UDC 004.8

Bolatov Sanzhar
Grade 10 student
Supervisor: Koishybayeva Assel
FIZTEX Science & IT School
(Almaty, Kazakhstan)

HOW MUCH DOES KAZAKH SPEECH RECOGNITION LOSE WHEN THE MODEL GETS SMALLER? KAZAKH AGAINST 82 LANGUAGES IN THE PUBLISHED WHISPER RESULTS

Abstract: Small speech recognition models are the ones that run on ordinary laptops and phones, so how a language fares in a small model decides whether its speakers can use offline voice tools. This paper asks whether Kazakh loses more than other languages when the Whisper speech recognition model is made smaller, and whether this depends on how much Kazakh speech the model was trained on. No model was run: the study uses the word error rates (WER) that the Whisper authors published for 82 languages and six models from 39 million to 1.55 billion parameters on the FLEURS benchmark, and the training hours they published for each language. Before any calculation, the author forecast that Kazakh would fall below 50% WER only at the largest size and that its nearest neighbours would be other Turkic languages with little data. Kazakh WER falls from 165.2% in the smallest model to 37.7% in the largest; the 50% line is crossed one size earlier than forecast, at medium (48.8%). Measured in percentage points, Kazakh loses more than any of 68 comparable languages between the largest and the smallest model (127.5 points), but relative to its own best result its loss is ordinary (33rd of 68). Across 60 languages, more training hours go with lower error at every size (Spearman ρ from −0.88 to −0.92), and Kazakh's nearest languages by training hours and by the shape of its error curve are not Turkic (for example Hindi, Armenian and Icelandic). Only the smallest model does clearly worse on Kazakh than its training hours predict. The results agree with the author's conclusion that the amount of training data matters more than the language itself, although published tables cannot prove a cause.

Keywords: speech recognition, Kazakh language, Whisper, model size, low-resource languages, word error rate

Introduction

The author's interest in the topic began at home. In his own words: "I first noticed that AI understands Kazakh poorly when I was helping my younger brother with his homework. We translated a text from Kazakh to understand what it meant, but the translation came out strange and meaningless. This did not happen with Russian or English, and I wondered why it happened with Kazakh in particular. I started to investigate this question and assumed that how well a model works depends on how much data in the language it saw during training".

Speech recognition turns spoken words into text. It is used in voice assistants, automatic subtitles and dictation, and for a language with few written resources it is also a way into digital services for people who prefer to speak rather than type. Whisper is an open speech recognition model released by OpenAI together with its weights [1, 3]. It was trained on 680,000 hours of audio with transcripts collected from the internet and was released in

several sizes, from tiny (39 million parameters) to large (1,550 million) [1]. The sizes differ greatly in what they need to run: according to the official repository, tiny needs about 1 GB of video memory and large about 10 GB [3]. A developer who wants an application to work offline on an ordinary laptop or phone is therefore pushed towards the smaller models.

The Whisper authors reported the accuracy of every size on FLEURS, a benchmark of read sentences in many languages [1, 2], and they also published how many hours of transcribed speech each language had in training [1]. For Kazakh this was 12 hours; for Russian it was 9,761 hours, about 800 times more [1]. They found that across languages the logarithm of the error rate falls almost linearly with the logarithm of the training hours, with a squared correlation of 0.83, and estimated that the error halves for every 16-fold increase in training data [1]. What they did not report is how this relation changes with model size: whether a language with little data loses more than other languages when the model is made smaller.

This question matters for Kazakh in particular. In text models, it is already known that open models do worse in Kazakh than in Russian or English [8], and that compressing a model hurts languages with less pretraining data more strongly as the compression becomes harder [7]. Speech recognition in small models has not been examined for Kazakh in this way. This paper uses only numbers that are already published, so every result can be checked against the original tables, and asks two questions: (1) at which Whisper size Kazakh speech recognition becomes usable, taking an error below 50% as the line, and (2) whether Kazakh loses more than other languages when the model gets smaller, and whether this depends on training hours.

Related work

Whisper and its evaluation. Radford et al. [1] trained Whisper on weakly supervised data and evaluated it without any fine-tuning on each test set. For multilingual speech recognition they used FLEURS [2] and reported results for 82 languages in an appendix table, for all six released models. They noted that many of the languages that do worse than their training hours predict "have unique scripts and are more distantly related to the Indo-European languages" [1, p. 7]. Since the paper, a further large model (large-v3) and a faster version of it (turbo) have been released; the official repository lists no newer versions of tiny, base, small or medium [3].

Language data and model compression. The author's earlier reading concerned language models that are compressed by quantization, that is, by storing their weights with fewer bits. Marchisio et al. [7] found that quantization damages non-Latin-script languages more, and that the link between the amount of a language's pretraining data and how well it survives quantization becomes stronger as the compression becomes harder. Smaller Whisper models are not compressed versions of the large model: each size was trained separately [1]. The question is similar, however: when the model has less capacity, which languages pay the price?

