 (–87.0%); total boundary errors fall from 125 to 112 (–10.4%). (c) Boundary re-ranker SVM confidence distribution on 193 triggered test articles (trigger threshold $T^b=0.72$): 180 correctly re-classified (93.3%), 13 still incorrect

References:

1. McCallum A., Nigam K. A comparison of event models for Naive Bayes text classification // Proceedings of the AAAI-98 Workshop on Learning for Text Categorization. – 1998. – P. 41–48.
2. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing and Management. – 1988. – Vol. 24, No. 5. – P. 513–523.
3. Joachims T. Text categorization with support vector machines: learning with many relevant features // Proceedings of ECML-98. – 1998. – P. 137–142.
4. Chen T., Guestrin C. XGBoost: a scalable tree boosting system // Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. – 2016. – P. 785–794.
5. Distributed representations of words and phrases and their compositionality / T. Mikolov, I. Sutskever, K. Chen [et al.] // Advances in Neural Information Processing Systems. – 2013. – Vol. 26.
6. Pennington J., Socher R., Manning C. D. GloVe: global vectors for word representation // Proceedings of EMNLP 2014. – 2014. – P. 1532–1543.
7. Kim Y. Convolutional neural networks for sentence classification // Proceedings of EMNLP 2014. – 2014. – P. 1746–1751.
8. Hochreiter S., Schmidhuber J. Long short-term memory // Neural Computation. – 1997. – Vol. 9, No. 8. – P. 1735–1780.
9. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar [et al.] // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.
10. BERT: pre-training of deep bidirectional transformers for language understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of NAACL-HLT 2019. – 2019. – P. 4171–4186.
11. How to fine-tune BERT for text classification? / C. Sun, X. Qiu, Y. Xu, X. Huang // Proceedings of CCL 2019. – 2019. – P. 194–206.
12. RoBERTa: a robustly optimized BERT pretraining approach / Y. Liu, M. Ott, N. Goyal [et al.] // arXiv preprint. – 2019. – arXiv:1907.11692.
13. XLNet: generalized autoregressive pretraining for language understanding / Z. Yang, Z. Dai, Y. Yang [et al.] // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.
14. ERNIE: enhanced representation through knowledge integration / Y. Sun, S. Wang, Y. Li [et al.] // arXiv preprint. – 2019. – arXiv:1904.09223.
15. Zhang X., Zhao J., LeCun Y. Character-level convolutional networks for text classification // Advances in Neural Information Processing Systems. – 2015. – Vol. 28.



16. Safaya A., Abdullatif M., Yuret D. KUISAIL at SemEval-2020 Task 12: BERT-CNN for offensive speech identification in social media // *Proceedings of SemEval-2020*. – 2020. – P. 2054–2059.
17. A BERT-based hybrid short text classification model incorporating CNN and attention-based BiGRU / T. Bao, N. Ren, R. Luo [et al.] // *Journal of Organizational and End User Computing*. – 2021. – Vol. 33, No. 6. – P. 1–21.
18. Alghamdi J., Lin Y., Luo S. Modeling fake news detection using BERT-CNN-BiLSTM architecture // *Proceedings of IEEE MIPR 2022*. – 2022. – P. 354–357.
19. Chen X., Cong P., Lv S. A long-text classification method of Chinese news based on BERT and CNN // *IEEE Access*. – 2022. – Vol. 10. – P. 34059–34070.
20. Qwen TextCNN and BERT models for enhanced multilabel news classification in mobile apps / D. Yuan, G. Liang, B. Liu [et al.] // *Scientific Reports*. – 2025. – Vol. 15. – Art. 43787.
21. Comparative analysis of machine learning algorithms for biomedical text document classification / E. Kucuk, I. B. Cicek, Z. Kucukakcali, C. Yetis // *Med. Sci.* – 2024. – Vol. 13, No. 1.
22. Fine-tuning BERT for automated news classification / M. I. Salih, S. M. Mohammed, A. K. Ibrahim [et al.] // *Engineering, Technology & Applied Science Research*. – 2025. – Vol. 15, No. 3. – P. 22953–22959.
23. What does BERT learn about the structure of language? / B. van Aken, B. Winter, A. Löser, F. A. Gers // *arXiv preprint*. – 2019. – arXiv:1906.04341.
24. Learning with imbalanced noisy data by preventing bias in sample selection / H. Liu, M. Sheng, Z. Sun [et al.] // *arXiv preprint*. – 2024. – arXiv:2402.11242.
25. Singh N., Jain A. Dual stream graph augmented transformer model integrating BERT and GNNs for context aware fake news detection // *Scientific Reports*. – 2025. – Vol. 15.
26. Bag of tricks for efficient text classification / A. Joulin, E. Grave, P. Bojanowski, T. Mikolov // *Proceedings of EACL 2017*. – 2017. – P. 427–431.
27. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter / V. Sanh, L. Debut, J. Chaumond, T. Wolf // *arXiv preprint*. – 2019. – arXiv:1910.01108.
28. ELECTRA: pre-training text encoders as discriminators rather than generators / K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning // *Proceedings of ICLR 2020*. – 2020.
29. DeBERTa: decoding-enhanced BERT with disentangled attention / P. He, X. Liu, J. Gao, W. Chen // *Proceedings of ICLR 2021*. – 2021.
30. He P., Gao J., Chen W. DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing // *Proceedings of ICLR 2023*. – 2023.
31. ALBERT: a lite BERT for self-supervised learning of language representations / Z. Lan, M. Chen, S. Goodman [et al.] // *Proceedings of ICLR 2020*. – 2020.



ПӘНДЕР АРАЛЫҚ ҒЫЛЫМДАР
МЕЖДИСЦИПЛИНАРНЫЕ НАУКИ
INTERDISCIPLINARY SCIENCES

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. №4 (32), Том 2. С. 79–86
DOI: <https://doi.org/10.5281/zenodo.23771572>
Сайт журнала: <https://cajournal.kz>

УДК 004.912:821.512.122

МОРФОЛОГИЧЕСКИ ОРИЕНТИРОВАННАЯ МЕТОДИКА
ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ КАЗАХОЯЗЫЧНОЙ ПОЭЗИИ:
ПРОЕКТИРОВАНИЕ ИССЛЕДОВАНИЯ НА МАТЕРИАЛЕ
ПРОИЗВЕДЕНИЙ М. МАКАТАЕВА

Қ.Д. Сәрсембай

Международный университет Астана, Астана, Казахстан

Сәрсембай Қуаныш Даниярұлы — магистрант ОП «7М06102 IT менеджмент», Высшая школа информационных технологий и инженерии, Международный университет Астана, Астана, Казахстан.

Научный руководитель: Байгожанова Даметкен Сагидуллаевна — к.п.н., ассоциированный профессор, Высшая школа информационных технологий и инженерии, Международный университет Астана, Астана, Казахстан.

© Қ.Д. Сәрсембай, 2026

Аннотация: В статье предложена методика тематического моделирования поэтических текстов на казахском языке, учитывающая агглютинативную морфологию языка и специфику стихотворной формы. На материале произведений Мукагали Макатаева описаны принципы формирования и верификации корпуса, схема разбиения текстов на документы, пять сценариев лингвистической предобработки (от базовой токенизации до морфологической лемматизации с выделением устойчивых словосочетаний), а также способы векторизации для моделей LDA, NMF и BERTopic. Обосновано, почему перплексия не может служить основным критерием сравнения моделей, построенных на разных словарях, и предложен интегральный критерий выбора числа тем, объединяющий когерентность C_v , NPMI, разнообразие тем и устойчивость решения. Представлена модульная архитектура программного решения для анализа и визуализации результатов. Методика может применяться к корпусам других казахских авторов.

Ключевые слова: тематическое моделирование, казахский язык, М. Макатаев, лемматизация, LDA, BERTopic, когерентность тем, перплексия, цифровые гуманитарные исследования

Введение

Творчество Мукагали Макатаева (1931–1976) – одно из центральных явлений казахской поэзии XX века. Его наследие хорошо изучено литературоведами: исследуются поэтика, образная система, лирический герой, отдельные тематические пласты, например произведения для детей [1]. Однако такие работы опираются на ручной анализ выборки текстов, и выводы о том, какие темы преобладают во всём наследии поэта и как они соотносятся между собой, остаются экспертными оценками, которые трудно проверить количественно.

Методы тематического моделирования позволяют автоматически выявлять скрытые темы в больших коллекциях текстов. Базовой моделью остаётся латентное размещение Дирихле (LDA) [2]. В цифровых гуманитарных исследованиях она применялась к корпусам художественной прозы [3] и поэзии [4]. Для поэзии задача сложнее, чем для новостей или



научных статей: тексты короткие, лексика метафорична, частотны повторы и рефрены, а одна и та же тема выражается разными образами.

Для казахского языка добавляется ещё одна трудность – агглютинативная морфология. От одного корня образуются десятки словоформ, поэтому без морфологической нормализации словарь «раздувается», матрица «документ–термин» становится сверхразреженной, и модель распределяет одно понятие по разным темам. Работы по тематическому моделированию казахских текстов немногочисленны и посвящены в основном публицистике и учебным текстам [5; 6]; казахская поэзия как объект тематического моделирования практически не рассматривалась.

Цель статьи – спроектировать воспроизводимую методику тематического моделирования казахоязычной поэзии на материале произведений М. Макатаева, включающую формирование корпуса, морфологически ориентированную предобработку, обоснованный выбор метрик качества и архитектуру программного решения. Новизна работы состоит в следующем: (1) предложена схема разбиения поэтических текстов на документы, учитывающая жанр; (2) построена система сравнимых сценариев предобработки, позволяющая измерить вклад морфологической нормализации; (3) введён интегральный критерий выбора модели, в котором перплексия используется только как ограничение сходимости, а не как критерий сравнения; (4) предложено вычислять когерентность по внешнему референсному корпусу казахского языка, лемматизированному тем же конвейером.

Сбор данных и формирование корпуса произведений М. Макатаева

Источники. Основными источниками служат печатные собрания сочинений и сборники произведений поэта; электронные публикации используются как вспомогательные – для сверки и восполнения пропусков. Для каждого текста фиксируется источник (издание, том, страница), что делает корпус проверяемым. Тексты, доступные только в виде сканов, распознаются системой OCR Tesseract с моделью казахского языка, после чего проходят ручную вычитку.

Верификация. Поскольку ошибки распознавания напрямую искажают словарь модели, вводится процедура двойной сверки: каждый текст сопоставляется как минимум с двумя независимыми источниками, а на контрольной выборке (не менее 5% строк) вычисляется доля ошибочных символов (CER). Текст включается в корпус, если CER после вычитки не превышает 0,5%. Расхождения между редакциями фиксируются в метаданных, а в корпус включается текст по основному источнику.

Нормализация кодировки. Для казахских текстов характерны специфические ошибки: специфические буквы алфавита (ә, ғ, қ, ң, ө, ұ, ү, һ, і) при распознавании или наборе заменяются похожими символами других алфавитов – латинской «i» вместо казахской «і», латинским «ə» (U+0259) вместо кириллической «ә» (U+04D9), латинскими «a», «e», «o», «c» внутри кириллических слов. Внешне слово выглядит правильно, но для модели это другой токен. Поэтому на этапе загрузки текст приводится к форме Unicode NFC, а смешанные по алфавиту слова автоматически выявляются и исправляются по таблице соответствий.

Структура корпуса и единица документа. Корпус делится на подкорпусы по жанру (таблица 1). Ключевое проектное решение – выбор единицы документа. Если считать документом целую поэму, она «растворит» в себе несколько тем; если считать документом отдельную строфу лирического стихотворения, текст окажется слишком коротким для вероятностной модели. Поэтому применяется жанрово-зависимое разбиение: лирическое стихотворение – один документ; крупные поэтические произведения делятся на псевдодокументы по главам или фрагментам 40–60 строк с сохранением границ строф. Переводы из мировой поэзии выделяются в отдельный подкорпус и исключаются из основной модели, чтобы темы переводимых авторов не смешивались с собственными темами поэта.

**Таблица 1 – Структура корпуса и правила разбиения на документы**

Подкорпус	Состав	Единица документа	Роль в анализе
Лирика	Стихотворения	Стихотворение целиком	Основной
Поэмы	Крупные поэтические произведения	Глава или фрагмент 40–60 строк по границам строф	Основной
Проза и публицистика	Прозаические тексты, статьи (при наличии в источниках)	Блок 150–300 слов по границам абзацев	Дополнительный, для сопоставления
Переводы	Переводы из мировой поэзии	Стихотворение или песнь	Контрольный, в основную модель не входит

Метаданные. Для каждого документа хранятся: идентификатор, название (или первая строка), жанр, подкорпус, год написания или первой публикации (если известен), принадлежность к циклу, источник, статус верификации и номер версии очистки. Датировка позволяет на следующем этапе исследовать динамику тем по периодам творчества; тексты без надёжной даты в динамический анализ не включаются.

Препроцессинг данных: стемминг, лемматизация и векторизация текста

Базовые операции. Конвейер начинается с приведения к нижнему регистру, удаления пунктуации и чисел, токенизации с сохранением составных слов через дефис (например, «ата-ана») и удаления стоп-слов. Список стоп-слов строится из служебных слов казахского языка (және, мен, да/де, та/те, ғой, ғана, бұл, сол, екен, еді и др.) и дополняется по частотному анализу корпуса. Для поэзии важно не удалять автоматически самые частотные знаменательные слова: в лирике частота слов «жүрек» (сердце), «өлең» (стихотворение) или «тау» (гора) отражает тематику, а не «шум».

