

UDC 004.8:811.512.122

Zhumagali Yerbol

12th-grade student

Nazarbayev Intellectual School of Science and Mathematics in Turkistan
(Turkistan, Kazakhstan)

DOES A BIGGER MODEL CLOSE THE GAP? KAZAKH-LANGUAGE ACCURACY ON KAZAKHSTAN-SPECIFIC VERSUS UNIVERSAL SCHOOL SUBJECTS IN SMALL OPEN-WEIGHT LANGUAGE MODELS

Abstract: Open-weight language models are usually less accurate in Kazakh than in Russian, but a single Kazakh-versus-Russian score cannot tell whether the model struggles with the language itself or with knowledge about Kazakhstan. We separate the two with a difference-in-differences design on KazMMLU, a bilingual school-level benchmark. Kazakh History, a Kazakhstan-specific subject available in both languages, is compared with mathematics, physics and chemistry, which are Kazakhstan-neutral. We evaluate three sizes of the same open model family (Qwen2.5-Instruct 1.5B, 3B and 7B) with deterministic multiple-choice scoring. On universal subjects the gap between Kazakh and Russian stays within about 1.5 percentage points at every size. On Kazakh History, Kazakh is 13 to 18 points behind Russian, and the gap does not shrink as the model grows: the extra Kazakh penalty on Kazakhstan content is -12.6 , -12.9 and -16.7 points (95% CI excludes zero for all three). Scaling within this family improves knowledge about Kazakhstan mainly through Russian. We discuss training-data imbalance as the likely cause and the limits of a single-family, single-subject design.

Keywords: Kazakh language, language models, KazMMLU, low-resource languages, regional knowledge, model scaling

1. Introduction

Language models are now used by students in Kazakhstan for homework, exam preparation and everyday questions, often in a mix of Kazakh and Russian. Models are known to perform worse in Kazakh than in Russian. What is less clear is where that gap comes from. A model may be weaker in Kazakh because it handles the language poorly, or because it knows less about Kazakhstan when the question is asked in Kazakh. The two explanations call for different fixes, yet a single aggregate score mixes them.

KazMMLU was built to measure exactly this kind of knowledge. It contains 23,000 questions from authentic Kazakhstani educational materials, 10,969 in Kazakh and 12,031 in Russian, covering STEM, humanities and social sciences [1]. Its authors report that even strong multilingual models struggle in both languages. Work that uses KazMMLU usually reports a pooled score. For example, Mussa et al. [2] evaluate eleven open-weight checkpoints and use the Kazakh portion of KazMMLU (12 subjects, 9,870 questions) as one measure of general knowledge, reporting Kazakh-only accuracy. Related work on Kazakh and Russian does separate regional from general content, but for safety rather than accuracy (Qorgau [3]). To our knowledge, no study has split the Kazakh-versus-Russian accuracy gap by Kazakhstan-specific versus universal subjects, or asked how that split changes with model size.

We ask: does scaling a small open-weight model close the Kazakh-versus-Russian accuracy gap equally for universal school subjects and for Kazakhstan-specific ones?

2. Data and Method

Benchmark. We use the Kazakh and Russian subsets of KazMMLU. The two subsets are not translations of each other: they have no shared item identifiers and come from different sources. A raw Kazakh-versus-Russian gap therefore mixes the language effect with a difference in item difficulty.

Design. To separate the two, we compare two kinds of subjects, each available in both languages. The Kazakhstan-specific cell is Kazakh History. The universal cell combines Mathematics, Physics and Chemistry, which do not depend on knowledge about Kazakhstan and carry the least language load. The quantity of interest is the difference-in-differences:

$DiD = (\text{Kazakh} - \text{Russian accuracy} \mid \text{Kazakh History}) - (\text{Kazakh} - \text{Russian accuracy} \mid \text{universal subjects})$

Any item-source difference that is common to all subjects cancels out. What remains is the additional penalty that Kazakh carries on Kazakhstan content. The subject split and the analysis were fixed before the results were seen. Geography, Law and World History mix Kazakhstan-specific and general content and were kept out of the main comparison.

Models. Qwen2.5-Instruct at 1.5B, 3B and 7B parameters [4], 4-bit quantised, run locally. Using one family at three sizes lets us look at scale while holding the training recipe roughly constant.

Scoring. We use standard multiple-choice scoring. Each prompt is written in the item's own language, with three examples from that subset's development split. The model reads the question once, and the predicted answer is the option letter with the highest probability among the letters the item actually has. No text is generated and nothing is sampled, so the run is fully reproducible.