Kazakh language technology. Kazakh has its own open speech corpora. The Institute of Smart Systems and Artificial Intelligence (ISSAI) at Nazarbayev University released the Kazakh Speech Corpus (KSC) [6] and then KSC2, an open corpus of more than a thousand hours [4]; a 2026 extension of KSC2 contains 1,151 hours of speech and 627,382 utterances [5]. For text, the KazMMLU benchmark showed that strong language models lag behind on Kazakh questions and do worst on the Kazakh language and reading subjects [8]. Studies

published in this journal have measured how open language models handle mixed Kazakh-Russian text [9]. The present paper adds the speech side and the dimension of model size.

The author's forecast

On 13 October 2026, after choosing this question and before any number was calculated, the author was given the five Whisper sizes and asked at which size Kazakh becomes usable, which languages will be near Kazakh and why. He wrote: "I think that Kazakh will reach an acceptable level (an error below 50%) only at the large size. All the smaller sizes will lag behind, because they do not have enough size to fully learn Kazakh: the space in the model goes mostly to languages with thousands of hours of data. Next to Kazakh there will be other Turkic languages with little data and few training hours, for example Uzbek".

The forecast contains three claims that the published tables can test: (F1) Kazakh WER falls below 50% only at large; (F2) smaller sizes lag because their capacity goes to languages with much more data; (F3) Kazakh's nearest languages are Turkic languages with little data, such as Uzbek. The threshold of 50% is the author's own choice of a usable level, not a standard from the literature.

Data and methods

Error rates. Word error rate (WER) is the number of word substitutions, deletions and insertions needed to turn the model's transcript into the reference transcript, divided by the number of words in the reference. Because inserted words count as errors, WER can exceed 100%. The WER values come from Table 13 "WER (%) on Fleurs" of the Whisper paper (arXiv 2212.04356v1, Appendix D.2.4, p. 24) [1]. It gives 82 languages for six models: tiny (39M parameters), base (74M), small (244M), medium (769M), large (1550M) and large-v2 (1550M). One value is missing (Javanese, large-v2). For languages other than English, the Whisper authors computed WER after a basic standardization of the text (removing punctuation and symbols and making the text lowercase) [1, Appendix C].

Training hours. The number of hours of transcribed speech per language comes from Figure 11 "Training dataset statistics" (Appendix E, p. 27, panel "Multilingual Speech Recognition") [1]. The paper has no table of these hours, so the numbers printed next to each bar were used, not bar lengths. The panel lists 74 languages. Of these, 67 also appear in Table 13 ("Burmese" in the figure and "Myanmar" in the table were treated as the same language); English has no hours in this panel.

Extraction. Both sources were read from the PDF by the word coordinates of every printed number, and each number was assigned to the nearest column heading; the program stops if any language does not receive exactly six values. The extracted values for Kazakh, Russian, Turkish, Uzbek and Tajik agree with the values read separately from the PDF on 12 October 2026. As a check, the log-log fit between training hours and large-v2 WER over the 67 languages was recomputed: it gives $R^2 = 0.83$, the same value the Whisper authors report for this relation [1, p. 7].

Samples. A language with an error above 100% even in its best model is not being recognised at all, and ratios to such a value mean little. Comparisons of loss therefore use the languages whose WER in the reference model is below 100%: 68 languages with large-v2 as the reference and 67 with large. Analyses that need training hours use the 60 languages that have hours and a large-v2 WER below 100% (the "main sample").

Measures of loss. The loss from shrinking the model was measured in two ways: in percentage points (WER of the smaller model minus WER of the reference) and relatively (WER of the smaller model divided by WER of the reference, minus one). The main reference is large-v2, the best model in the table. Because large-v2 was not only larger than tiny but also trained for 2.5 times more epochs with added regularization [1, p. 4], every comparison was repeated with the original large model as the reference, which differs from the smaller models only in size.

Statistics. Associations with training hours were measured with Spearman's rank correlation; p-values were obtained from 10,000 random permutations (two-sided). To ask whether Kazakh does worse than its training hours predict, a straight line was fitted to log10 WER (or log10 of the loss ratio) against log10 training hours over the main sample, and Kazakh's distance from the line (its residual) was ranked among the 60 languages and divided by the standard deviation of all residuals (z). The languages nearest to Kazakh were found in four ways: by training hours, by large-v2 WER, by the whole six-size curve (distance between the log10 WER values) and by the ratio of tiny to large-v2 WER. All calculations were done with the Python standard library; no speech or language model was run.

Results

Kazakh by model size

Table 1 and Figure 1 show Kazakh WER for every size next to Russian. Kazakh WER falls from 165.2% in tiny to 37.7% in large-v2. The first size below 50% is medium, with 48.8%, 1.2 points under the line; the same holds if large-v2 is left out. Even the best Kazakh result (37.7%) is worse than the result of the smallest model on Russian (31.1%).