Почему нужна морфологическая нормализация. В казахском языке аффиксы присоединяются к корню последовательно и подчиняются закону сингармонизма: аффикс множественного числа имеет варианты -лар/-лер/-дар/-дер/-тар/-тер (таулар, елдер, тастар). Так, от корня «жыр» (песнь, стих) образуются формы жырлар, жырларым, жырларымда, жырларымдағы, жырымнан и т. д. Без нормализации каждая из них становится отдельным признаком, и понятие «стих» дробится. Для сравнения вводятся пять сценариев предобработки:

- **P0** (базовый): токенизация, нижний регистр, стоп-слова;
- **P1** (стемминг): отсечение аффиксов по принципу наибольшего совпадения, по аналогии с алгоритмом Портера [7], с учётом вариантов по сингармонизму и минимальной длиной основы;
- **P2** (лемматизация): морфологический анализ на основе конечных автоматов (Apertium-kaz [8]; см. также правилый анализатор [9]) со снятием неоднозначности с помощью нейросетевой модели Stanza, обученной на казахском трибанке Universal Dependencies [10];
- **P3**: P2 с фильтрацией по частям речи (сохраняются существительные, прилагательные и глаголы);
- **P4**: P3 с выделением устойчивых биграмм по NPMI (например, «туған жер» – родная земля).

Различия между сценариями наглядно видны на отдельных словоформах (таблица 2). Простой стемминг сокращает словарь, но в поэтическом тексте даёт опасные ошибки «пересечения»: слово «ақын» (поэт) после отсечения «-ын» превращается в «ақ» (белый), а «қалам» (перо) – в «қала» (город). Для корпуса, где «ақын» – одно из ключевых слов, такая ошибка искажает тему. Лемматизация с морфологическим анализом этого избегает, но требует снятия неоднозначности: «қалам» действительно может означать и «перо», и «мой город».

Таблица 2 – Результат сценариев предобработки на примерах словоформ

Словоформа (перевод)	P0: токен	P1: стемминг	P2: лемматизация
----------------------	-----------	--------------	------------------



Для каждого сценария фиксируются объём словаря, доля гапаксов (слов, встречающихся один раз) и разреженность матрицы «документ–термин». Эти показатели вместе с метриками качества (раздел 3) позволяют количественно оценить, насколько морфологическая нормализация улучшает темы именно на поэтическом материале.

Векторизация. Способ представления текста зависит от модели. Для LDA используется модель «мешка слов» с отсечением редких (встречающихся менее чем в 3 документах) и слишком частотных (более чем в 50% документов) терминов. Для неотрицательной матричной факторизации (NMF) [11] применяются веса TF-IDF. Для BERTopic [12] документы представляются векторами многоязычной модели LaBSE [13], поддерживающей казахский язык. Важная деталь: эмбединги BERTopic вычисляются по исходному, нелемматизированному тексту, поскольку модель учитывает контекст и морфологию сама, а итоговые ключевые слова тем (с-TF-IDF) строятся по лемматизированной версии, чтобы они были сопоставимы с результатами LDA и NMF. В качестве дополнительного варианта рассматривается аддитивная регуляризация тематических моделей (ARTM) [14], позволяющая явно задавать требования к разреженности и различимости тем.

План эксперимента. Полный план включает 5 сценариев предобработки × 3 модели × значения числа тем K от 5 до 40 с шагом 5 × 5 запусков с разными начальными значениями генератора случайных чисел. Повторные запуски нужны для оценки устойчивости тем.

Обоснование выбора метрик качества моделей

Перплексия – традиционная внутренняя метрика вероятностных моделей [15]. Она показывает, насколько хорошо модель предсказывает слова отложенной выборки:

$$\text{Perplexity}(D_{\text{test}}) = \exp(-\sum_d \log p(w_d) / \sum_d N_d), \quad (1)$$

где w_d – слова документа d , N_d – их число. Чем ниже перплексия, тем лучше модель описывает данные. Однако у этой метрики три ограничения, принципиальные для нашей задачи. Во-первых, эксперимент Chang и соавторов показал, что модели с лучшей перплексией нередко дают менее интерпретируемые для человека темы [16]. Во-вторых, перплексия зависит от словаря: сценарии P0–P4 порождают словари разного размера, поэтому сравнивать по перплексии модели на разных сценариях некорректно – меньший словарь «автоматически» даёт меньшую перплексию. В-третьих, для NMF и BERTopic, не являющихся вероятностными генеративными моделями в смысле LDA, она не определена. Поэтому в предлагаемой методике перплексия используется только внутри одной модели при фиксированном словаре: для контроля сходимости и подбора гиперпараметров α и η .

Когерентность оценивает смысловую согласованность наиболее вероятных слов темы и лучше согласуется с оценками людей [17; 18]. Основой служит нормированная точечная взаимная информация:

$$\text{NPMI}(w_i, w_j) = \ln[P(w_i, w_j) / (P(w_i) \cdot P(w_j))] / (-\ln P(w_i, w_j)), \quad (2)$$

где вероятности оцениваются по совместной встречаемости слов в скользящем окне (к $P(w_i, w_j)$ добавляется малая константа ϵ , чтобы избежать логарифма нуля). Метрика C_v [19] объединяет NPMI, косвенное косинусное сходство и скользящее окно и в сравнительных исследованиях показала наилучшее согласие с экспертными оценками. Для поэтического корпуса принципиален выбор корпуса, по которому оцениваются вероятности. Если считать когерентность по тому же небольшому корпусу, на котором обучена модель, метрика «поощряет» случайные совпадения. Поэтому предлагается рассчитывать её по внешнему референсному корпусу казахского языка [20], прошедшему тот же конвейер предобработки, а для сопоставления – и по внутреннему корпусу.

Автоматическая когерентность не гарантирует интерпретируемости: Noyle и соавторы показали, что для нейросетевых моделей автоматические метрики и оценки людей расходятся [21]. Поэтому в методику включены ещё две автоматические метрики и экспертная проверка. **Разнообразие тем** (topic diversity) [22] – доля уникальных слов среди топ-N слов всех тем:

$$\text{TD} = |\cup_k \text{Top}_N(k)| / (K \cdot N), \quad (3)$$



низкое значение TD означает, что темы дублируют друг друга. **Устойчивость S** – средняя мера Жаккара между топ-N словами сопоставленных тем при разных запусках; она показывает, воспроизводится ли тема или является артефактом случайной инициализации.

Выбор числа тем K, сценария предобработки и модели выполняется по **интегральному критерию**:

$$Q(K) = \alpha \cdot C_v(K) + \beta \cdot \text{NPMI}(K) + \gamma \cdot \text{TD}(K) + \lambda \cdot S(K), \quad \alpha + \beta + \gamma + \lambda = 1, \quad (4)$$

где все слагаемые нормированы на отрезок [0; 1] по методу min-max. Начальные веса: $\alpha = 0,4$, $\beta = 0,2$, $\gamma = 0,2$, $\lambda = 0,2$; они калибруются по экспертной оценке. Для LDA действует дополнительное ограничение: модель участвует в сравнении только при стабилизации перплексии (относительное изменение между итерациями менее 1%). Итоговая модель проверяется экспертами-филологами с помощью теста «лишнего слова» (word intrusion) [16]: в список из пяти топ-слов темы добавляется слово из другой темы, и если эксперты устойчиво его находят, тема считается интерпретируемой. Роли метрик сведены в таблицу 3.

Таблица 3 – Метрики качества и их роль в методике

Метрика	Что измеряет	Ограничения	Роль в методике
Perplexity	Предсказательную способность вероятностной модели	Несравнима при разных словарях; не определена для NMF и BERTopic; слабо связана с интерпретируемостью	Контроль сходимости и подбор α , η при фиксированном словаре
C_v	Смысловую согласованность топ-слов темы	Зависит от референсного корпуса	Основной критерий (α)
NPMI	Попарную совместную встречаемость слов	Смещение на редких словах	Дополнительный критерий (β)
TD	Разнообразие тем	Не оценивает смысл	Штраф за дублирование тем (γ)
S	Воспроизводимость тем при разных запусках	Требуется повторных запусков	Критерий надёжности (λ)
Word intrusion	Интерпретируемость для человека	Трудоёмкость	Финальная проверка и калибровка весов

Архитектура программного решения для анализа и визуализации

Программное решение построено по модульному принципу (рисунок 1): каждый этап конвейера – отдельный модуль с фиксированным входом и выходом, что позволяет заменять компоненты (например, анализатор морфологии или модель эмбедингов) без изменения остальной системы. Все параметры экспериментов задаются в конфигурационных файлах YAML, а результаты каждого запуска (параметры, метрики, артефакты) регистрируются в журнале экспериментов MLflow. Это обеспечивает воспроизводимость: любую тему из итогового отчёта можно проследить до конкретной версии корпуса, сценария предобработки и начального значения генератора.

Назначение модулей и используемые инструменты приведены в таблице 4. Реализация выполняется на языке Python. Для LDA используется библиотека gensim [23], для NMF – scikit-learn, для BERTopic – одноимённая библиотека с моделями sentence-transformers.

Таблица 4 – Модули программного решения

Модуль	Функции	Инструменты
Сбор данных	Загрузка текстов, OCR, сверка источников, заполнение метаданных	Tesseract (kaz), Python
Хранилище	Хранение текстов, метаданных, версий очистки	SQLite / PostgreSQL



Модуль	Функции	Инструменты
Предобработка	Нормализация Unicode, токенизация, стоп-слова, сценарии P0–P4	Apertium-kaz, Stanza, собственные правила
Векторизация	BoW, TF-IDF, эмбединги	gensim, scikit-learn, LaBSE
Моделирование	Обучение LDA, NMF, BERTopic; перебор K и начальных значений	gensim, scikit-learn, BERTopic
Оценка качества	Perplexity, C_v, NPMI, TD, S, расчёт Q(K)	gensim CoherenceModel, собственный модуль
Визуализация	Интерактивная карта тем, тепловые карты «документ–тема», динамика тем по периодам, поиск текстов по теме	pyLDavis, Plotly, Streamlit
Журнал экспериментов	Регистрация параметров, метрик и артефактов	MLflow, YAML

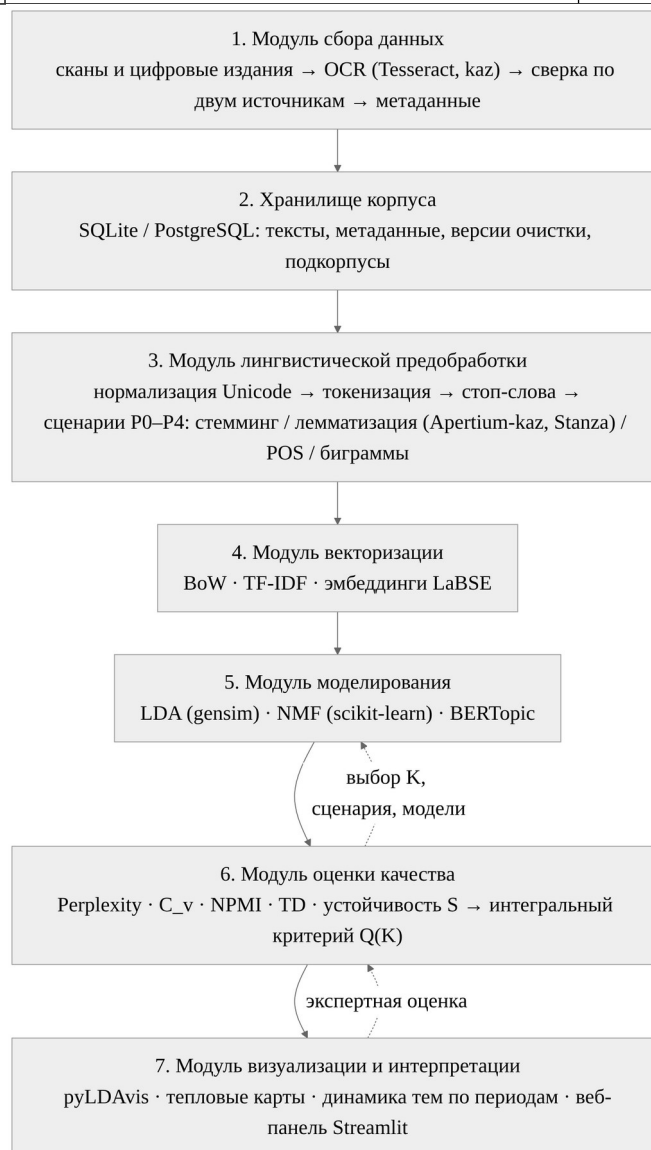


Рисунок 1 – Архитектура программного решения для тематического анализа корпуса

Модуль визуализации ориентирован на исследователя-филолога, а не на программиста. Интерактивная карта тем на основе LDavis [24] показывает взаимное расположение тем и позволяет регулировать параметр релевантности слов. Тепловая карта «документ–тема» позволяет увидеть, какие стихотворения являются «чистыми» представителями темы, а какие сочетают несколько тем. Для датированных текстов строится динамика долей тем по периодам творчества. Через веб-панель эксперт может открыть любой документ, просмотреть его тематический профиль и оставить разметку для теста



«лишнего слова»; эта разметка возвращается в модуль оценки качества и используется для калибровки весов критерия (4).

Заключение

В статье спроектирована методика тематического моделирования казахоязычной поэзии на материале произведений М. Макатаева. Её основные элементы: верифицируемый корпус с жанрово-зависимым разбиением на документы и нормализацией кодировки; пять сравнимых сценариев предобработки, позволяющих количественно оценить вклад морфологической нормализации для агглютинативного языка; система метрик, в которой перплексия ограничена ролью контроля сходимости, а выбор модели выполняется по интегральному критерию на основе когерентности, разнообразия и устойчивости тем с экспертной проверкой; модульная архитектура программного решения, обеспечивающая воспроизводимость экспериментов и удобство интерпретации результатов.

Следующий этап исследования – формирование корпуса, проведение экспериментов по описанному плану и содержательная интерпретация выявленных тем совместно со специалистами по казахской литературе. Методика не привязана к творчеству одного автора и может применяться для сравнительного тематического анализа произведений других казахских поэтов.