Sample. For each model we score 300 four-option Kazakh History items and 900 universal items (300 per subject) in each language. The main analysis uses four-option items only, because the two languages differ in their share of four- and five-option items and therefore in the chance level. Results on all items are reported as a robustness check. Standard errors follow the binomial formula; the DiD standard error combines the four cells.

Check for answer bias. For each model and language we record the share of answers that go to the most frequently chosen letter (0.25 means no letter preference).

3. Results

Table 1 – Accuracy on four-option items, by model, subject group and language

Model	Kazakh History, Kazakh	Kazakh History, Russian	Gap	Universal gap	DiD (95% CI)
Qwen2.5-1.5B	24.7%	38.7%	−14.0 pp	−1.4 pp	−12.6 pp (−21.1 to −4.0)
Qwen2.5-3B	26.0%	39.3%	−13.3 pp	−0.4 pp	−12.9 pp (−21.6 to −4.2)
Qwen2.5-7B	29.7%	47.3%	−17.7 pp	−1.0 pp	−16.7 pp (−25.6 to −7.7)

Three results stand out. First, on universal subjects the language gap is small at every size (−1.4, −0.4 and −1.0 points). The models read and answer Kazakh school questions about

mathematics, physics and chemistry about as well as Russian ones. Second, on Kazakh History the Kazakh penalty is large, 13 to 18 points, and the difference-in-differences is negative and clearly different from zero for all three models. Third, the penalty does not shrink with scale. From 1.5B to 7B, Kazakh History accuracy rises by 8.6 points in Russian but only 5.0 points in Kazakh, so the largest model shows the largest gap.

The pattern holds on all items, four- and five-option together: DiD -10.3 points ($z = -2.71$), -9.3 points ($z = -2.41$) and -16.1 points ($z = -4.08$).

The answer-bias check shows that the smallest models prefer one letter more often in Kazakh than in Russian (1.5B: 44% vs 31%; 3B: 35% vs 28%), while the 7B model shows no difference (30% vs 30%). Part of the small models' Kazakh gap may therefore reflect guessing behaviour rather than missing knowledge. The 7B result, where this bias disappears, is the cleanest estimate, and it is also the largest.

4. Discussion

Before running the experiment, we expected Kazakh to fall behind everywhere, in history and in the exact sciences alike, and expected the gap to narrow as the model grew, since larger models usually handle rare languages better. Neither expectation held. The gap of about one point in mathematics, physics and chemistry shows that the models understand questions in Kazakh reasonably well; the problem is knowledge of Kazakhstan-specific content. And growing the model did not close that gap: at 7B it was wider than at 1.5B (18 points against 14).

Why would a larger model gain about nine points in Russian on Kazakh History but only about five in Kazakh? We think the answer lies in the training data. Far more text about the history of Kazakhstan exists in Russian than in Kazakh in the corpora such models learn from. A larger model can hold more facts, but it can only remember what it has seen, so in Russian the extra capacity turns into knowledge, while in Kazakh there is little to remember. Kazakh is also tokenised less efficiently than Russian, which may further lower quality. We did not measure the training data or tokenisation directly, so this explanation remains a hypothesis that fits the results rather than a tested claim.

The practical consequence is concrete. A Kazakhstani student who asks a small open model about the history of their own country gets a noticeably worse answer in Kazakh than in Russian, and switching to a bigger model of the same family does not fix it. Making regional knowledge available in Kazakh, rather than only increasing model size, appears to be what matters.

5. Limitations

Sample size. Each cell contains about 300 questions, so every accuracy carries an uncertainty of roughly ± 2.5 to 3 points. The comparison between a nine-point and a five-point gain may be partly noise, and the 3B model barely improves over 1.5B in Russian ($+0.6$ points), so the growth appears mostly at 7B. The main finding, a 13 to 18 point gap on Kazakh History against about one point on universal subjects, is well outside this uncertainty. A repeated run or a larger item sample would sharpen the scale comparison.

One model family. Only Qwen2.5 was tested. Whether the pattern holds for other families remains open.

One Kazakhstan-specific subject. Kazakh History is the only subject that is clearly about Kazakhstan; other mixed subjects were excluded from the main contrast.

Non-parallel subsets. The design removes only the item-source difference that is shared across subjects.

Quantisation. All models were run in 4-bit form, which may lower absolute accuracy.

Answer bias in the smallest models, described above.

6. Conclusion

In three sizes of the same open model family, the gap between Kazakh and Russian accuracy is close to zero for universal school subjects but 13 to 18 points for Kazakh History, and it does not close as the model grows. The weakness is not the Kazakh language as such but knowledge about Kazakhstan expressed in Kazakh. Scaling alone does not solve it.