Table 1. Whisper WER (%) on FLEURS for Kazakh and Russian by model size

Language (training hours)	tiny	base	small	medium	large	large-v2
Kazakh (12)	165.2	109.2	70.3	48.8	43.8	37.7
Russian (9,761)	31.1	20.5	11.4	7.2	6.4	5.6

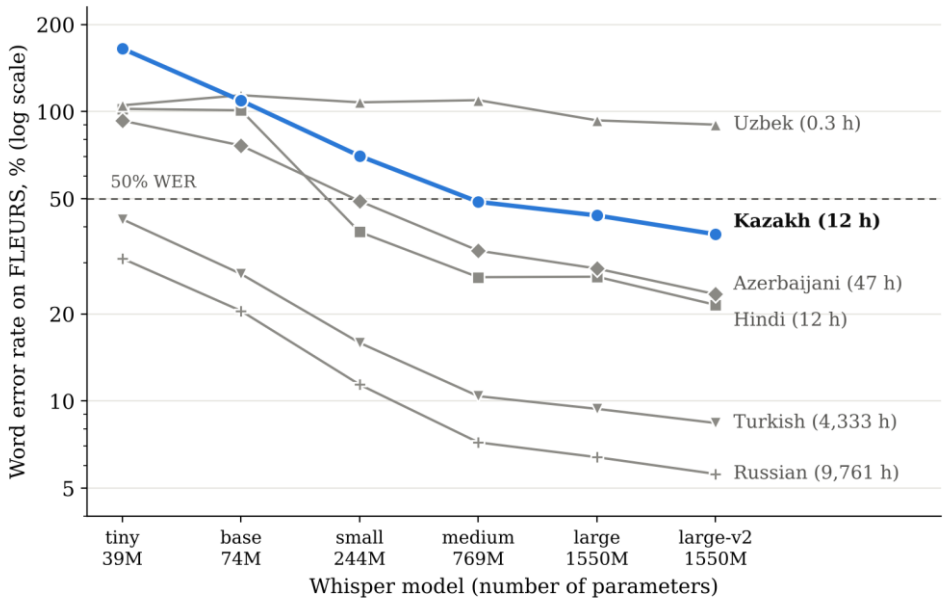


Figure 1. Whisper WER on FLEURS by model size for Kazakh and five other languages (log scale). Training hours from Figure 11 of [1] in brackets; the dashed line marks the author's 50% threshold. Kazakh crosses it at medium.

Across all 82 languages, the first size with WER below 50% is tiny for 14 languages, base for 11, small for 17, medium for 8, large for 3 and large-v2 for 4, and 25 languages never fall below 50%. The more training hours a language had, the smaller the size at which it crosses the line (Spearman $\rho = -0.93$, $n = 67$, $p < 0.001$). Table 2 shows the 12 languages with between 4 and 36 training hours, that is, within three times Kazakh's 12 hours in either direction. Medium is a common crossing point in this group: four languages, Kazakh among them, cross there, five cross earlier (four at small, one at base), and three cross later or never.

Table 2. First model size with WER below 50% for languages with 4 to 36 training hours

Language	Training hours	First size below 50%	large-v2 WER, %
Serbian	28	small	33.9
Persian	24	medium	32.9
Icelandic	16	medium	38.2
Macedonian	16	small	16.5
Armenian	13	large-v2	44.6
Hindi	12	small	21.5
Kazakh	12	medium	37.7
Bosnian	11	small	15.7
Galician	8.9	base	15.4
Swahili	5.4	large	39.3
Telugu	4.3	never	99.0
Afrikaans	4.1	medium	36.7

Does Kazakh lose more than other languages?

The two measures of loss give opposite answers (Table 3). In percentage points, Kazakh loses more than any other language when large-v2 is replaced by tiny: 127.5 points, against 82.4 for Tamil and 80.8 for Hindi, which are next. Relative to its own best result, the same loss is ordinary: tiny is 4.38 times worse than large-v2 on Kazakh (+338%), which places Kazakh 33rd of 68, close to the median of +315%. For every other size Kazakh's relative loss is slightly below the median. Russian loses more than Kazakh in relative terms (tiny is 5.55 times worse than large-v2), and the largest relative losses belong to languages with a lot of data, such as Polish (+744%), Italian (+645%) and Dutch (+631%).

Table 3. Kazakh's loss when large-v2 is replaced by a smaller model, compared with 68 languages (rank 1 = largest loss)

Smaller model	Relative loss, Kazakh	Rank	Median, all	Loss in points, Kazakh	Rank
tiny	+338%	33	+315%	127.5	1
base	+190%	41	+225%	71.5	3
small	+86%	41	+93%	32.6	6
medium	+29%	40	+32%	11.1	12
large	+16%	36	+17%	6.1	9

With the original large model as the reference (67 languages), the picture is the same: Kazakh is first in points (121.4) and 31st in relative loss (+277%). Much of Kazakh's large loss in points comes from the part of the tiny error above 100%, where the model produces

more wrong or extra words than the reference contains. If every WER is capped at 100%, the level of a model that outputs nothing, Kazakh's relative loss for tiny falls to 46th of 68.

Training hours, error and loss

In the main sample of 60 languages, more training hours go with lower WER at every size (Spearman ρ between −0.88 and −0.92, all $p < 0.001$). Training hours are also linked to the loss, but in opposite directions for the two measures. Languages with fewer hours lose more percentage points when the model shrinks (tiny versus large-v2: $\rho = -0.35$, $p = 0.007$; for the other sizes ρ between −0.53 and −0.64). Languages with more hours lose more relative to their best result (tiny versus large-v2: $\rho = +0.67$, $p < 0.001$; base +0.51; small +0.33), while for medium and large against large-v2 there is no link (ρ = +0.02 and +0.09). Low-resource languages start from a high error, so even a moderate relative increase adds many points; high-resource languages start from a very low error, which a small model multiplies several times.