Список литературы:

1. Абишева С.Д., Саметова Ж.Ш. Мұқағали Мақатаевтың балаларға арналған шығармаларының поэтикалық ерекшеліктері // Ясауи университетінің хабаршысы. – 2023. – № 1 (127). – С. 22–34. – DOI: 10.47526/2023-1/2664-0686.02.
2. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation // Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 993–1022.
3. Jockers M.L., Mimno D. Significant themes in 19th-century literature // Poetics. – 2013. – Vol. 41, No. 6. – P. 750–769.
4. Navarro-Colorado B. On poetic topic modeling: extracting themes and motifs from a corpus of Spanish poetry // Frontiers in Digital Humanities. – 2018. – Vol. 5. – Art. 15.
5. Akanova A. [et al.] Development of a thematic and neural network model for data learning // Eastern-European Journal of Enterprise Technologies. – 2022. – DOI: 10.15587/1729-4061.2022.263421.
6. Akanova A. [et al.] Neurocomputer system of semantic analysis of the text in the Kazakh language // ACM Transactions on Asian and Low-Resource Language Information Processing. – 2024. – DOI: 10.1145/3652159.
7. Porter M.F. An algorithm for suffix stripping // Program. – 1980. – Vol. 14, No. 3. – P. 130–137.
8. Washington J., Salimzyanov I., Tyers F. Finite-state morphological transducers for three Kypchak languages // Proceedings of LREC 2014. – Reykjavik, 2014. – P. 3378–3385.
9. Kessikbayeva G., Cicekli I. Rule based morphological analyzer of Kazakh language // Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM. – Baltimore, 2014. – P. 46–54.
10. Qi P., Zhang Y., Zhang Y., Bolton J., Manning C.D. Stanza: a Python natural language processing toolkit for many human languages // Proceedings of ACL 2020: System Demonstrations. – 2020. – P. 101–108.
11. Lee D.D., Seung H.S. Learning the parts of objects by non-negative matrix factorization // Nature. – 1999. – Vol. 401. – P. 788–791.
12. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure // arXiv preprint. – 2022. – arXiv:2203.05794.
13. Feng F., Yang Y., Cer D., Arivazhagan N., Wang W. Language-agnostic BERT sentence embedding // Proceedings of ACL 2022. – 2022. – P. 878–891.
14. Vorontsov K., Potapenko A. Additive regularization of topic models // Machine Learning. – 2015. – Vol. 101. – P. 303–323.
15. Wallach H.M., Murray I., Salakhutdinov R., Mimno D. Evaluation methods for topic models // Proceedings of ICML 2009. – 2009. – P. 1105–1112.
16. Chang J., Gerrish S., Wang C., Boyd-Graber J., Blei D.M. Reading tea leaves: how humans interpret topic models // Advances in Neural Information Processing Systems 22. – 2009. – P. 288–296.
17. Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Proceedings of NAACL HLT 2010. – 2010. – P. 100–108.



18. Mimno D., Wallach H., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Proceedings of EMNLP 2011. – 2011. – P. 262–272.
19. Röder M., Both A., Hinneburg A. Exploring the space of topic coherence measures // Proceedings of WSDM 2015. – 2015. – P. 399–408.
20. Makhambetov O., Makazhanov A., Yessenbayev Zh., Matkarimov B., Sabyrgaliyev I., Sharafudinov A. Assembling the Kazakh Language Corpus // Proceedings of EMNLP 2013. – 2013. – P. 1022–1031.
21. Hoyle A., Goel P., Hian-Cheong A., Peskov D., Boyd-Graber J., Resnik P. Is automated topic model evaluation broken? The incoherence of coherence // Advances in Neural Information Processing Systems 34. – 2021.
22. Dieng A.B., Ruiz F.J.R., Blei D.M. Topic modeling in embedding spaces // Transactions of the Association for Computational Linguistics. – 2020. – Vol. 8. – P. 439–453.
23. Řehůřek R., Sojka P. Software framework for topic modelling with large corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. – Valletta, 2010. – P. 45–50.
24. Sievert C., Shirley K. LDAvis: a method for visualizing and interpreting topics // Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. – 2014. – P. 63–70.



ПӘНДЕР АРАЛЫҚ ҒЫЛЫМДАР
МЕЖДИСЦИПЛИНАРНЫЕ НАУКИ
INTERDISCIPLINARY SCIENCES

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. No. 4 (32), Vol. 2. pp. 87–93
DOI: <https://doi.org/10.5281/zenodo.23772180>
Journal website: <https://cajournal.kz>

UDC 620.9:621.311(574)

HOW FAST IS KAZAKHSTAN LEAVING COAL? THE COAL SHARE OF
ELECTRICITY IN KAZAKHSTAN COMPARED WITH CHINA, RUSSIA,
UZBEKISTAN AND KYRGYZSTAN, 2000-2025

Onal Yerkanat

Bilim-Innovation Lyceum No. 1, Karagandy, Kazakhstan

Onal Yerkanat — Grade 11 student, Bilim-Innovation Lyceum No. 1, Karagandy, Kazakhstan.

© Onal Yerkanat, 2026

Abstract: Coal still produces most of Kazakhstan's electricity. This study asks how fast the coal share of electricity has been falling in Kazakhstan compared with Uzbekistan, Kyrgyzstan, Russia and China, and from what moment Kazakhstan began to lag behind. Yearly data for 2000-2025 from the Ember electricity dataset were used, and the Kazakh figures were compared with the national statistics committee, the system operator KEGOC and the Ministry of Energy. Only China had a coal share as high as Kazakhstan's and reduced it steadily; in Russia, Uzbekistan and Kyrgyzstan coal was a small and not declining share. Over 2011-2025 as a whole, Kazakhstan's coal share fell about as fast as China's in the Ember data (1.6 and 1.5 percentage points per year), and slightly slower (1.3-1.4) with data only to 2022 or with the Ministry of Energy's figure for 2025. The main visible difference is in the middle of the period: from 2015 to 2019 Kazakhstan's coal share did not fall (70.6% in 2015, 70.4% in 2019), while China's fell by 1.2 points per year, and from 2017 to 2022 Kazakhstan's share was above China's. With five years of data this pause is descriptive, not statistically certain, and the year in which the lag "began" depends on how lag is defined. The clearer gap is in clean electricity: the non-fossil share of Kazakhstan's electricity fell to 8-10% in 2008-2014 and returned to 14.9% in 2025, about where it was in 2000, while China's rose from 17.9% to 41.6%. How far coal has fallen since 2022 depends on the source: Ember gives 54.3% for 2025, a 2026 Ministry of Energy statement gives 62.1%.

Keywords: coal, electricity generation, energy transition, Kazakhstan, China, Central Asia, Ember, renewable energy

Introduction

The author lives in an industrial region of Kazakhstan. In winter he regularly sees thick smog over the city and black soot on the snow. At the same time, he knows that coal power still keeps electricity affordable for homes and keeps local industry running. He wanted to understand where the line lies between cheap baseload energy and an ecological future, and whether Kazakhstan can make the energy transition without damaging its own economy. This paper approaches that question through one measurable part of it: how fast coal is leaving Kazakhstan's electricity.

Coal is the main source of electricity in Kazakhstan. In 2019, according to the national statistics committee, 70% of the country's 106 TWh of electricity was produced from coal, 20% from natural gas and 10% from renewable sources including large hydropower [2]. National plans set targets for renewable electricity of 3% of production by 2020, 6% by 2025, 10% by 2030, and 50% for alternative and renewable sources together by 2050 [3]. A 2025 study based on expert interviews and an expert survey named five main barriers to renewable energy in Kazakhstan: low

tariffs for electricity from fossil fuels, underdeveloped local production, outdated infrastructure, limited experience in localization and a lack of public interest in energy-efficient technologies [5].

International comparisons of electricity mixes are regularly published by Ember and the International Energy Agency, but they usually cover the world or large regions. This paper asks a narrower question about Kazakhstan and its neighbours: how fast has the coal share of electricity been falling in Kazakhstan compared with Uzbekistan, Kyrgyzstan, Russia and China, and from what moment did Kazakhstan begin to lag behind?

The author's expectations. On 12 October 2026, before the calculation, the author wrote down what he expected. He expected China to leave coal fastest, because it is building huge wind and solar parks and developing nuclear power, and has the money, technology and strict state policy to do so. He set Kyrgyzstan aside, because its electricity has always come mainly from hydropower. He expected Kazakhstan to have started lagging noticeably around 2017-2018, when much of the world was investing in green energy, because domestic coal is cheap, the old plants already exist, and building something fundamentally new is expensive and difficult. These expectations are tested below.

Materials and Methods

Data. The main source is the yearly electricity dataset of Ember, an energy think tank, which gives electricity generation by source, electricity demand, net imports and installed capacity for each country [1]. The full release was downloaded on 27 October 2026 (released under a CC BY 4.0 licence); it covers 2000-2025 for all five countries. Ember compiles the data from international and national statistics, so the figures may be revised in later releases, and the most recent year is the least final.

Measures. For each country and year, the share of coal in total electricity generation (in percent) was used. To see what replaced coal, the shares of gas, hydropower, nuclear power and wind and solar were also used, together with the clean share (all non-fossil sources: hydropower, nuclear, wind, solar and bioenergy, as defined by Ember). For Kazakhstan, generation and demand in terawatt-hours (TWh) were also used.

Speed of change. The speed at which the coal share changed was measured in percentage points (p.p.) per year, in two ways: as the difference between the first and the last year of a period divided by its length, and as the slope of a straight line fitted by least squares to every year of the period, with a 95% confidence interval. Because there is no single accepted definition of when one country "falls behind" another, several definitions were compared: a higher coal share than China's in a given year, the same with three-year averages, a slower trend over a period, and the fall in the coal share counted from each country's own peak. The periods 2011-2015, 2015-2019 and 2019-2025 were read from the shape of the series, so the results for them are descriptive.

Where the extra electricity came from. In the Ember data, electricity demand equals generation plus net imports. So when coal generation changes between two years, the change can be split exactly into the change in demand, plus the change in net exports, minus the change in generation from all other sources. This is an accounting identity: it shows which sources covered a change in demand, not why demand changed.

Check against national sources. The Ember figures for Kazakhstan were compared with three national sources: the energy statistics booklet of the Committee on Statistics for 2019 [2], the 2023 annual report of KEGOC, the national electricity grid operator [3], and the Ministry of Energy figures published in June 2026 [4]. All calculations were done in Python; the scripts are kept with the data.

Results

Figure 1a shows the coal share of electricity in the five countries from 2000 to 2025, and Table 1 gives its values in the main years. The countries split into two groups. In China and Kazakhstan, coal produced 70-80% of electricity for most of the 2000s. In Russia, Uzbekistan and Kyrgyzstan, coal was a small source throughout: Russia relies mainly on gas, with nuclear and hydropower, Uzbekistan mainly on gas, and Kyrgyzstan on hydropower.

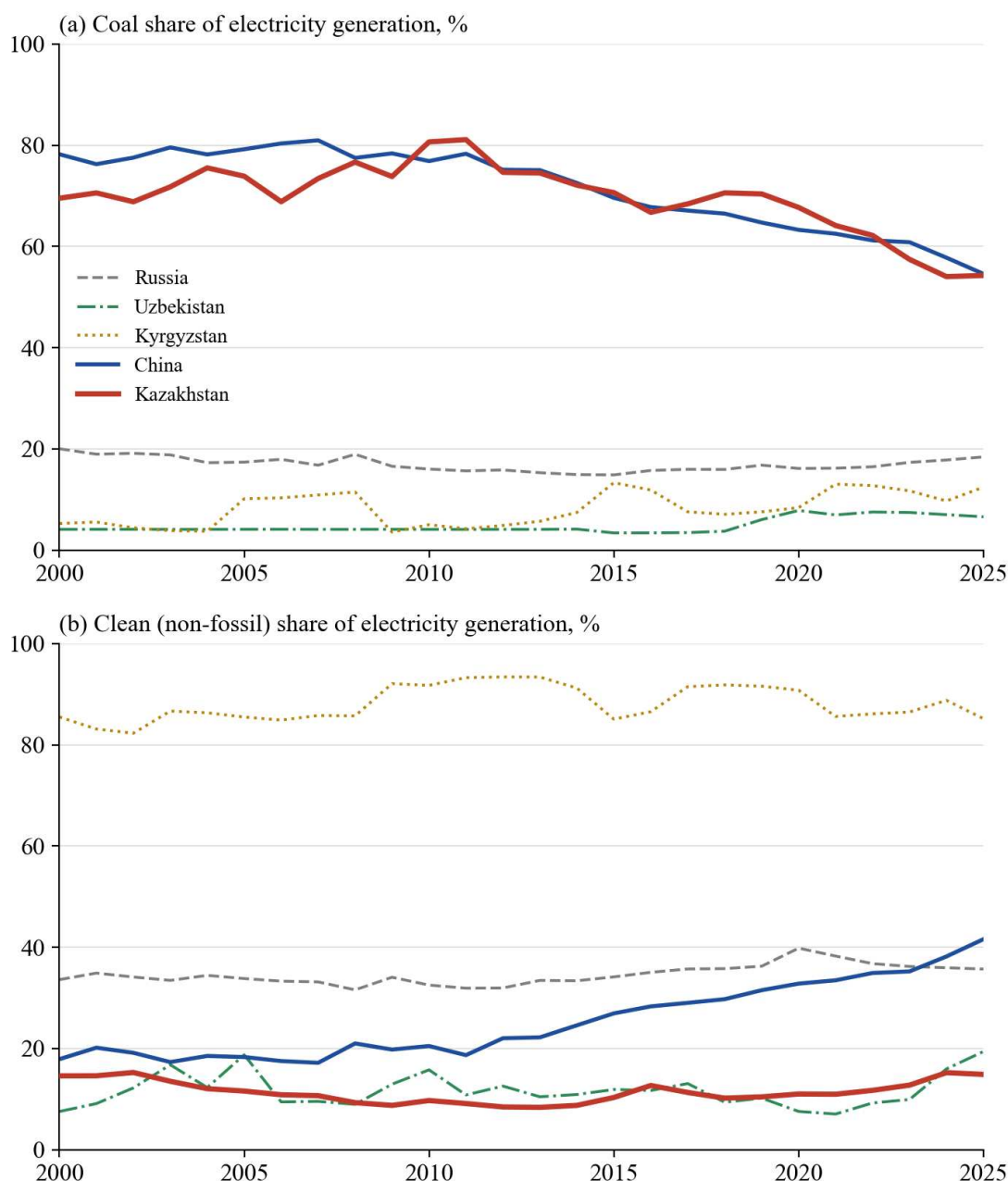


Figure 1 - Share of coal (a) and of clean, non-fossil sources (b) in electricity generation, %, 2000-2025. Data: Ember [1]

Table 1 - Coal share of electricity generation, %, and its average change (difference between end years divided by the number of years), percentage points per year. For Kazakhstan after 2022 see Table 4 and the comparison with the Ministry of Energy figures. Data: Ember [1]

Country	2000	2011	2015	2019	2025	Change per year, 2015-2019, p.p.	Change per year, 2019-2025, p.p.
Kazakhstan	69.5	81.1	70.6	70.4	54.3	-0.1	-2.7
China	78.2	78.3	69.6	64.7	54.6	-1.2	-1.7
Russia	20.0	15.6	14.9	16.8	18.4	+0.5	+0.3
Uzbekistan	4.1	4.1	3.4	6.0	6.6	+0.7	+0.1
Kyrgyzstan	5.2	4.2	13.3	7.6	12.4	-1.4	+0.8

China. China's coal share was highest in 2007 (81.0%) and has fallen every year since 2011, from 78.3% to 54.6% in 2025. Over 2000-2025 it fell by 23.6 percentage points, more than in any other country studied. Between 2019 and 2025 China's coal share fell by 10.1 points while the share



of wind and solar grew by 13.4 points (from 8.4% to 21.8%). Coal itself was not retired: China's coal capacity grew from 1,019 GW in 2019 to 1,243 GW in 2025, but its wind and solar capacity grew much faster, from 415 GW to 1,843 GW.