References:

1. Togmanov M., Mukhituly N., Turmakhan D. [et al.] KazMMLU: Evaluating Language Models on Kazakh, Russian, and Regional Knowledge of Kazakhstan // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Vol. 1. – 2025. – P. 14403–14416. – URL: <https://arxiv.org/abs/2502.12829> (accessed: 05.10.2026).
2. Mussa A., Tuimebayev Z., Mansurova M., Shihab A.H. When Does Fine-Tuning Matter for a Low-Resource Language? Evaluating Open-Weight LLMs on Kazakh Benchmarks // Information. – 2026. – Vol. 17, No. 9. – Art. 921. – URL: <https://www.mdpi.com/2078-2489/17/9/921> (accessed: 05.10.2026).
3. Goloburda M., Laiyk N., Turmakhan D. [et al.] Qorgau: Evaluating LLM Safety in Kazakh-Russian Bilingual Contexts [Electronic resource]. – 2025. – arXiv:2502.13640. – URL: <https://arxiv.org/abs/2502.13640> (accessed: 05.10.2026).
4. Qwen Team. Qwen2.5 Technical Report [Electronic resource]. – 2024. – arXiv:2412.15115. – URL: <https://arxiv.org/abs/2412.15115> (accessed: 05.10.2026).

МАЗМҰНЫ – СОДЕРЖАНИЕ – CONTENT

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

Төлегенов Тимур Сейфуллаулы ДИНАМИКА ТУРИСТСКИХ ПОТОКОВ И ЭФФЕКТИВНОСТЬ ИНФРАСТРУКТУРЫ ТУРКЕСТАНА: ОТ ЭКСТЕНСИВНОГО РОСТА К СМАРТ-УПРАВЛЕНИЮ	10
Alexander Gorevoy WHICH INEQUALITY STATISTICS SURVIVE AN ADMINISTRATIVE BOUNDARY CHANGE? EVIDENCE FROM KAZAKHSTAN, 2010–2023	14
Rakhimov Djalil DOES THE SHIFT TO CASHLESS PAYMENTS GO TOGETHER WITH FASTER GROWTH IN HOUSEHOLD CREDIT? EVIDENCE FROM THE REGIONS OF KAZAKHSTAN, 2023-2025	26
Kazhgerey Asan HOW MANY FOOD BASKETS DOES AN AVERAGE WAGE BUY? WAGES AND FOOD PRICES IN THE 20 REGIONS OF KAZAKHSTAN, 2024-2026	37
Tagimova Alua WHO WAS LEFT BEHIND? EDUCATION AND FINANCIAL INCLUSION IN KAZAKHSTAN AND THREE CENTRAL ASIAN NEIGHBOURS, 2021-2024	44
Wei Wei, Mynzhanova Gulzhakan, Niyetalina Gaukhar DIGITAL–PHYSICAL INTEGRATION AND ECONOMIC PERFORMANCE IN INDUSTRIAL INTEGRATED ENERGY SYSTEMS: A SCENARIO-BASED ANALYSIS OF SMART GRID CONTROL, BATTERY ENERGY STORAGE, AND DIGITAL TWINS	51
Xu Fumin, Wei Wei, Ruziyeva Elvira, Moldabekova A. FROM CULTURAL RESOURCES TO GLOBAL CONSUMER VALUE: THE BRIDGE–IMC MODEL AND THE INTERNATIONALIZATION OF BLACK MYTH: WUKONG UNDER THE BELT AND ROAD INITIATIVE	64
Ibraimkhan Alikhan WHICH CRIMES CAN A 14-YEAR-OLD BE TRIED FOR? THE LIST IN ARTICLE 15 OF KAZAKHSTAN'S CRIMINAL CODE COMPARED WITH 30 OTHER JURISDICTIONS	80
Yerboluly Madi DID IMPORTED CARS REALLY GET CHEAPER? CHINESE CARS AND THE AVERAGE CUSTOMS VALUE OF PASSENGER CARS IMPORTED INTO KAZAKHSTAN, 2015-2025	91
Салимбаев Ербол Аманжолович АНАЛИЗ ИНВЕСТИЦИОННОЙ АКТИВНОСТИ КАК ИНДИКАТОР ИЗМЕНЕНИЙ ПОТЕНЦИАЛЬНОГО СПРОСА НА КОНСАЛТИНГОВЫЕ УСЛУГИ	99
Rambergenov Ramil REGIONAL NEW-FIRM FORMATION IN UZBEKISTAN AFTER THE 2017 REFORMS: PERSISTENCE IN SCALE, CONVERGENCE IN INTENSITY, AND RECOMPOSITION TOWARD TRADE	107
Abdesh Dinara FROM A WOMEN'S LEAD TO A MEN'S LEAD: THE GENDER GAP IN DIGITAL PAYMENTS IN KAZAKHSTAN AND THREE CENTRAL ASIAN NEIGHBOURS, 2021-2024	116
Amangeldi Aidana DID SMALL BUSINESS IN KAZAKHSTAN REALLY SHRINK IN 2026? THE NEW TAX CODE, REGIONAL TAX RATES AND THE FALL IN THE NUMBER OF INDIVIDUAL ENTREPRENEURS	124