Table 4 compares Kazakh with what its 12 hours predict. From small upward, Kazakh WER is about 1.2 to 1.3 times higher than predicted, which is within the normal scatter (z between +0.5 and +0.8). The clear exception is tiny: Kazakh tiny WER is 1.82 times higher than predicted, the largest excess among the 60 languages ($z = +2.25$). Its loss ratio from large-v2 to tiny is also above the line (1.47 times the predicted ratio, 7th of 60), but this excess disappears when WER is capped at 100%. From medium upward, Kazakh loses exactly as much as its training hours predict.

Table 4. Kazakh against a log-linear fit on training hours (60 languages; rank 1 = furthest above the line)

Measure	Kazakh	Predicted from 12 hours	Actual / predicted	Rank of 60
WER tiny, %	165.2	90.9	1.82	1
WER base, %	109.2	79.3	1.38	7
WER small, %	70.3	54.5	1.29	14
WER medium, %	48.8	39.8	1.23	14
WER large-v2, %	37.7	30.5	1.24	17
Ratio tiny / large-v2	4.38	2.98	1.47	7
Ratio medium / large-v2	1.29	1.30	0.99	38
Ratio tiny (capped at 100) / large-v2	2.65	2.90	0.92	39

Which languages are nearest to Kazakh?

By training hours, the languages nearest to Kazakh are Hindi (12 hours), Armenian (13), Bosnian (11), Icelandic and Macedonian (16 each). By large-v2 WER they are Icelandic (38.2%), Marathi (38.3%), Kannada (37.0%), Maori (38.5%) and Afrikaans (36.7%). By the shape of the whole curve over six sizes they are Icelandic, Armenian, Swahili, Nepali and Marathi. No Turkic language appears in any of these lists. Of the Turkic languages, Table 13 contains only Azerbaijani, Turkish and Uzbek besides Kazakh (Table 5). Azerbaijani is the closest of them, 14th of 80 by the shape of the curve, and its ratio of tiny to large-v2 WER (3.98) is the sixth nearest to Kazakh's (4.38). Uzbek, with 0.3 training hours, never falls below 50% and barely improves with size (from 105.2% to 90.2%).

Table 5. The Turkic languages in Table 13 of [1]

Language	Training hours	WER tiny → large-v2, %	First size below 50%
Kazakh	12	165.2 / 109.2 / 70.3 / 48.8 / 43.8 / 37.7	medium
Azerbaijani	47	93.1 / 76.4 / 49.1 / 33.1 / 28.7 / 23.4	small
Uzbek	0.3	105.2 / 114.0 / 107.7 / 109.6 / 93.3 / 90.2	never
Turkish	4,333	42.5 / 27.5 / 15.9 / 10.4 / 9.4 / 8.4	tiny

Discussion

The forecast against the data. The first claim (F1) was wrong by one size: Kazakh falls below 50% at medium, although only by 1.2 points; the author was right that tiny, base and small are all above the line. The second claim (F2), about capacity going to languages with more data, cannot be tested with published tables, because they show results, not what happens inside the model. What the tables do show sits uneasily with it: in relative terms the smallest models lose most on languages with a lot of data, and Kazakh's relative loss is average. The third claim (F3) was wrong: Kazakh's neighbours by hours and by error are Hindi, Armenian and Icelandic rather than Turkic languages, and Uzbek, with 0.3 hours, is in a different situation from Kazakh. When the results were shown to him on 16 October 2026, the author wrote: "The forecast about errors almost came true: tiny, base and small have more than 50% errors, but medium already has 48.8%, so the result turned out slightly better than the forecast. Kazakh loses more than all other languages from the largest size to the smallest, most likely because of the small amount of training data. Next to it by hours there were not Turkic languages but Hindi, Armenian and Icelandic, and Uzbek has only 0.3 hours and never drops below 50%. This suggests that the amount of data matters more than the language itself". His statement that Kazakh loses more than all other languages holds for the loss in percentage points (Table 3).

What the results mean. Three findings stand out. First, Kazakh is usable, by the author's threshold, only from the medium model (769M parameters) upward, and medium is a common crossing point for languages with similar training hours. Second, Kazakh is the language that loses the most percentage points when the model shrinks, but this is mostly what a language with 12 hours and a high starting error is expected to lose: relative to its best result and to its training hours, Kazakh is ordinary from base upward. Third, the only clear Kazakh-specific excess is in the tiny model, where the error is above 100% and the model inserts words that were not said. A tiny model is therefore not a usable choice for Kazakh, and a developer who needs an offline Kazakh tool has to plan for at least the medium model, which according to the official repository needs about 5 GB of video memory [3].

The results agree with the author's conclusion that the amount of data matters more than the language itself: Kazakh sits among languages with similar training hours, most of them unrelated to it, and close to the trend line for its hours. They cannot prove it. Training hours go together with other differences, such as script, language family and the amount of text on the internet, and the Whisper authors themselves noted that distantly related languages with unique scripts tend to do worse than their hours predict [1]. Kazakh is written in Cyrillic, like Russian, and is not Indo-European. Kazakh WER is somewhat above the line predicted by its hours at every size (Table 4), so a small effect of the language itself cannot be ruled out.