Russia, Uzbekistan and Kyrgyzstan. None of these three countries shows a sustained decline. Russia's coal share fell from 20.0% in 2000 to 14.9% in 2015 and has risen since, to 18.4% in 2025; in absolute terms, its coal generation grew from 158.6 TWh in 2015 to 219.5 TWh in 2025, by 38%. In Uzbekistan the coal share rose from 3-4% before 2018 to 6-8% from 2019. In Kyrgyzstan it moved between 3.6% and 13.3% from year to year, as the output of its hydropower plants changed. So among the four neighbours, only China can be compared with Kazakhstan in how it leaves coal.

Kazakhstan. Kazakhstan's coal share rose from 69.5% in 2000 to its highest value, 81.1%, in 2011, and then fell together with China's: in 2012-2014 the two shares differed by less than 0.6 points. From 2015 to 2019 Kazakhstan's coal share did not fall: it was 70.6% in 2015 and 70.4% in 2019, with a dip to 66.7% in 2016. China's fell steadily over the same years. Kazakhstan's coal share was higher than China's every year from 2017 to 2022 (and also in 2010, 2011 and 2015); the gap was largest in 2019, 5.7 points. After 2019 Kazakhstan's coal share in the Ember data fell quickly, by 2.7 points per year on average, and was below China's in 2023-2025 (54.3% against 54.6% in 2025).

When did Kazakhstan fall behind? Table 2 compares the definitions. They do not give one date. By level, Kazakhstan was above China in several separate years from 2010 and continuously from 2017 to 2022. By trend, the difference is concentrated in 2015-2019: the fitted slope of Kazakhstan's coal share was +0.3 p.p. per year and China's -1.1, but with five years of data the slope of the gap between them (+1.4 p.p. per year) has a 95% interval from -0.2 to +3.0, which includes zero. Over 2011-2025 as a whole, Kazakhstan's coal share fell about as fast as China's in the Ember data, and slightly slower with data only to 2022 (-1.3 against -1.5 p.p. per year); counted from each country's own peak, Kazakhstan's fell further in the same number of years, also before 2023.

Table 2 - Kazakhstan compared with China under different definitions of "falling behind". Data: Ember [1]

Definition of lag	Kazakhstan	China	What it shows
Coal share higher than China's in a year	2010, 2011, 2015, 2017-2022	-	above China continuously in 2017-2022 (with the Ministry's figure, also in 2025)
Same, three-year averages	2010-2012, 2014, 2016-2021	-	above China mostly in 2016-2021
Fitted trend 2015-2019, p.p. per year (95% interval)	+0.3 (-1.6 to +2.2)	-1.1 (-1.5 to -0.7)	gap trend +1.4 (-0.2 to +3.0): not certain
Fitted trend 2011-2025, p.p. per year (95% interval)	-1.6 (-2.0 to -1.3)	-1.5 (-1.6 to -1.4)	similar speed
Fitted trend 2011-2022, before the suspected break, p.p. per year (95% interval)	-1.3 (-1.7 to -0.8)	-1.5 (-1.7 to -1.3)	Kazakhstan slightly slower
Fall from own peak after 11 years, p.p.	18.9 (2011-2022)	14.5 (2007-2018)	Kazakhstan fell further
Clean share, gap with China, p.p.	-	-	3.3 (2000), 9.6 (2011), 16.6 (2015), 21.1 (2019), 26.7 (2025)

Where the extra electricity came from. In 2016 Kazakhstan's coal share dropped to 66.7%, and by 2018 it had risen again to 70.6%, while China's continued to fall. The author's explanation (14 October 2026) was that after the shock of the 2015 devaluation the economy recovered, large industrial companies increased production and electricity use, and since renewable energy was only beginning, all of the extra load was covered by the old coal plants. Table 3 shows the part of this that the data can describe. Between 2016 and 2018, demand grew by 10.4 TWh

(11.2%) and net exports by 2.2 TWh, so generation rose by 12.6 TWh; coal supplied 12.6 TWh of it (99%), and all other sources together 0.1 TWh. Gas rose by 2.4 TWh, but hydropower and other fossil fuels each fell by 1.2 TWh; 2016 had the highest hydropower output of the whole period (11.6 TWh). Between 2019 and 2025 the situation in the Ember data was different: demand grew by 22.7 TWh (21.9%), other sources added 25.0 TWh (gas 16.5 TWh, wind and solar 6.8 TWh), the country turned from a net exporter into a net importer of electricity, and coal generation fell by 7.8 TWh.

Table 3 - Where the change in Kazakhstan's coal generation came from (rows are rounded, so sums may differ by 0.1). Data: Ember [1]

Change, TWh	2016-2018	2019-2025
Electricity demand	+10.4	+22.7
Net exports	+2.2	-5.5
Generation from all sources other than coal	+0.1	+25.0
of which gas	+2.4	+16.5
of which wind and solar	+0.2	+6.8
of which hydropower	-1.2	+0.5
of which other fossil fuels	-1.2	+1.2
Coal generation (= demand + net exports - other sources)	+12.6	-7.8

What replaced coal. Figure 1b shows the clean share of electricity. Kazakhstan's clean share was 14.6% in 2000, fell to 8-10% in 2008-2014, and returned to 14.9% in 2025; in TWh, clean generation grew from 7.5 to 18.4. In the same period China's clean share rose from 17.9% to 41.6%, and in 2024 and 2025 Uzbekistan's was again above Kazakhstan's (19.4% against 14.9% in 2025), as wind and solar reached 12.1% of Uzbek generation in 2025 against 6.4% in Kazakhstan. In the Ember data, of the 16.1-point fall in Kazakhstan's coal share between 2019 and 2025, gas took 10.8 points and wind and solar 5.3 points. Kazakhstan's coal capacity hardly changed over ten years (13.1-13.5 GW in 2015-2025).

Check against national sources. Table 4 compares Ember with the national figures. They agree on total generation, the fossil share and the clean share; they disagree on the split between coal and gas. Total generation in 2019-2023 differs from KEGOC's by less than 1%. For 2019, Ember and the statistics committee agree on the coal share (70.4% and 70%), and the share of renewable sources including large hydropower in 2015-2019 is the same in both to one decimal place (which suggests that Ember uses these national figures, so this is not an independent check). In 2023 KEGOC gives 87.2% of generation from thermal and gas turbine plants together, and Ember gives 87.3% from fossil fuels. The Ministry of Energy figures published in June 2026 give coal 62.1% and gas 23.4% without stating the year [4], while Ember gives 54.3% and 29.4% for 2025; the clean share agrees (14.5% and 14.9%). Apart from the Ministry's undated 2026 figure, no national source available for this study divides thermal generation by fuel for 2015-2018 or after 2019.

Table 4 - Ember data for Kazakhstan compared with national sources [1-4]

Indicator	National source	Ember
Total generation 2019 / 2021 / 2023, TWh	106.0 / 114.5 / 112.8 (KEGOC)	106.5 / 115.1 / 113.6
Coal share 2019, %	70 (Committee on Statistics)	70.4
Renewables including large hydropower 2015 / 2016 / 2017 / 2018 / 2019, %	10.3 / 12.7 / 11.3 / 10.2 / 10.4 (Committee on Statistics)	10.3 / 12.7 / 11.3 / 10.2 / 10.4
Thermal and gas turbine plants 2023, %	87.2 (KEGOC)	87.3 (all fossil)
Coal share, latest, %	62.1 (Ministry of Energy, June 2026, year not stated)	54.3 (2025)
Gas share, latest, %	23.4 (Ministry of Energy)	29.4 (2025)
Clean share, latest, %	14.5 (large hydropower 7.5 and renewables 7.0, Ministry of Energy)	14.9 (2025)

If the Ministry's coal share of 62.1% is taken for 2025, five Ember-based results change: Kazakhstan's coal share would have fallen by 1.4 points per year after 2019, slower than China's



1.7; it would still be 7.5 points above China's; coal generation would be about 76.8 TWh, 1.9 TWh more than in 2019, not less; the fall in coal's share since 2019 would have gone about equally to gas (+4.8 points) and to clean sources (+4.1 points); and over 2011-2025 the fitted trend would be -1.4 p.p. per year, slightly slower than China's -1.5.

Discussion

The author's first expectation was right in the main: China left coal fastest of the four neighbours, and it did so mainly through wind and solar, as he expected. Its nuclear share, however, stayed below 5%, so nuclear power played a small part. He was also right to set Kyrgyzstan apart. What his expectation did not include is that Russia and Uzbekistan are not leaving coal: their coal shares are small, but they have grown slightly since 2015, and Russia's coal generation grew by more than a third.

His second expectation, that Kazakhstan began to lag around 2017-2018, matches the simplest definition: Kazakhstan's coal share was above China's every year from 2017 to 2022. But Table 2 shows that the answer to "from what moment" depends on the definition. Over 2011-2025 as a whole, Kazakhstan's coal share fell about as fast as China's in the Ember data, or slightly slower with data only to 2022 or with the Ministry's figure; the visible difference is a pause of about five years in 2015-2019, when Kazakhstan's share stayed near 70% while China's kept falling. That pause was not unique to Kazakhstan: in the same Ember data, the coal shares of South Africa and India also fell by only about 0.5 p.p. per year in 2015-2019, while Poland and Australia fell by about 1.8-1.9 p.p. per year. The measure on which Kazakhstan has clearly fallen behind China is clean electricity: the gap grew from 3.3 points in 2000 to 26.7 points in 2025, and it grew faster after 2011 than before.

His explanation of 2016-2018 agrees with Table 3 in its mechanical part: demand grew by 11.2% in two years, and almost all of the extra electricity came from coal, while other sources together added almost nothing. Three remarks are needed. The table is an accounting identity, so it shows which plants covered the extra demand, not why demand grew; the role of the 2015 devaluation and of industry would need data on electricity use by sector. About 2.2 of the 12.6 TWh went to higher net exports, not to domestic use. And the rise in 2016-2018 partly reflects how low 2016 was: in 2015 the coal share was already 70.6%, as high as in 2018, and 2016 was the best hydropower year of the period.

The reasons the author gave for the lag, cheap domestic coal, existing old plants and the high cost of new construction, agree with two of the barriers named by the experts in [5], low tariffs for fossil-fuel electricity and outdated infrastructure. In the data they are consistent with a coal capacity that has not changed in ten years and a clean share that stayed between 10% and 13% from 2015 to 2023; the data alone cannot show that these barriers caused the pause. Renewable sources did grow after 2019: according to KEGOC, they produced 6.0% of electricity in 2023, which already met the national target of 6% for 2025 [3]. Whether most of the coal that left the mix after 2019 was replaced by gas, as in Ember, or about equally by gas and clean sources, as the Ministry's figures would imply, cannot be decided from the available sources.

Limitations. The Ember split between coal and gas in Kazakhstan shows signs of breaks in the series. The gas share stays at 10.6-10.7% every year from 2000 to 2005, which looks like an estimate rather than a measurement, and Uzbekistan's coal share is 4.1% every year from 2000 to 2014. Kazakhstan's gas share jumps from 9.2% in 2011 to 16.1% in 2012, the year after the coal peak, and from 24.9% in 2022 to 29.3% in 2023, when total generation and installed gas capacity hardly changed (from 4.14 to 4.26 GW); this would mean gas plants running about 89% of the hours in 2023, an unusually high value. The 2011 peak and the post-2022 fall of coal should therefore be read with caution, while the total of fossil generation and the clean share, which national sources confirm, are more reliable. The latest year in Ember is the least final, and some 2025 values repeat the 2024 values (for example, Kazakhstan's net imports). Shares hide size: China generated about 85 times more electricity than Kazakhstan in 2025, and its coal generation in TWh was still growing until 2024. Finally, this is a description of yearly national data; it does not measure the effect of any policy.