Cheng Che	
BUILD, BUY, OR PARTNER? ENTRY MODE SELECTION AND RISK GOVERNANCE FOR WIND POWER INVESTMENT IN KAZAKHSTAN	133
Wang Liang	
FROM CAPACITY EXPANSION TO SYSTEM VALUE: COORDINATED DEVELOPMENT OF RENEWABLE ENERGY, GRID FLEXIBILITY AND ENERGY STORAGE IN KAZAKHSTAN	146
Chen Jun	
CONSTRUCTION-PERIOD CREDIT RISK MANAGEMENT FOR LARGE INFRASTRUCTURE PROJECT FINANCE IN KAZAKHSTAN	159
Alniyaz Danial	
HOW MUCH FUEL DOES A HYBRID REALLY SAVE? A SAME-MODEL COMPARISON OF HYBRID AND GASOLINE CARS IN US FUEL ECONOMY	173

ТЕХНИКАЛЫҚ ҒЫЛЫМДАР - ТЕХНИЧЕСКИЕ НАУКИ - TECHNICAL SCIENCE

Nurmakhan Almira	
ADVANTAGES OF FIBER OPTIC COMMUNICATION OVER TRADITIONAL COPPER CABLES	180
Kozhabai Bereke	
SECURITY CHALLENGES IN IOT-BASED TELECOMMUNICATION NETWORKS	188
Azamatuly Yerassyl	
WHY AI-GENERATED CODE FAILS A BENCHMARK: A FAILURE ANALYSIS OF THREE OPEN LANGUAGE MODELS ON 30 HUMANEVAL+ TASKS	197
Nurzhan Sanzhar	
A FEW SHELLS, A FEW OPERATORS: WHERE LOW EARTH ORBIT IS CROWDED IN AN ALTITUDE-INCLINATION MAP OF 16,620 ACTIVE SATELLITES	206
Askarbek Nurislam	
CAN THE WEATHER TELL WHICH DAYS IN ASTANA WILL BE POLLUTED? CLASSIFYING PM2.5 EXCEEDANCE DAYS WITH MACHINE LEARNING AND CHECKING THE DATA AGAINST A GROUND MONITOR	212
Iliyas Arman	
SHOOTING ON THE MOVE: HOW MUCH AIM A ROBOT LOSES BY IGNORING ITS OWN VELOCITY, AND WHETHER A VIRTUAL-TARGET CORRECTION RECOVERS IT	223
Ануар Әлімжан Бахытжанұлы, Сексенбаева Айша Курманбековна	
СРАВНИТЕЛЬНЫЙ АНАЛИЗ СТАНДАРТОВ И МЕТОДИК ОБУЧЕНИЯ КИБЕРБЕЗОПАСНОСТИ: NICE FRAMEWORK, ECSF И ПРАКТИКО- ОРИЕНТИРОВАННЫЕ ТЕХНОЛОГИИ	229
Mukhamedzhan Yerganat	
GROWTH OF GITHUB DEVELOPER COMMUNITIES IN KAZAKHSTAN, UZBEKISTAN AND KYRGYZSTAN, 2020-2026: EVIDENCE FROM THE GITHUB INNOVATION GRAPH	239

ПӘНДЕР АРАЛЫҚ ҒЫЛЫМДАР – МЕЖДИСЦИПЛИНАРНЫЕ НАУКИ – INTERDISCIPLINARY SCIENCES

Albiyev Asser	
HOW MUCH OF AN ARTICLE COUNT IS HUMAN WORK? DECOMPOSING THE CREATION OF FIVE WIKIPEDIAS, 2010–2025	251
Amangeldy Nurbolat	
HOW MUCH OF KAZAKHSTAN'S URBAN BUILT STOCK PREDATES ITS SEISMIC CODES, AND WHERE DOES THAT STOCK SIT RELATIVE TO HAZARD?	261

Torekhan Sabrina

WHERE ALMATY WALKS ITS OWN WAY: INFORMAL SHORTCUTS IN OPENSTREETMAP AND THE DETOURS THEY AVOID	267
---	------------

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

Мұқатаева Айзат Амантайқызы, Стамбекова Жанат Байғазықызы, Слямханова Гүлім Болатқызы, Қызырбек Аяулым Ерболқызы