Implications for Kazakhstan. The author drew the following conclusion: "The model's inaccuracy is linked to a lack of data, not to the language itself. So more high-quality recordings of Kazakh speech need to be collected. This can be done by ISSAI (Nazarbayev University), which has already collected about 1,200 hours, 100 times more than Whisper had, and also by the state, radio and television. I want everyone, including me and my brother, to have an equal opportunity to learn and use AI in Kazakh as in Russian". The size he cites agrees with the published figure: the 2026 extension of ISSAI's KSC2 corpus contains 1,151 hours [5], about 96 times Whisper's 12 Kazakh hours. Open Kazakh corpora many times larger than Whisper's 12 hours therefore already exist [4-6]. Whisper's training audio was collected from the internet [1], and the paper does not say whether such corpora were used. More recordings help only when they reach the models people actually use, either through the training data of new models or by fine-tuning existing ones.

Limitations. The study uses published numbers only, so it shares their limits. Table 13 gives no confidence intervals and no number of test sentences per language, so it is not known whether the 1.2-point margin of Kazakh medium below 50% is larger than the sampling error of FLEURS. WER for Kazakh was computed after a basic text standardization designed for all non-English languages [1], and WER does not show which kinds of error the model makes. The training hours are Whisper's own count of weakly labelled audio, printed as rounded numbers, and 15 languages in Table 13 have no hours and are left out of the analyses that need them. The comparison between large-v2 and the smaller models mixes size with a different training recipe, although the comparison with large, which differs only in size, gives the same picture. Finally, newer models such as large-v3 are not in Table 13, and the results describe Whisper only, not speech recognition in general.

Conclusion

In the published Whisper results, Kazakh speech recognition falls below 50% WER only from the medium model upward, one size earlier than the author forecast. In percentage points Kazakh loses more than any other comparable language when the model shrinks from the largest to the smallest size, but relative to its own best result and to its 12 training hours, its loss is ordinary except in the smallest model. Its nearest languages by training hours and by error are not Turkic but languages with similar training hours, such as Hindi, Armenian and Icelandic. These results support the author's view that the amount of training data matters more than the language itself, and they suggest that Kazakh speech recognition in small models will improve mainly when much more transcribed Kazakh speech, which already exists in open corpora, enters the training of widely used models.

Data availability

All data used in this study are published in the Whisper paper [1] (Table 13, p. 24, and Figure 11, p. 27). The extracted tables, the per-language results and the analysis code are available from the author on request.

Author contributions and use of AI tools

The author chose the research question from a list of five directions prepared by his mentors, read and summarised the literature on model compression and on Kazakh language models, wrote his forecast before any result was calculated, compared it with the results, and wrote the introductory paragraph and the conclusion for Kazakhstan; The numbers were extracted from the published paper and analysed by his mentoring team. No number, quotation or reference was produced by a language model: every statistic is computed by the

analysis scripts from the published tables, and each statement about a cited work was checked against its published text.

Conflicts of interest

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

References:

1. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision // arXiv preprint. 2022. arXiv:2212.04356v1. https://arxiv.org/abs/2212.04356

2. Conneau A., Ma M., Khanuja S., Zhang Y., Axelrod V., Dalmia S., Riesa J., Rivera C., Bapna A. FLEURS: Few-shot learning evaluation of universal representations of speech // arXiv preprint. 2022. arXiv:2205.12446. https://arxiv.org/abs/2205.12446

3. OpenAI. Whisper, GitHub repository. https://github.com/openai/whisper (accessed: 24.10.2026).

4. Mussakhojayeva S., Khassanov Y., Varol H.A. KSC2: An industrial-scale open-source Kazakh speech corpus // Proceedings of Interspeech 2022. P. 1367-1371.

5. IS2AI (Institute of Smart Systems and Artificial Intelligence, Nazarbayev University). KSC2++: Enhanced Kazakh speech corpus with LLM-based transcription restoration, GitHub repository. https://github.com/IS2AI/KSC2 (accessed: 24.10.2026).

6. Khassanov Y., Mussakhojayeva S., Mirzakhmetov A., Adiyev A., Nurpeiissov M., Varol H.A. A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL). 2021. https://arxiv.org/abs/2009.10334

7. Marchisio K., Dash S., Chen H., Aumiller D., Üstün A., Hooker S., Ruder S. How does quantization affect multilingual LLMs? // Findings of the Association for Computational Linguistics: EMNLP 2024. https://aclanthology.org/2024.findings-emnlp.935

8. Togmanov M., Mukhituly N., Turmakhan D., et al. KazMMLU: Evaluating language models on Kazakh, Russian, and regional knowledge of Kazakhstan // arXiv preprint. 2025. arXiv:2502.12829. https://arxiv.org/abs/2502.12829

9. Assan R. KazSwitch: A controlled benchmark for Kazakh-Russian code-switching in multilingual language models // Central Asian Scientific Journal. 2026. No. 2(30), Vol. 3. P. 65-72.