Conclusion

Among Kazakhstan's neighbours, only China had a coal share of electricity as high as Kazakhstan's and reduced it steadily, from 78.3% in 2011 to 54.6% in 2025, mainly through wind and solar. Russia, Uzbekistan and Kyrgyzstan never depended much on coal and are not reducing it. Over 2011-2025 Kazakhstan's coal share fell about as fast as China's in the Ember data (slightly slower with the Ministry's figure), but not evenly: it fell with China's in 2011-2015, paused near 70% in 2015-2019 (in 2016-2018 growing demand was covered almost entirely by coal), was continuously above China's in 2017-2022, and then fell quickly. So the answer to "from what moment did Kazakhstan begin to lag" is: in coal share, during a pause in 2015-2019, visible as a continuously higher share than China's from 2017; in clean electricity, Kazakhstan has been behind China for the whole period, and the gap has widened, with small interruptions, since 2011. Whether Kazakhstan's coal share is now at China's level depends on the source (54.3% in Ember, 62.1% in the Ministry's figures). A next step would be to obtain the official split of thermal generation between coal and gas by year, and to check, with data by sector, which industries drove the growth in demand.

Data availability. The Ember yearly electricity data are publicly available at the address in reference [1]; the national figures are from the documents in references [2-4]. The extracted table and the analysis scripts are kept by the author's research mentor and are available on request.

Author contributions and use of AI tools. The author chose the research question (6 October 2026), wrote his expectations before the calculation (12 October 2026), found reference [5], and wrote the explanation of 2016-2018 tested in this paper (14 October 2026) and the account of smog and soot in the Introduction (17 October 2026). The data were downloaded and analysed, and the English text was drafted, by his mentors with the help of an AI assistant (Claude); the author read the final text and confirmed that it describes his part correctly.

Acknowledgements. Electricity data from Ember (CC BY 4.0).

Conflict of interest. The author declares no conflict of interest.

References:

1. Ember. Yearly electricity data: full release, long format. URL: <https://ember-energy.org/data/yearly-electricity-data/> (accessed: 27.10.2026).
2. Committee on Statistics of the Ministry of National Economy of the Republic of Kazakhstan. Energy statistics in the Republic of Kazakhstan. Nur-Sultan, 2020. 6 p. URL: <https://stat.gov.kz/upload/medialibrary/519/vv3lgjx6mrp7g2e5nwtbxfx93lztadt5/Статистика%20энергетики.pdf> (accessed: 27.10.2026).
3. KEGOC JSC. Annual report 2023: Electricity balance. Astana, 2024. URL: <https://ar2023.kegoc.kz/ru/download/electricity-balance> (accessed: 27.10.2026).
4. Kazinform. What heats and lights Kazakhstan: the main source of electricity named [Chto obogrevaet i osveshchaet Kazakhstan: nazvan osnovnoi istochnik elektroenergii]. 15.06.2026. URL: <https://www.inform.kz/ru/ugol-gaz-ili-vie-chto-segodnya-obespechivaet-energobalans-kazahstanachto-obogrev-0253191f> (accessed: 27.10.2026).
5. Zhakiyev N., Burkhanova D., Nurkanat A., Zhussipkaliyeva S., Sospanova A., Khamzina A. Green energy in grey areas: The financial and policy challenges of Kazakhstan's energy transition // Energy Research & Social Science. 2025. Vol. 124. Art. 104046. DOI: 10.1016/j.erss.2025.104046.



ЕСТЕСТВЕННЫЕ НАУКИ
ЖАРАТЫЛЫСТАНУ ҒЫЛЫМДАРЫ
NATURAL SCIENCES

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. No. 4 (32), Vol. 2. pp. 94–100
DOI: <https://doi.org/10.5281/zenodo.23772145>
Journal website: <https://cajournal.kz>

UDC 512.53

HOW MANY MAPS OF $\mathbb{Z}/n\mathbb{Z}$ PRESERVE MULTIPLICATION? A
COUNTING FORMULA FOR OEIS SEQUENCE A117986

Onal Yerkanat

Bilim-Innovation Lyceum No. 1, Karagandy, Kazakhstan

Onal Yerkanat — Grade 11 student, Bilim-Innovation Lyceum No. 1, Karagandy, Kazakhstan.

Scientific supervisor: Kaldanov Serikzhan Abdildaevich

© Onal Yerkanat, 2026

Abstract: Sequence A117986 of the On-Line Encyclopedia of Integer Sequences (OEIS) gives the number $N(n)$ of maps $f: \mathbb{Z}/n\mathbb{Z} \rightarrow \mathbb{Z}/n\mathbb{Z}$ with $f(xy) = f(x)f(y)$ for all x and y . The entry lists 56 terms and no proved formula; its only formula line, added in 2019, says that "apparently" $N(p) = p + 1$ for every prime p . This paper confirms that remark with a short proof and gives a formula for $N(n)$ for every n . By the Chinese remainder theorem the count splits into local numbers $T(p_i, q_i)$, the numbers of multiplicative maps $g: \mathbb{Z}/p_i\mathbb{Z} \rightarrow \mathbb{Z}/q_i\mathbb{Z}$ with $g(1) = 1$, and $N(n) = \prod_i (1 + \prod_j T(q_j, p_i))$, where $q_1, \dots, q_r$ are the prime powers in the factorisation of n . Each local number is a finite sum of counts of solutions of explicit congruences; the prime 2 is treated separately, because the units modulo 2^k do not form a cyclic group for $k \geq 3$. This gives $N(p^k) = p^{2k-2}(p-1) + 2$ for odd p , $N(2^k) = 3 \cdot 2^{k-2} + 2$ for $k \geq 3$, and a closed formula for square-free n . The formula agrees with all 56 published terms and with exhaustive search for every $n \leq 72$, and it gives 16 new terms. The maps between prime-power monoids had already been described in a 2026 doctoral thesis by J. Atalaye (Stellenbosch University), whose explicit forms imply the local counts in many cases; for $\mathbb{Z}/8\mathbb{Z}$, however, one of its corollaries implies 65 maps with $g(1) = 1$, while a direct enumeration gives 49. The contribution of this paper is an element-by-element count of these maps for every pair of prime powers, checked against direct counts, the resulting formula for $N(n)$, and its connection to A117986.

Keywords: multiplicative map; residue ring; monoid homomorphism; Chinese remainder theorem; integer sequence; OEIS A117986

Introduction

Consider the residues modulo n with their usual multiplication, and call a map $f: \mathbb{Z}/n\mathbb{Z} \rightarrow \mathbb{Z}/n\mathbb{Z}$ multiplicative if $f(xy) = f(x)f(y)$ for all residues x and y . Let $N(n)$ be the number of such maps. The constant maps 0 and 1 are multiplicative, and so is every power map $x \rightarrow x^e$, but for most n there are many more. John W. Layman added the sequence of these numbers to the OEIS as A117986 in 2006 [1]. In 2019 Rémy Sigrist added more terms (56 are listed now), published a C++ program that computes them, and added the only formula line the entry has: "Apparently, $a(p) = p + 1$ for any prime number p " [1]. When this paper was written (24 October 2026) the entry still had no proof, no reference and no general formula. Its first terms are 1, 3, 4, 6, 6, 35, 8, 50, 20, 55, 12, 160.

The additive version of the problem is easy: $f(x + y) = f(x) + f(y)$ forces $f(x) = x \cdot f(1)$, so there are exactly n solutions, as the entry notes [1]. The multiplicative version is not, for two reasons. First, the number of all maps grows like n^n , so a naive search over all of them stops being possible long before the published range ends; a search that fills in forced values as it goes, used in this paper, still reaches $n = 72$. Second, N is not multiplicative: $N(2) = 3$ and $N(3) = 4$, but $N(6) =$



35, not 12. Splitting n into prime factors is therefore not enough, and the way the prime factors interact has to be understood.

The work was done in this order. I proved the prime case by hand by 6 August 2026 (it was checked by my mentor's AI-assisted review on 7 August). My first formula for prime powers was correct for odd primes but gave 18 for $n = 8$, where a computer count by my mentoring team gave 50. The reason was structural: the proof used a primitive root, and the units modulo 2^k have no primitive root when $k \geq 3$, so powers of 2 needed their own argument. The same hidden assumption, now on the side of the values, spoiled my first formulas for maps between powers of different primes; those errors were found by comparing the formulas with computer counts over several rounds in August 2026. This paper gives the complete argument and the checks behind it.

Theorems 1 to 3 and Lemma 2 below were obtained in August 2026. In September 2026, after that work, my mentoring team found a doctoral thesis by J. Atalaye, completed in March 2026 at Stellenbosch University [2]. The thesis studies $\mathbb{Z}/n\mathbb{Z}$ as a monoid under multiplication. Its Proposition 4.4.1 [2, p. 79] describes all monoid homomorphisms from $(\mathbb{Z}/p^e\mathbb{Z}, \cdot)$ to $(\mathbb{Z}/q^f\mathbb{Z}, \cdot)$ through the homomorphisms of the unit groups that can be lifted, and Theorem 2.3.2 [2, p. 21] writes the endomorphisms of a finite direct product of monoids as matrices of homomorphisms between the factors; two papers by the same group determine the automorphism groups of these monoids [3, 4]. The local numbers T below are the sizes of the sets described in Proposition 4.4.1. Proposition 4.3.4, Corollary 4.4.2 and Lemma 4.4.4 of [2] (pp. 61-62, 80, 83) give explicit forms of these sets, which imply the local counts in many cases; for odd p , for example, Corollary 2 below also follows from [2, Corollary 4.4.2]. In some cases the counts implied by [2] differ from a direct count. Statement 2 of Proposition 4.3.4 [2, p. 62] says that for $e \geq 3$ every homomorphism $U(2e) \rightarrow U(2f)$ can be lifted, and Corollary 4.4.2 [2, p. 80] uses it to describe all multiplicative maps of $\mathbb{Z}/2^e\mathbb{Z}$ to itself. For $\mathbb{Z}/8\mathbb{Z}$ this would give 65 multiplicative maps g with $g(1) = 1$: the constant map 1, the 16 maps with $g(2) = 0$, and 16 maps for each of the three values 2, 4 and 6 of $g(2)$. An enumeration of all 88 maps of $\mathbb{Z}/8\mathbb{Z}$ to itself gives 49, and then $N(8) = 50$, which is the published term. One example shows the difference: there is no multiplicative map with $g(2) = 2$ and $g(5) = 3$, because $2 \cdot 5 \equiv 2 \pmod{8}$ would force $g(2) = 2 \cdot 3 = 6$. The thesis does not mention the OEIS. The prime case is easy from either description, so the proof of Theorem 1 below only confirms the remark recorded in A117986. What this paper adds is: element-by-element counts of the local maps for every pair of prime powers, including the cases $p = 2$ and $q = 2$ (Propositions 1 and 2), each checked against a direct count; a formula for $N(n)$ for every n that also counts the maps with $f(1) \neq 1$ (these are not monoid homomorphisms); closed forms for powers of 2 and for square-free n ; and 16 new terms.

Materials and methods

Definitions. For a prime power $s = p^a$ let $U(s)$ be the group of units modulo s and ϕ the Euler function. For prime powers $s = p^a$ and $t = q^b$ (the case $p = q$ is allowed) let $T(s, t)$ be the number of multiplicative maps $g: \mathbb{Z}/s\mathbb{Z} \rightarrow \mathbb{Z}/t\mathbb{Z}$ with $g(1) = 1$. A homomorphism of groups $\chi: U(p^a) \rightarrow U(q^b)$ is called a character in this paper. For a nonzero integer y , $vq(y)$ is the exponent of q in y .

Standard facts. (F1) For an odd prime p , and for $p^a = 2$ or 4 , the group $U(p^a)$ is cyclic of order $\phi(p^a)$; a generator is called a primitive root. (F2) For $a \geq 3$ every unit modulo 2^a can be written in exactly one way as $(-1)^\epsilon 5^t$ with $\epsilon \in \{0, 1\}$ and $0 \leq t < 2^{a-2}$; the residue 5 has order 2^{a-2} . Both facts are in [5, Ch. 4]. (F3) A cyclic group of order A has exactly one subgroup of each order dividing A (an elementary fact of group theory). (F4) If $n = q_1 \cdots q_r$ with pairwise coprime q_i , then $\mathbb{Z}/n\mathbb{Z}$ is isomorphic as a ring to the product of the rings $\mathbb{Z}/q_i\mathbb{Z}$ (the Chinese remainder theorem [5, Ch. 3]). (F5) Every odd w satisfies $w^{2s} \equiv 1 \pmod{2^{s+2}}$ for $s \geq 1$: for $s = 1$ because $w^2 - 1 = (w - 1)(w + 1)$ is a product of two consecutive even numbers, and then by induction, since $w^{2s+1} - 1 = (w^{2s} - 1)(w^{2s} + 1)$ and the second factor is even.

Computation. Two independent programs were compared. The first counts multiplicative maps directly: it assigns values one residue at a time, after each choice fills in every value forced by $f(xy) = f(x)f(y)$, and abandons a branch at the first contradiction. It uses no formula. It was run for $N(n)$ with $n \leq 72$ and for $T(s, t)$ for all 324 ordered pairs of prime powers $s, t \leq 32$. The second



program evaluates the formulas of Theorems 2 and 3 with the counts of Propositions 1 and 2. Both were compared with the 56 terms published in the OEIS [1], read on 24 October 2026.

Results

Lemma 1. In $\mathbb{Z}/p\mathbb{Z}$ the only solutions of $z^2 = z$ are 0 and 1.

Proof. The equation says that pa divides $z(z - 1)$. The numbers z and $z - 1$ are coprime, so p divides only one of them, and then pa divides that one. Hence $z \equiv 0$ or $z \equiv 1$.

Theorem 1 (the prime case). For every prime p , $N(p) = p + 1$.