ЖАСАНДЫ ИНТЕЛЛЕКТ ДӘУІРІНДЕГІ ҒЫЛЫМИ ТАНЫМ: ПОППЕР, КУН ЖӘНЕ ЛАКАТОС ИДЕЯЛАРЫ ТҮРҒЫСЫНАН ТАЛДАУ	272
--	------------

Мұқатаева Айзат Амантайқызы, Сапарханова Мадина Төлегенқызы, Окенова Аружан Муратхановна, Тұрсынғазыева Ақерке Ерболатқызы

ШӘКӘРІМ ФИЛОСОФИЯСЫНДАҒЫ ЕРКІНДІК ЖӘНЕ МОРАЛЬДЫҚ ТАҢДАУ МӘСЕЛЕСІ	282
---	------------

Komekova Saida

DO UNHAPPY REVIEWERS MIX IN MORE RUSSIAN? STAR RATINGS AND KAZAKH-RUSSIAN CODE-SWITCHING IN 175,002 ONLINE REVIEWS	291
---	------------

Zhumagali Yerbol

DOES A BIGGER MODEL CLOSE THE GAP? KAZAKH-LANGUAGE ACCURACY ON KAZAKHSTAN-SPECIFIC VERSUS UNIVERSAL SCHOOL SUBJECTS IN SMALL OPEN-WEIGHT LANGUAGE MODELS	304
---	------------

Abdrakhmanova Madina

DOES A MEDAL BUY ATTENTION? WIKIPEDIA VIEWS OF KAZAKHSTAN'S ATHLETES BEFORE AND AFTER THE PARIS 2024 OLYMPIC	308
---	------------

ПЕДАГОГИКАЛЫҚ ЖӘНЕ ПСИХОЛОГИЯЛЫҚ ҒЫЛЫМДАР – ПЕДАГОГИЧЕСКИЕ И ПСИХОЛОГИЧЕСКИЕ НАУКИ – PEDAGOGICAL AND PSYCHOLOGICAL SCIENCES

Kaldarbek Mergen

FAMILY OR SCHOOL? SOCIOECONOMIC COMPOSITION AND MATHEMATICS ACHIEVEMENT IN KAZAKHSTAN, PISA 2022 AND 2025	316
--	------------

Zhenis Adilzhan

WHICH OLYMPIAD PROBLEMS DO STUDENTS TAKE TO THE BOARD? CHOICES AND DEFENCES IN THE 8TH-GRADE LEAGUE OF KARAGANDY MATHBATTLES 2026	323
--	------------

Базарбаева Айнура Бектуровна

ДЕТСКИЙ ФОЛЬКЛОР И ПОДВИЖНЫЕ ИГРЫ КАК ИНСТРУМЕНТЫ ФОРМИРОВАНИЯ ОСНОВ ПРАВОВОЙ КУЛЬТУРЫ У ДЕТЕЙ СТАРШЕГО ДОШКОЛЬНОГО ВОЗРАСТА	330
---	------------

БИОЛОГИЯ ҒЫЛЫМДАР - БИОЛОГИЧЕСКИЕ НАУКИ - BIOLOGICAL SCIENCES

Bazhenova Diana

GROWTH DYNAMICS OF STERLET (<i>Acipenser ruthenus</i> L.) REARED IN A RECIRCULATING AQUACULTURE SYSTEM	333
---	------------

ТУРИСТИЧЕСКИЕ И СПОРТИВНЫЕ НАУКИ - ТУРИСТІК ЖӘНЕ СПОРТ ҒЫЛЫМДАРЫ - TOURISM AND SPORTS SCIENCES

Төлеп Жазира Маратқызы

ГАСТРОНОМИЯЛЫҚ ФЕСТИВАЛЬДЕРДІҢ ТУРИСТІК ДЕСТИНАЦИЯ ТАРТЫМДЫЛЫҒЫН ҚАЛЫПТАСТЫРУДАҒЫ РӨЛІ МЕН ДАМУ ҮРДІСТЕРІ	340
--	------------

ЕСТЕСТВЕННЫЕ НАУКИ - ЖАРАТЫЛЫСТАНУ ҒЫЛЫМДАРЫ - NATURAL SCIENCES

Нургельдина Асель Ермаковна, Серикбаева Жанерке Алимжановна

ШЕКАРАСЫЗ ФИЗИКА: ӘР ЕЛДЕГІ ФИЗИКАЛЫҚ БІЛІМДІ САЛЫСТЫРУ 347