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 12

Alexander Gorevoy
WHICH INEQUALITY STATISTICS SURVIVE AN ADMINISTRATIVE BOUNDARY CHANGE? EVIDENCE FROM KAZAKHSTAN, 2010–2023 16

Rakhimov Djalil
DOES THE SHIFT TO CASHLESS PAYMENTS GO TOGETHER WITH FASTER GROWTH IN HOUSEHOLD CREDIT? EVIDENCE FROM THE REGIONS OF KAZAKHSTAN, 2023-2025 28

Kazhgerey Asan
HOW MANY FOOD BASKETS DOES AN AVERAGE WAGE BUY? WAGES AND FOOD PRICES IN THE 20 REGIONS OF KAZAKHSTAN, 2024-2026 39

Tagimova Alua
WHO WAS LEFT BEHIND? EDUCATION AND FINANCIAL INCLUSION IN KAZAKHSTAN AND THREE CENTRAL ASIAN NEIGHBOURS, 2021-2024 46

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 53

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 66

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 82

Yerboluly Madi
DID IMPORTED CARS REALLY GET CHEAPER? CHINESE CARS AND THE AVERAGE CUSTOMS VALUE OF PASSENGER CARS IMPORTED INTO KAZAKHSTAN, 2015-2025 93

Салимбаев Ербол Аманжолович
АНАЛИЗ ИНВЕСТИЦИОННОЙ АКТИВНОСТИ КАК ИНДИКАТОР ИЗМЕНЕНИЙ ПОТЕНЦИАЛЬНОГО СПРОСА НА КОНСАЛТИНГОВЫЕ УСЛУГИ 101

Rambergenov Ramil
REGIONAL NEW-FIRM FORMATION IN UZBEKISTAN AFTER THE 2017 REFORMS: PERSISTENCE IN SCALE, CONVERGENCE IN INTENSITY, AND RECOMPOSITION TOWARD TRADE 109

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 118

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 126

Cheng Che
BUILD, BUY, OR PARTNER? ENTRY MODE SELECTION AND RISK GOVERNANCE FOR WIND POWER INVESTMENT IN KAZAKHSTAN **135**

Wang Liang
FROM CAPACITY EXPANSION TO SYSTEM VALUE: COORDINATED DEVELOPMENT OF RENEWABLE ENERGY, GRID FLEXIBILITY AND ENERGY STORAGE IN KAZAKHSTAN **148**

Chen Jun
CONSTRUCTION-PERIOD CREDIT RISK MANAGEMENT FOR LARGE INFRASTRUCTURE PROJECT FINANCE IN KAZAKHSTAN **161**

Alniyaz Danial
HOW MUCH FUEL DOES A HYBRID REALLY SAVE? A SAME-MODEL COMPARISON OF HYBRID AND GASOLINE CARS IN US FUEL ECONOMY **175**

Assem Yergali, Maral Assanova
DIGITAL TRANSFORMATION AND SUSTAINABLE BUSINESS MODEL INNOVATION IN SMALL AND MEDIUM-SIZED ENTERPRISES: A SYSTEMATIC REVIEW OF DYNAMIC CAPABILITIES, VALUE CREATION, AND THE CREDIBILITY GAP IN EMERGING ECONOMIES **182**

Бейбіт Инара Алибекқызы, Едилов Саназар
ВОДА ИЗ ТУМАНА ИЛИ ВОДА ИЗ МОРЯ: ПОРОГОВАЯ СТОИМОСТЬ, ПРИ КОТОРОЙ ТУМАНОУЛОВИТЕЛИ КОНКУРЕНТОСПОСОБНЫ В МАНГИСТАУСКОЙ ОБЛАСТИ **195**

Zhenisbekova Aknur
DID A TENGE DEPOSIT OR THREE KAZAKHSTANI SHARES PROTECT SAVINGS FROM INFLATION? REAL RETURNS OF BANK DEPOSITS, KAZATOMPROM, KEGOC AND KASPI.KZ, 2020-2025 ..... **207**

Абилдаева Аселя
ОРГАНИЗАЦИОННАЯ ГОТОВНОСТЬ ВУЗОВ КАЗАХСТАНА К ВНЕДРЕНИЮ ЭНДАУМЕНТ-ФОНДОВ: КОНЦЕПТУАЛЬНАЯ РАМКА НА ОСНОВЕ ОБЗОРА ЛИТЕРАТУРЫ И АНАЛИЗА ДОКУМЕНТОВ **215**

Yang Juntao
DIGITAL ESG GOVERNANCE AND NEXT-GENERATION DIGITAL COMPETENCY AS DRIVERS OF SUCCESSION IN FAMILY AGRICULTURAL ENTERPRISES: A CONCEPTUAL FRAMEWORK ... **229**

Kairbek Mariya Islamkyzy, Kylbaev Erlan Serikovich
STATE SUPPORT FOR THE DIGITALIZATION OF MANUFACTURING IN THE REPUBLIC OF KAZAKHSTAN: AN ANALYSIS OF ITS COVERAGE OF MARKETING AND SALES ACTIVITIES **242**

Tilep Aiaru
WHEN DID IT EXPORTS TAKE OFF? ICT SERVICE EXPORTS OF KAZAKHSTAN, ARMENIA, GEORGIA AND THREE NEIGHBOURS, 2015-2024 ..... **255**