Proof. Let f be multiplicative on $\mathbb{Z}/p\mathbb{Z}$. From $f(x) = f(x \cdot 1) = f(x)f(1)$, either $f(1) = 0$, and then f is the zero map, or $f(1) = 1$. Assume $f(1) = 1$. By Lemma 1, $f(0) = 0$ or $f(0) = 1$. If $f(0) = 1$, then $f(x) = f(0)f(x) = f(0 \cdot x) = f(0) = 1$ for every x , so f is the constant map 1. If $f(0) = 0$, then $f(x)f(x-1) = f(1) = 1$ for every nonzero x , so f maps the nonzero residues to nonzero residues and is a homomorphism of the group $U(p)$. This group is cyclic of order $p - 1$ (F1); if g is a primitive root, f is determined by $h = f(g)$, and $f(gk) = hk$. Every nonzero h is allowed, because $hp-1 = 1$ by Fermat's little theorem applied to the nonzero residue h , so the value hk depends only on k modulo $p - 1$. Conversely, each such map is multiplicative: if one factor is 0 both sides of $f(xy) = f(x)f(y)$ are 0, and otherwise $f(gkg'l) = hk+l = f(gk)f(gl)$. This gives $p - 1$ maps, and together with the two constant maps $N(p) = 1 + 1 + (p - 1) = p + 1$. If $h = ge$ with $1 \leq e \leq p - 1$, the map is the power map $x \rightarrow xe$ ($h = 1$ gives x^{p-1}), so apart from the constants every solution is a power map.

The structure of a local map. Let $g: \mathbb{Z}/pa\mathbb{Z} \rightarrow \mathbb{Z}/qb\mathbb{Z}$ be multiplicative with $g(1) = 1$. Every nonzero x can be written as $x = pku$ with $0 \leq k \leq a - 1$ and u a unit; here k is unique, and $pku = pku'$ exactly when $u \equiv u' \pmod{pa-k}$.

(i) By Lemma 1, $g(0)$ is 0 or 1. If $g(0) = 1$, then $g(x) = g(0)g(x) = g(0 \cdot x) = 1$ for all x , which is the constant map 1.

(ii) For a unit u , $g(u)g(u-1) = g(1) = 1$, so g maps units to units, and its restriction χ to $U(pa)$ is a character.

(iii) Let $g(0) = 0$. Then $g(p)a = g(pa) = g(0) = 0$. If $g(p) = 0$, then g vanishes on all non-units, and every character χ gives such a map: a product with a non-unit factor is a non-unit, so both sides of the equation are 0. These maps are counted by the number of all characters $U(pa) \rightarrow U(qb)$.

(iv) If $g(0) = 0$ and $g(p) \neq 0$, write $g(p) = qmv$ with $0 \leq m \leq b - 1$ and v not divisible by q . Then $g(p)a = 0$ exactly when $am \geq b$, that is, $m \geq b/a$. There are $\phi(qb-m)$ residues modulo qb of this form for a given m . In every case $g(pku) = g(p)k\chi(u)$, so g is determined by $g(p)$ and χ .

Lemma 2 (consistency). Let m be an integer with $b/a \leq m \leq b - 1$, let $y = qmv$ with v not divisible by q , and let χ be a character. A multiplicative map g with $g(0) = 0$, $g(p) = y$ and $g(u) = \chi(u)$ for all units u exists if and only if, for every k with $1 \leq k \leq a - 1$ and $km < b$, $\chi(u) \equiv 1 \pmod{qb-km}$ for every unit $u \equiv 1 \pmod{pa-k}$. Such a map is unique. The condition does not depend on v .

Proof. Necessity. Let $u \equiv 1 \pmod{pa-k}$. Then $pku = pk$, so $y k \chi(u) = yk$, that is, $qkmv k(\chi(u) - 1) \equiv 0 \pmod{qb}$. Since v is a unit and $km < b$, this gives $\chi(u) \equiv 1 \pmod{qb-km}$.

Sufficiency. Put $g(0) = 0$ and $g(pku) = y k \chi(u)$. The value does not depend on the choice of u : if $pku = pku'$, then $u' = uw$ with $w \equiv 1 \pmod{pa-k}$, and $y k \chi(u') - y k \chi(u) = y k \chi(u)(\chi(w) - 1)$. If $km < b$ this is divisible by $qkm \cdot qb - km = qb$; if $km \geq b$, then $yk = 0$. Now take $x = pku$ and $x' = plu'$. If $k + l < a$, then $xx' = pk+luu'$ and $g(xx') = yk+l\chi(uu') = g(x)g(x')$. If $k + l \geq a$, then $xx' = 0$ and $g(x)g(x') = yk+l\chi(uu') = 0$, because $(k + l)m \geq am \geq b$. If one factor is 0, both sides are 0. Finally $g(1) = \chi(1) = 1$, and uniqueness follows from (iv).

The condition of Lemma 2 is essentially the condition defining the liftable homomorphisms in [2, Definition 4.3.1] (cf. the proof of [2, Proposition 4.4.1]); what is new here is the element-by-element count in Propositions 1 and 2 below. Let $H_m = H_m(p, a; q, b)$ be the number of characters that satisfy the condition of Lemma 2.

Theorem 2 (local formula). For primes p, q and integers $a, b \geq 1$,

$T(pa, qb) = 1 + \#\text{Hom}(U(pa), U(qb)) + \sum_{b/a \leq m \leq b-1} \phi(qb-m) \cdot H_m(p, a; q, b)$, where $\#\text{Hom}$ is the number of all characters and the sum runs over the integers m with $b/a \leq m \leq b - 1$ (it is empty when $a = 1$ or $b = 1$).



Proof. By (i) to (iv) the maps fall into three disjoint classes: the constant map 1; the maps with $g(p) = 0$, one for each character; and the maps with $vq(g(p)) = m$, where $g(p)$ takes $\phi(qb-m)$ values and, by Lemma 2, each value goes with exactly H_m characters.

Proposition 1 (p odd, or $pa = 2$ or 4). Let r be a primitive root modulo pa and $A = \phi(pa)$. A character is determined by $h = \chi(r)$, and the allowed h are exactly the units h modulo qb with $hA \equiv 1 \pmod{qb}$; in particular $\#Hom$ is the number of such h . Moreover,

$H_m = \#\{h : hA \equiv 1 \pmod{qb}, \text{ and } hA/pk \equiv 1 \pmod{qb-km} \text{ for all } k \text{ with } 1 \leq k \leq a-1 \text{ and } km < b\}$.

Proof. The first statement holds because $U(pa)$ is cyclic of order A (F1). For $1 \leq k \leq a-1$, the units $u \equiv 1 \pmod{pa-k}$ form the kernel of the reduction map from $U(pa)$ onto $U(pa-k)$, a subgroup of order $\phi(pa)/\phi(pa-k) = pk$. By (F3) it is the subgroup generated by rA/pk , and since the units $\equiv 1$ modulo a power of q also form a subgroup, the condition of Lemma 2 for this k holds exactly when $\chi(rA/pk) = hA/pk$ satisfies it.

Proposition 2 ($p = 2$, $a \geq 3$). By (F2) a character is determined by $c = \chi(-1)$ and $d = \chi(5)$, and the allowed pairs are exactly those with $c^2 \equiv 1$ and $d^{2a-2} \equiv 1 \pmod{qb}$. Moreover, H_m is the number of such pairs with $d^{2a-k-2} \equiv 1 \pmod{qb-km}$ for all k with $1 \leq k \leq a-2$ and $km < b$, and, when $(a-1)m < b$, also $c \equiv d \equiv 1$ modulo the power of q with exponent $b - (a-1)m$.

Proof. By (F2) the units $\equiv 1 \pmod{4}$ are exactly the powers of 5. For $1 \leq k \leq a-2$ the units $\equiv 1 \pmod{2a-k}$ form a subgroup of order $2k$ of the cyclic group generated by 5, so by (F3) they are generated by 5^{2a-k-2} . For $k = a-1$ the condition concerns all units, which are generated by -1 and 5.

Remark. Propositions 1 and 2 count elements of $U(qb)$ one by one, so they need no assumption about the structure of $U(qb)$. This matters. The shortcut $\#Hom = \gcd(\phi(pa), \phi(qb))$ is guaranteed only when both unit groups are cyclic, and otherwise it can fail: there are 4 characters $U(8) \rightarrow U(3)$ and 4 characters $U(3) \rightarrow U(8)$, while $\gcd(4, 2) = 2$. Assuming cyclicity on one side or the other was the source of the errors described in the introduction.

Theorem 3 (product formula). Let $n = q_1 \cdots q_r$, where $q_i = p_i^{a_i}$ are powers of distinct primes p_i . Then

$$N(n) = \prod_{i=1..r} (1 + \prod_{j=1..r} T(q_j, q_i)).$$

Proof. By (F4), f corresponds to its coordinate maps $F_i: \mathbb{Z}/n\mathbb{Z} \rightarrow \mathbb{Z}/q_i\mathbb{Z}$, and f is multiplicative exactly when every F_i is. By Lemma 1, $F_i(1)$ is 0 or 1. If $F_i(1) = 0$, then $F_i(x) = F_i(x)F_i(1) = 0$ for all x . If $F_i(1) = 1$, let ij send a residue y modulo q_j to the residue modulo n whose j -th coordinate is y and whose other coordinates are 1. The map ij is multiplicative, so $g_{ji}(y) = F_i(ij(y))$ defines a multiplicative map from $\mathbb{Z}/q_j\mathbb{Z}$ to $\mathbb{Z}/q_i\mathbb{Z}$ with $g_{ji}(1) = 1$. Every x is the product of the $ij(x_j)$, where x_j are its coordinates, so $F_i(x) = \prod_j g_{ji}(x_j)$. Conversely, for any multiplicative maps g_{ji} with $g_{ji}(1) = 1$ this formula defines a multiplicative F_i with $F_i(1) = 1$, and the g_{ji} are recovered as $g_{ji}(y) = F_i(ij(y))$. So the i -th coordinate has $1 + \prod_j T(q_j, q_i)$ possibilities, chosen independently for different i . For $n = 1$ the empty product gives $N(1) = 1$.

The proof shows why N is not multiplicative: the coordinate modulo q_i depends on every prime factor of n , not only on p_i . For $n = 6$, $T(2, 2) = T(3, 2) = T(2, 3) = 2$ and $T(3, 3) = 3$, so $N(6) = (1 + 2 \cdot 2)(1 + 2 \cdot 3) = 5 \cdot 7 = 35$. For $n = 24$, $T(8, 8) = 49$, $T(3, 8) = T(8, 3) = 5$ and $T(3, 3) = 3$, so $N(24) = (1 + 49 \cdot 5)(1 + 5 \cdot 3) = 246 \cdot 16 = 3936$, which is the published term.

Corollary 1 (square-free n). For primes p and q , $T(p, q) = 1 + \gcd(p-1, q-1)$. Hence for distinct primes $p_1, \dots, p_r$,

$$N(p_1 \cdots p_r) = \prod_i (1 + \prod_j (1 + \gcd(p_j - 1, p_i - 1))).$$

Proof. For $b = 1$ the sum in Theorem 2 is empty, and the characters between the cyclic groups $U(p)$ and $U(q)$ of orders $p-1$ and $q-1$ number $\gcd(p-1, q-1)$. For $r = 1$ the formula gives $1 + p$, in agreement with Theorem 1.

Corollary 2 (powers of an odd prime). For an odd prime p and $k \geq 1$, $N(pk) = p^{2k-2}(p-1) + 2$. The same formula holds for 2 and 4: $N(2) = 3$ and $N(4) = 6$.

Proof. By Theorem 3, $N(pk) = 1 + T(pk, pk)$. Let $U = U(pk)$, cyclic of order $A = pk-1(p-1)$. There are A characters $U \rightarrow U$. Let $1 \leq m \leq k-1$ and let $1 \leq j \leq k-1$ with $jm < k$. The units $u \equiv$



$1 \pmod{pk-j}$ form the subgroup of U of order pj ; its image under any character is a subgroup of U of order dividing pj . The subgroup of U of order pj is cyclic and contains a subgroup of every order dividing pj , so by the uniqueness in (F3) the image lies in it; and that subgroup consists of the units $\equiv 1 \pmod{pk-j}$. Since $jm \geq j$, these are $\equiv 1 \pmod{pk-jm}$. So every character satisfies the condition of Lemma 2, and $H_m = A$. Theorem 2 gives $T = 1 + A + A \cdot (\varphi(pk-1) + \dots + \varphi(p)) = 1 + A \cdot pk-1$, because the $\varphi(pe)$ for $1 \leq e \leq k-1$ add up to $pk-1-1$. Hence $N(pk) = 2 + p^{2k-2}(p-1)$. For 2 and 4 the group of units is also cyclic, the same argument applies, and the formula gives 3 and 6. For odd p the corollary also follows from [2, Corollary 4.4.2]; for $p = 2$ and $k \geq 3$ the count below differs from the one implied by [2, Corollary 4.4.2] (see the Introduction).

Corollary 3 (powers of 2). For $k \geq 3$, $N(2k) = 3 \cdot 2^{2k-2} + 2$.

Proof. Let $U = U(2k)$. By (F2) a character $U \rightarrow U$ is given by $c = \chi(-1)$ with $c^2 = 1$ (4 choices: ± 1 and $\pm 1 + 2k-1$) and $d = \chi(5)$, any element of U ($2k-1$ choices, since every unit has order dividing $2k-2$); so there are $2k+1$ characters. Take $m = 1$ in Proposition 2. The condition for $k' = k-1$ concerns residues modulo 2, so it always holds. For $1 \leq k' \leq k-3$ the condition $d^{2s} \equiv 1 \pmod{2s+2}$ with $s = k-k'-2 \geq 1$ always holds by (F5). For $k' = k-2$ it says $d \equiv 1 \pmod{4}$, which leaves half of the $2k-1$ choices of d . So $H_1 = 4 \cdot 2^{k-2} = 2^k$. For $m \geq 2$ every condition holds: $k' = k-1$ does not occur, because $(k-1)m \geq 2k-2 \geq k$; for $k' \leq k-3$ the modulus $2k-k'm$ divides $2k-k'$, which (F5) covers; and $k' = k-2$ occurs only when $k = 3$ and $m = 2$, where the condition is modulo 2. So $H_m = 2^{k+1}$. Theorem 2 gives $T = 1 + 2^{k+1} + 2^{k-2} \cdot 2^k + 2^{k+1} \cdot (2^{k-3} + \dots + 1) = 1 + 2^{k+1} + 2^{2k-2} + 2^{k+1}(2^{k-2} - 1) = 1 + 3 \cdot 2^{2k-2}$, and $N(2k) = 1 + T$. For example, $N(8) = 50$, $N(16) = 194$ and $N(32) = 770$.

Checks. The local formula of Theorem 2, with Propositions 1 and 2, agreed with the direct count of multiplicative maps for all 324 ordered pairs of prime powers $s, t \leq 32$. Theorem 3 agreed with the direct count of all multiplicative maps for every $n \leq 72$ and with all 56 terms published in the OEIS. Table 1 gives some local numbers, and Table 2 gives the 16 terms after the published range; each of them was obtained both from the formula and by the direct count.

Table 1. Local numbers $T(s, t)$ for small prime powers (s is the source, t is the target)

$s \setminus t$	2	4	8	3	9	5
2	2	2	2	2	2	2
4	2	5	9	3	5	3
8	2	9	49	5	9	5
3	2	3	5	3	3	3
9	2	5	9	3	19	3
5	2	3	5	3	3	5

Table 2. The terms $N(57)$ to $N(72)$, beyond the 56 published in A117986

n	57	58	59	60	61	62	63	64
$N(n)$	580	295	60	59248	62	315	6700	3074
n	65	66	67	68	69	70	71	72
$N(n)$	1716	11457	68	832	700	11997	72	76024

Discussion

The answer to the question in the title is: $N(n)$ is determined by the prime factorisation of n , but not as a product of separate answers for the prime powers. Each prime power contributes a factor $1 + \prod_j T(q_j, q_i)$ that depends on all the other prime powers, and the "1 +" counts the choice of an identically zero coordinate; the maps with $f(1) \neq 1$ are exactly those with at least one zero coordinate. The local numbers T come from three kinds of maps: the constant map 1, the maps that kill every non-unit, and the maps that send p to an element of exact q -valuation m , which survive only if the character on the units cannot see the difference between u and u' modulo $qb-km$ whenever $pku = pku'$. Lemma 2 makes this last condition exact.

The case $p = 2$ is the delicate one. Modulo $2k$ with $k \geq 3$ there is no primitive root: 3, 5 and 7 all square to 1 modulo 8. An argument that writes every unit as a power of one element therefore loses solutions, which is why my first formula gave 18 maps modulo 8 instead of 50. The same



mistake reappeared later on the side of the values, where it is less visible. In every case the computer comparison found the error and the proof had to explain it, which is the main methodological lesson of this work: a check up to 72 does not prove anything about 73, but it is very good at showing where a proof is wrong.

The relation to the thesis [2] is as follows. $T(pa, qb)$ is the size of the set of monoid homomorphisms described in Proposition 4.4.1 there; the condition $m \geq b/a$ corresponds to its condition $v \geq f/e$, and the characters counted by H_m correspond to its liftable homomorphisms. Theorem 3 is the counting form of the matrix description of [2, Theorem 2.3.2], extended to maps with $f(1) \neq 1$. The thesis studies the algebraic structure of these sets and of the automorphism groups; this paper counts all multiplicative maps and gives conditions that can be checked by hand and that were compared with a direct count in every case tested. For maps of $\mathbb{Z}/p\mathbb{Z}$ to itself with p odd, [2, Corollary 4.4.2] gives the same count as Corollary 2; for $\mathbb{Z}/8\mathbb{Z}$, [2, Corollary 4.4.2] implies 65 maps with $g(1) = 1$, while the direct count is 49. The statement recorded in A117986 for primes follows easily from either description; the contribution here is to write the count out for every n and to extend the table.

Limitations

For arbitrary prime powers the number H_m is given as the number of solutions of an explicit system of congruences, not by one elementary expression. It is computed quickly, but a single closed formula for all pairs of prime powers was found only in the special cases of Corollaries 1 to 3. The direct count was run for $n \leq 72$ and for prime powers up to 32; the formula is proved for every n , but the programs check it only on this range. The proofs were checked by computer comparisons, by my own AI-assisted checks and by an AI-assisted review by my mentoring team; they have not yet been reviewed by a specialist. The exponents in Theorem 2 and in the propositions are easy to get wrong, and the history of this work shows it: several earlier versions failed on particular pairs before the present form passed every test.

Conclusion

The statement "apparently, $a(p) = p + 1$ " recorded in OEIS A117986 is true: apart from the two constant maps, every multiplicative map of $\mathbb{Z}/p\mathbb{Z}$ is a power map, and there are $p - 1$ of them. For every n , the number of multiplicative maps of $\mathbb{Z}/n\mathbb{Z}$ is $N(n) = \prod_i (1 + \prod_j T(q_j, q_i))$, where the local numbers T are given by Theorem 2 and Propositions 1 and 2. For prime powers this gives $N(pk) = p^{2k-2}(p - 1) + 2$ for odd p and $N(2k) = 3 \cdot 2^{2k-2} + 2$ for $k \geq 3$, and for square-free n an explicit product of greatest common divisors. The formula agrees with all 56 published terms and with exhaustive search up to 72, and it extends the sequence by the 16 terms of Table 2. The maps between prime-power monoids had already been described in [2]; the element-by-element counts checked against direct counts, the formula for $N(n)$, its closed forms and the new terms are what this paper adds. A natural next step is to submit the formula and the new terms to the OEIS entry, with a reference to this paper and to [2].

Author contributions and use of AI tools. The author chose functional equations as the field and this counting question from options proposed by his mentor in the ADMIT mentoring programme (July 2026); this mentor is not the scientific supervisor named above. The author proved Theorem 1 by hand (sent on 6 August 2026 and checked by the mentor's AI-assisted review on 7 August); the proof above follows his, with the write-up corrected (Fermat's theorem applied only to nonzero residues, the converse direction added). He found the formula for powers of an odd prime and, after a computer count by the mentor showed that his first formula failed at $n = 8$, proved the case of powers of 2 separately (9 August, checked by the mentor's AI-assisted computer count on 10 August), using the structure of the units modulo $2k$ that the mentor had pointed out on 8 August. The author first proved both prime-power formulas directly; the derivation of Corollaries 2 and 3 from Theorem 2 given here is the mentor's. The mentor, working with a generative AI tool (Claude, Anthropic), proposed the reduction of the composite case to local numbers (Theorem 3), the formula for $T(p, q)$ and Corollary 1, and described the structure of a local map (10 August 2026); identified the consistency condition for every k , which became Lemma 2 (11 August); wrote and ran all programs except the author's own, including the comparisons with the OEIS, the 324



local pairs and the search up to $n = 72$ that produced the new terms; found the thesis [2] (September 2026); and drafted the English text of this paper from the author's handwritten proofs, his Russian and Kazakh drafts and his messages. The author corrected the summation limits in the local formula, wrote out the explicit counts of Propositions 1 and 2, proved Lemma 2 in both directions and wrote out the proof of Theorem 3 in both directions (18 August 2026, checked on 19 August), after several rounds in which his formulas were compared with computer counts; he also wrote a short program that counts $N(n)$ directly for $n \leq 24$. The author states that he used generative AI tools to check his work, and he describes the rest as about 90% an olympiad-style solution. Before submission the text and the proofs were checked in two independent reviews by separate AI agents. Every number in the paper comes from the output of the programs. The author's mentoring team read the OEIS entry [1] and the thesis [2] in full and the abstracts of [3] and [4]; [5] is cited for standard textbook facts.

Acknowledgments. The scientific supervisor, Kaldanov Serikzhan Abdildaevich, mathematics teacher at Bilim-Innovation Lyceum No. 1, agreed on 1 August 2026 to be named as the scientific supervisor of this project. An earlier Russian version of this work was entered in the 2026-2027 Republican competition of research projects of school students, organised by the Republican Scientific and Practical Centre "Daryn", jointly with Miras Kenesbek, who is the speaker for the joint entry and supports its defence; he took no other part in the work and is not an author of this paper. The Russian text of that version was prepared by the author's mentoring team from the author's work. The author thanks his mentor in the ADMIT mentoring programme for the work described above.

Data availability. The published terms are available at <https://oeis.org/A117986>. The programs that produce every number in this paper, and their output, are held by the author's mentor and are available on request from the author.

Conflict of interest. The author declares no conflict of interest.

References:

1. Sequence A117986: Number of functions $f:[n] \rightarrow [n]$ such that $f[(x*y) \bmod n] = [f(x)*f(y)] \bmod n$ for all x, y in $[n]$ // The On-Line Encyclopedia of Integer Sequences. OEIS Foundation Inc. URL: <https://oeis.org/A117986> (accessed: 24.10.2026).
2. Atalaye J. Exploring the multiplicative structure of $\mathbb{Z}/n\mathbb{Z}$: PhD dissertation. Stellenbosch: Stellenbosch University, 2026. 120 p. URL: <https://scholar.sun.ac.za/server/api/core/bitstreams/34ac3e49-a014-49c7-b376-84aa27df66cc/content> (accessed: 24.10.2026).
3. Atalaye J., Baker L., Marques S. On the automorphism group of the monoid of the integers modulo a prime power // arXiv. 2024. arXiv:2408.06278. URL: <https://arxiv.org/abs/2408.06278> (accessed: 24.10.2026).
4. Atalaye J., Baker L., Marques S. Automorphism groups of direct products of multiplicative monoids of certain rings // arXiv. 2026. arXiv:2605.06119. URL: <https://arxiv.org/abs/2605.06119> (accessed: 24.10.2026).
5. Ireland K., Rosen M. A Classical Introduction to Modern Number Theory. 2nd ed. New York: Springer, 1990. (Graduate Texts in Mathematics; Vol. 84). DOI: 10.1007/978-1-4757-2103-4.



ӘЛЕУМЕТТІК ЖӘНЕ ГУМАНИТАРЛЫҚ ҒЫЛЫМДАР
СОЦИАЛЬНО-ГУМАНИТАРНЫЕ НАУКИ
SOCIAL SCIENCES AND HUMANITIES

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. No. 4 (32), Vol. 2. pp. 101–106
DOI: <https://doi.org/10.5281/zenodo.23776973>
Journal website: <https://cajournal.kz>

UDC 81

LANGUAGE OF SCHOOLING AND PERCEPTIONS OF ENGLISH
ACCENTED SPEECH IN KAZAKHSTAN

Saltanat Mussina, Torgyn Yeralinova

Ye. A. Buketov, Kazakhstan

Saltanat Mussina — Department of Foreign Philology, Karaganda National Research University named after academician, Ye. A. Buketov, Kazakhstan.

Torgyn Yeralinova — Department of Foreign Philology, Karaganda National Research University named after academician, Ye. A. Buketov, Kazakhstan.

© Saltanat Mussina, Torgyn Yeralinova, 2026

Abstract: This exploratory study examined whether language of secondary schooling was associated with university students' evaluations of English-accented speech in Kazakhstan. Twenty undergraduates, comprising 10 students with predominantly Kazakh-medium schooling and 10 with predominantly Russian-medium schooling, evaluated eight recordings by different speakers. Data were collected in September 2026 during a university phonetics course. Each recording lasted 30–60 seconds and was played once through classroom loudspeakers in a fixed order. Participants rated comprehensibility, perceived accent strength, and perceived speaker competence on separate nine-point scales. Complete responses yielded 160 listener–recording observations per outcome. Ratings were averaged within listeners before two-sided Welch tests, with Holm adjustment across three outcomes. Kazakh-medium listeners gave numerically higher ratings for comprehensibility (6.55 vs. 6.25), accent strength (5.78 vs. 5.58), and competence (6.85 vs. 6.38). All confidence intervals for mean differences included zero, and no Welch comparison was statistically significant after adjustment. Mann–Whitney sensitivity analyses likewise yielded no significant adjusted comparisons. The findings provide preliminary listener-level estimates but do not establish reliable group differences or a causal effect of schooling language. Interpretation is limited by the small convenience sample, fixed presentation order, unadjusted proficiency differences, and the absence of an objective intelligibility measure.

Keywords: accented English, comprehensibility, perceived accentedness, listener judgments, language of schooling, Kazakhstan

1. Introduction

Perception of additional-language English cannot be explained solely by the speaker's pronunciation. Research distinguishes accentedness (the perceived strength of a non-native accent), comprehensibility (the listener's perceived ease of understanding), and intelligibility (what the listener actually recognizes). Munro and Derwing (1995) found that heavily accented speech can nevertheless remain intelligible, and their later synthesis emphasizes that the three dimensions are related but not interchangeable (Munro & Derwing, 2022). Listener characteristics are therefore a legitimate object of investigation rather than merely background variation.

Expectations can also shape how listeners evaluate speech. Kang and Rubin (2009) examined reverse linguistic stereotyping and showed why information about a speaker's presumed group membership may influence speech evaluations. Research designs that separate auditory evidence from social expectations are consequently important when studying judgments of accented



speech. A rating of competence is especially sensitive to this issue: it represents a listener's perception and should not be treated as an objective measure of a speaker's language ability.

Kazakhstan provides a relevant multilingual educational context in which Kazakh, Russian, and English may occupy different positions in a person's language-learning history. Research on language learning in Kazakhstan has examined L1/L2 transfer in L3 English acquisition (Knol et al., 2026), while the present study addresses a related but distinct question: whether listeners with different school-language backgrounds evaluate the same English speech samples differently. The study does not assume in advance that either group will be more tolerant of an accent, and it is not intended to represent all students in Kazakhstan.

2. Research Questions

RQ1. Do students with Kazakh-medium and Russian-medium secondary-school backgrounds differ in their ratings of the comprehensibility of the same English speech recordings?

RQ2. Do the two groups differ in perceived accent strength and perceived speaker competence?

RQ3. Do the conclusions from mean-based comparisons remain consistent with rank-based sensitivity analyses?

3. Method

3.1. Design and participants

This exploratory study examined the association between language of secondary schooling and university students' evaluations of English-accented speech. Twenty undergraduate students enrolled in English Philology and Translation Studies at Karaganda Buketov University participated in the study. Recruitment and data collection took place in September 2026 during the course Theory of the Studied Language (Phonetics).