Anuarbek Murager
DID THE TENGE STOP FOLLOWING OIL? BRENT PRICES AND THE TENGE-DOLLAR RATE BEFORE AND AFTER FEBRUARY 2022 ..... **266**

Kaiyrbek Zhanadil
HOW ACCURATELY CAN OBSERVABLE LISTING DATA ESTIMATE FAIR PRICES FOR USED CARS IN KAZAKHSTAN? A HEDONIC MODEL OF 2,546 ONLINE LISTINGS **276**

Wen Yali
PERCEIVED ORGANIZATIONAL SUPPORT AND TURNOVER INTENTION AMONG KNOWLEDGE WORKERS IN A CHINESE PRIVATE TRAINING INSTITUTION: THE MEDIATING ROLE OF AFFECTIVE ORGANIZATIONAL COMMITMENT **289**

Құрманалиева Дана
ИСТОЧНИКИ ФИНАНСИРОВАНИЯ ИННОВАЦИОННЫХ ПРОЕКТОВ И КРИТЕРИИ ВЫБОРА ФИНАНСОВОЙ МОДЕЛИ **300**

ТЕХНИКАЛЫҚ ҒЫЛЫМДАР - ТЕХНИЧЕСКИЕ НАУКИ - TECHNICAL SCIENCE

Nurmakhan Almira
ADVANTAGES OF FIBER OPTIC COMMUNICATION OVER TRADITIONAL COPPER CABLES **307**

Kozhabai Bereke
SECURITY CHALLENGES IN IOT-BASED TELECOMMUNICATION NETWORKS **315**

Azamatuly Yerassyl
WHY AI-GENERATED CODE FAILS A BENCHMARK: A FAILURE ANALYSIS OF THREE OPEN LANGUAGE MODELS ON 30 HUMANEVAL+ TASKS **324**

Nurzhan Sanzhar
A FEW SHELLS, A FEW OPERATORS: WHERE LOW EARTH ORBIT IS CROWDED IN AN ALTITUDE-INCLINATION MAP OF 16,620 ACTIVE SATELLITES **333**

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 **339**

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 **350**

Ануар Әлімжан Бахытжанұлы, Сексенбаева Айша Курманбековна
СРАВНИТЕЛЬНЫЙ АНАЛИЗ СТАНДАРТОВ И МЕТОДИК ОБУЧЕНИЯ КИБЕРБЕЗОПАСНОСТИ: NICE FRAMEWORK, ECSF И ПРАКТИКО- ОРИЕНТИРОВАННЫЕ ТЕХНОЛОГИИ **356**

Mukhamedzhan Yerganat
GROWTH OF GITHUB DEVELOPER COMMUNITIES IN KAZAKHSTAN, UZBEKISTAN AND KYRGYZSTAN, 2020-2026: EVIDENCE FROM THE GITHUB INNOVATION GRAPH **366**

Kashkinov Alden
WHAT DECIDES HOW LONG A LANDER WORKS ON ANOTHER WORLD? SURFACE LIFETIMES OF 55 LANDERS ON THE MOON, VENUS, MARS, TITAN AND SMALL BODIES IN THE GCAT CATALOGUE, 1966-2025 ..... **378**

Kenesbek Miras
HOW MANY HERONIAN TRIANGLES HAVE A GIVEN LONGEST SIDE? A COMPLETE COUNT UP TO 2000 AND THE RIGHT TRIANGLES HIDDEN INSIDE **390**

Amirlan Seitkadyrov, Akezhan Omashev
AUTOMATIC MULTIPLE-CHOICE QUESTION GENERATION USING ARTIFICIAL INTELLIGENCE **401**

Idrissov Amirzhan
WHERE DO SMALL OPEN LANGUAGE MODELS BREAK ON OLYMPIAD MATHEMATICS? DOMAIN, DIFFICULTY AND MODEL SIZE ON OMNI-MATH **410**

Маматаева Рамина Адилгазыевна, Бидуәлі Алмат Алматұлы
ОБЗОР И СРАВНЕНИЕ ПОПУЛЯРНЫХ КРОССПЛАТФОРМЕННЫХ ФРЕЙМВОРКОВ ДЛЯ РАЗРАБОТКИ МОБИЛЬНЫХ ПРИЛОЖЕНИЙ **415**

Қасен Ғауһар Ғалымжанқызы, Алкеева Дана Мирасовна
КОМПЬЮТЕРЛІК КӨРУДІ ПАЙДАЛАНЫП,МЕХАНИКАЛЫҚ ДЕНЕ ЗАҢЫН АНЫҚТАУ АЛГОРИТМДЕРІН ӘЗІРЛЕУ **421**

Khalilzade Seymur Fazil oglu
ASSESSING THE IMPACT OF ELECTRIC VEHICLES ON ENERGY DEMAND IN HIGH-RISE RESIDENTIAL BUILDINGS **434**

Igilikov Adilkhan Igilikovich
PRE-TRAINED REPRESENTATIONS OF PRIOR-ADMISSION HISTORY IMPROVE ONE-YEAR, BUT NOT IN-HOSPITAL, MORTALITY PREDICTION IN ICU PATIENTS WITH CHRONIC DISEASE: A MIMIC-IV STUDY **449**