Participants received an oral invitation during class, with an explanation of the study's purpose and the voluntary nature of participation. Convenience sampling with quotas based on language of secondary schooling was used to recruit 10 students with predominantly Kazakh-medium backgrounds and 10 with predominantly Russian-medium backgrounds. Eligibility criteria were enrolment in one of the two-degree programs, consent to participate, and provision of information about the predominant language of secondary schooling.

Each schooling group included five English Philology students and five Translation Studies students. The mean age was 21.8 years in each group. Participants had studied English for a mean of 8.8 years ($SD = 0.42$) in the Kazakh-medium group and 8.6 years ($SD = 0.84$) in the Russian-medium group.

Schooling background was classified according to the predominant language of secondary education reported by each participant. These categories were not treated as indicators of ethnicity, first language, or monolingual status. Participants' home languages and language-use patterns varied within both groups. Self-rated English levels were Intermediate ($n = 4$), Upper-intermediate ($n = 5$), and Advanced ($n = 1$) in the Kazakh-medium group, and Intermediate ($n = 4$), Upper-intermediate ($n = 2$), and Advanced ($n = 4$) in the Russian-medium group.

3.2. Speech materials and measures

Eight English speech recordings, each lasting between 30 and 60 seconds, were selected from the author's own recording archive and used with the speakers' permission. The recordings featured eight different speakers and were selected based on audio quality, duration, and suitability for the phonetics listening task. They were assigned anonymous codes C01–C08. No comparisons between accent varieties were undertaken.

Participants evaluated each recording on three separate nine-point scales. Comprehensibility ranged from 1 (very difficult to understand) to 9 (very easy to understand). Perceived accent strength ranged from 1 (not noticeable) to 9 (very strong). Perceived speaker competence ranged from 1 (very low) to 9 (very high).

Comprehensibility represented the listener's subjective ease of understanding the speech. Accent strength referred to the perceived prominence of the accent, whereas perceived competence represented the listener's impression of the speaker's language ability. These judgments were not



treated as objective measures of proficiency. No usable transcription responses or scores were available; therefore, objective intelligibility was not assessed.

3.3. Procedure and data screening

Participants listened together in a teaching classroom during the course Theory of the Studied Language (Phonetics). The instructor played all eight recordings through classroom loudspeakers in the same fixed order for all participants. Each recording was played once. Instructions were provided in English. Participants recorded their evaluations on paper rating forms containing recording codes C01–C08 and the three nine-point scales. Responses were subsequently entered into an Excel workbook for analysis.

The dataset contained 160 participant–recording observations, corresponding to 20 listeners evaluating eight recordings. No duplicate participant–recording pairs were identified. All three primary rating variables were complete and within the permitted range of 1–9. Each participant contributed eight ratings per outcome, and all 20 participants were retained.

The optional transcription and exclusion-reason fields contained the label “Actual” rather than usable transcription responses, numeric transcription scores, or substantive exclusion reasons. These entries were treated as metadata and were not used as outcomes or as grounds for exclusion. The Clip_Order field contained values inconsistent with the fixed presentation procedure and was not used in the analyses.

3.4. Statistical analysis

The eight ratings for each outcome were averaged within each listener. The independent unit for between-group inference was therefore the listener ($n = 10$ per group), not the individual listener–recording observation. Group means and standard deviations were calculated from listener averages. Two-sided Welch independent-samples t tests compared groups, with Satterthwaite degrees of freedom and unadjusted 95% confidence intervals for mean differences. Differences were calculated as Kazakh-medium minus Russian-medium. Holm correction was applied across the three Welch tests, with a familywise significance threshold of .05.

Hedges’ g was calculated from the mean difference divided by the pooled listener-level standard deviation, using the small-sample correction $1 - 3/(4df - 1)$, where $df = 18$. Positive values indicate higher ratings in the Kazakh-medium group. Confidence intervals presented in Table 1 refer to raw mean differences, not to standardized effect sizes.

Two-sided asymptotic Mann–Whitney U tests, with tie correction and continuity correction, were used as a sensitivity analysis. Holm correction was applied separately across these three tests. Rank-based tests address relative ordering of observations rather than the same mean-difference estimand as Welch tests. All analyses were exploratory. No covariate adjustment was undertaken with this small sample, and no causal inference was intended. Inference is conditional on the eight selected recordings; the analysis does not estimate variation across a wider population of speakers.

Calculations were checked using Python with pandas 2.2.3, NumPy 2.3.5, and SciPy 1.17.0. Holm adjustment and the Hedges’ g correction were implemented directly.

3.5. Ethical considerations

Participation was voluntary, and all participants provided informed consent before taking part. The purpose of the study and the voluntary nature of participation were explained during recruitment. The speech recordings were drawn from the author’s own archive and used with the speakers’ permission. Participant codes were used instead of names in the analysis dataset. No claim of institutional ethics approval or formal exemption is made, as a documented institutional determination was not available for reporting.

4. Results

4.1. Data completeness

All 20 participants contributed complete ratings for all eight recordings and all three primary outcomes. Thus, 160 observations were available per outcome. No primary rating values were missing, and no participant–recording duplicates were detected. The mean-based comparisons below use 20 participant-level averages for each outcome, with 10 participants in each schooling group.



4.2. Descriptive and inferential comparisons

The Kazakh-medium group had a higher mean rating than the Russian-medium group on all three outcomes, although the size and uncertainty of the differences varied. The largest raw mean difference was observed for perceived competence (+0.48 scale points), followed by comprehensibility (+0.30) and perceived accent strength (+0.20). Table 1 summarizes the group-level participant averages and Welch tests.

Table 1. Participant-level rating averages and comparisons by language of schooling

Outcome	Kazakh-medium M (SD)	Russian-medium M (SD)	Mean difference (95% CI)	Welch test; p; Holm p; Hedges g
Comprehensibility	6.55 (0.48)	6.25 (0.40)	0.30 [-0.12, 0.72]	$t(17.54) = 1.52$; $p = .146$; $p_{\text{Holm}} = .293$; $g = 0.65$
Perceived accent strength	5.78 (0.76)	5.58 (0.43)	0.20 [-0.39, 0.79]	$t(14.21) = 0.72$; $p = .481$; $p_{\text{Holm}} = .481$; $g = 0.31$
Perceived competence	6.85 (0.46)	6.38 (0.66)	0.48 [-0.06, 1.01]	$t(16.12) = 1.88$; $p = .079$; $p_{\text{Holm}} = .237$; $g = 0.80$

Note. $n = 10$ per group. Higher scores indicate greater ease of understanding, stronger perceived accent, or higher perceived competence, respectively. Mean differences are Kazakh-medium minus Russian-medium. Confidence intervals are 95% Welch intervals. Holm-adjusted p values control familywise error across the three primary tests. Positive Hedges' g values indicate higher mean ratings in the Kazakh-medium group.

Welch's tests did not show statistically significant group differences at the .05 level after correction for the three primary outcomes. For comprehensibility, the estimated difference was 0.30 points, 95% CI [-0.12, 0.72], $t(17.54) = 1.52$, $p = .146$, Holm-adjusted $p = .293$. For perceived accent strength, the estimated difference was 0.20 points, 95% CI [-0.39, 0.79], $t(14.21) = 0.72$, $p = .481$, Holm-adjusted $p = .481$. For perceived competence, the estimated difference was 0.48 points, 95% CI [-0.06, 1.01], $t(16.12) = 1.88$, $p = .079$, Holm-adjusted $p = .237$. The confidence intervals include zero for every outcome, so the mean-based results are compatible with both negligible differences and potentially meaningful positive differences for the Kazakh-medium group.

4.3. Sensitivity analysis and variation across recordings

The Mann–Whitney sensitivity analysis yielded $U = 79.0$, $p = .029$, Holm-adjusted $p = .088$ for comprehensibility; $U = 62.5$, $p = .360$, Holm-adjusted $p = .360$ for perceived accent strength; and $U = 72.5$, $p = .095$, Holm-adjusted $p = .189$ for perceived competence. No comparison remained below .05 after adjustment. The smaller unadjusted rank-based p value for comprehensibility should be interpreted in light of the different estimand and the small sample, rather than as confirmation of a mean difference.

Descriptively, the Kazakh-medium group's mean comprehensibility rating was higher for seven recordings and equal for one. Mean perceived competence ratings were higher for the Kazakh-medium group for all eight recordings. In contrast, perceived accent-strength differences varied by recording, with the direction changing for some clips and no difference for one clip. These recording-level patterns are descriptive only; with eight stimuli and 20 listeners, they should not be interpreted as evidence of a general speaker- or accent-level effect.

5. Discussion

This exploratory analysis asked whether the language of secondary schooling was associated with how undergraduate listeners evaluated English-accented speech. Across the three ratings, the Kazakh-medium group gave descriptively higher scores than the Russian-medium group. However, the primary participant-level Welch comparisons were not statistically significant after correcting for the three outcomes, and each 95% confidence interval included zero. The data therefore do not



provide robust evidence that the two schooling groups differed in their mean evaluations in this sample.

The perceived-competence difference was largest in scale-point and standardized terms (Hedges' $g = 0.80$), but the confidence interval for the raw mean difference included zero and the adjusted p value was .237. In a sample of only 10 participants per group, effect sizes can be unstable and uncertainty is substantial. This pattern should not be translated into the claim that one group is more accurate or fair in evaluating speakers. Perceived competence is a social judgment, not an objective test of language proficiency.

The comprehensibility outcome merits particular caution. The mean-based test suggested a positive difference of 0.30 points in favor of the Kazakh-medium group, while the rank-based sensitivity test produced a smaller unadjusted p value. The latter result was no longer below .05 after Holm correction. This dependence on the analytical method is a reason to avoid a binary conclusion and to prioritize replication with a larger sample, prespecified analyses, and transparent documentation of the ratings.

The distinction between perceived comprehensibility and intelligibility remains important. Because no usable transcription responses were available, the present study cannot determine whether the groups differed in actual word recognition. Subjective ease-of-understanding ratings do not directly measure this ability. This distinction is consistent with the conceptual separation of accentedness, comprehensibility, and intelligibility in speech-perception research (Munro & Derwing, 1995, 2022). The present comparison also cannot establish causation: English proficiency, everyday exposure, accent familiarity, and other aspects of multilingual experience may be associated with both schooling history and listener judgments.

The study contributes preliminary evidence about listener evaluations in a local multilingual university setting. Its estimates concern the selected recordings and the sampled students, rather than all Kazakh- or Russian-medium learners. Future sample-size planning should allow for substantial uncertainty in these pilot effect estimates rather than relying directly on a single observed standardized difference.

6. Limitations and Future Research

Several limitations constrain interpretation. First, the sample comprised only 20 students from one university, recruited through convenience sampling with schooling-language quotas, which limits precision and generalizability. Second, schooling medium is a broad proxy for a complex multilingual biography. The groups differed descriptively in self-rated English proficiency, and proficiency and exposure were not controlled statistically. Third, inference is conditional on the eight selected recordings and cannot establish a general effect across speakers or accent varieties. Fourth, no usable transcription responses or scores were available, preventing objective intelligibility analysis. Fifth, all participants heard the recordings in the same fixed order, so recording-specific differences cannot be separated from possible order effects, such as fatigue or increasing familiarity with the task. Listening through classroom loudspeakers may also have introduced differences in audibility depending on seating position.

Future research should preregister the primary outcome and analysis plan, recruit a larger sample across institutions, and model the repeated-measures structure directly, ideally with mixed-effects models including listener and speaker/recording effects were supported by the design. Recording-level variation should be sampled systematically rather than relying on a small set of stimuli. Adding a transcription task would allow perceived comprehensibility to be compared with an objective intelligibility measure. Researchers should also document ethics approval or exemption, informed consent, stimulus permissions, and data-handling procedures.

7. Conclusion

In this 20-listener exploratory study, ratings by students with Kazakh-medium school backgrounds were descriptively higher than those by students with Russian-medium backgrounds for comprehensibility, perceived accent strength, and perceived competence. The group differences were not statistically robust after correction for multiple comparisons, and the comprehensibility result varied by analytical approach. The findings are therefore preliminary and do not establish a



reliable effect of schooling language. A larger, preregistered study combining subjective ratings with an objective transcription task is needed to determine whether schooling background is meaningfully associated with the perception of English-accented speech in Kazakhstan.

Data Availability Statement

The dataset contains coded participant characteristics and listener ratings. It is not publicly available because of participant privacy considerations. Requests for access may be directed to the corresponding author and will be considered subject to participant consent and applicable institutional requirements.

References:

1. Kang, O., & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, 28(4), 441–456. <https://doi.org/10.1177/0261927X09341950>
2. Knol, M., Assanova, D., Smagulova, G., Uteubayeva, E., & Akbaeva, G. (2026). Reducing L1 and L2 transfer in L3 English acquisition: An intervention study in Kazakhstani trilingual education. *Frontiers in Education*, 11, Article 1850753. <https://doi.org/10.3389/feduc.2026.1850753>
3. Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 73–97. <https://doi.org/10.1111/j.1467-1770.1995.tb00963.x>
4. Munro, M. J., & Derwing, T. M. (2022). Foreign accent, comprehensibility and intelligibility, redux. In J. M. Levis, T. M. Derwing, & M. J. Munro (Eds.), *The evolution of pronunciation teaching and research: 25 years of intelligibility, comprehensibility, and accentedness* (pp. 7–32). John Benjamins. <https://doi.org/10.1075/bct.121.02mun>

Электронный научный журнал «Central Asian Scientific Journal»

Редактор: Байдильдинов Т.Ж.
Комп.верстка: Хусаинов Е.М.

Электронный научный журнал «Central Asian Scientific Journal»
–2026–4(32)–Астана–ИП ДОС
Зарегистрировано и выдано свидетельство
Министерством Информации и Общественного Развития РК
№KZ77VPY00147053 от 17.04.2026 г.
ISSN: 3135–3061

*За достоверность публикуемой информации, цитат и
иных изложений ответственность несет автор*