Nurgali Dias, Nurgali Miras, Akashov Aslan
PREDICTIVE TRAFFIC MANAGEMENT: EXPLORING SOLUTIONS TO TRAFFIC CONGESTION IN ASTANA, KAZAKHSTAN **461**

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 ..... **467**

Amangeldy Nurbolat
HOW MUCH OF KAZAKHSTAN'S URBAN BUILT STOCK PREDATES ITS SEISMIC CODES, AND WHERE DOES THAT STOCK SIT RELATIVE TO HAZARD? **477**

Torekhan Sabrina
WHERE ALMATY WALKS ITS OWN WAY: INFORMAL SHORTCUTS IN OPENSTREETMAP AND THE DETOURS THEY AVOID **483**

Ailan Salybekova
PARENTAL SUPPORT, PARENTAL CONTROL, STRESS AND EARLY ALCOHOL USE IN ADOLESCENT GIRLS: A NARRATIVE REVIEW WITH A FOCUS ON KAZAKHSTAN **488**

Yerboluly Madi
WHY IS KAZAKHSTAN'S GENDER GAP AT THE TOP OF CHESS SO SMALL? A BIRTH-COHORT COMPARISON ON THE FIDE RATING LIST **496**

Omashev Akezhan
REGIONAL HEALTH AND SOCIAL INDICATOR PATTERNS AND PCA- BASED MORTALITY MODELING IN KAZAKHSTAN, 2011–2024. A TWO- PART EXPLORATORY ECOLOGICAL ANALYSIS **501**

Мадиярова Жанел Биржановна
ДИПФЕЙК ТЕХНОЛОГИЯСЫНЫҢ ҚОҒАМДЫҚ ПІКІРГЕ ЫҚПАЛЫ ЖӘНЕ ОНЫ ҚҰҚЫҚТЫҚ РЕТТЕУ ТӘЖІРИБЕСІ **511**

Bolatov Sanzhar
HOW MUCH DOES KAZAKH SPEECH RECOGNITION LOSE WHEN THE MODEL GETS SMALLER? KAZAKH AGAINST 82 LANGUAGES IN THE PUBLISHED WHISPER RESULTS **521**

Ilya Tsoy
WHO PAYS MORE TOKENS FOR THE SAME SENTENCE? RUSSIAN, KOREAN AND ENGLISH IN FOUR LANGUAGE MODEL TOKENIZERS ON 1,012 FLORES-200 SENTENCES **530**

Assan Ramazan
WHICH NAME DO READERS STILL USE? FORMER NAMES OF RENAMED KAZAKHSTANI CITIES IN RUSSIAN AND KAZAKH WIKIPEDIA PAGE VIEWS, 2015–2026 ..... **543**

Nigmadilova Zhasmin
WHAT REPLACES PARABENS? THE PARABEN-FREE LABEL ACROSS PRODUCT CATEGORIES IN THE OPEN BEAUTY FACTS DATABASE **556**

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 ТҰРҒЫСЫНАН ТАЛДАУ **561**

Мұқатаева Айзат Амантайқызы, Сапарханова Мадина Төлегенқызы, Окенова Аружан Муратхановна, Тұрсынғазыева Ақерке Ерболатқызы
ШӘКӘРІМ ФИЛОСОФИЯСЫНДАҒЫ ЕРКІНДІК ЖӘНЕ МОРАЛЬДЫҚ ТАҢДАУ МӘСЕЛЕСІ **571**

Komekova Saida
DO UNHAPPY REVIEWERS MIX IN MORE RUSSIAN? STAR RATINGS AND KAZAKH-RUSSIAN CODE-SWITCHING IN 175,002 ONLINE REVIEWS **580**

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 **593**

Abdrakhmanova Madina
DOES A MEDAL BUY ATTENTION? WIKIPEDIA VIEWS OF KAZAKHSTAN'S ATHLETES BEFORE AND AFTER THE PARIS 2024 OLYMPIC **597**

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 ..... **605**

Zhenis Adilzhan
WHICH OLYMPIAD PROBLEMS DO STUDENTS TAKE TO THE BOARD? CHOICES AND DEFENCES IN THE 8TH-GRADE LEAGUE OF KARAGANDY MATHBATTLES 2026 ..... **612**

Базарбаева Айнур Бектуровна
ДЕТСКИЙ ФОЛЬКЛОР И ПОДВИЖНЫЕ ИГРЫ КАК ИНСТРУМЕНТЫ ФОРМИРОВАНИЯ ОСНОВ ПРАВОВОЙ КУЛЬТУРЫ У ДЕТЕЙ СТАРШЕГО ДОШКОЛЬНОГО ВОЗРАСТА **619**

БИОЛОГИЯ ҒЫЛЫМДАР - БИОЛОГИЧЕСКИЕ НАУКИ - BIOLOGICAL SCIENCES

Bazhenova Diana
GROWTH DYNAMICS OF STERLET (Acipenser ruthenus L.) REARED IN A RECIRCULATING AQUACULTURE SYSTEM **622**

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 **629**

ЕСТЕСТВЕННЫЕ НАУКИ - ЖАРАТЫЛЫСТАНУ ҒЫЛЫМДАРЫ - NATURAL SCIENCES

Нургельдина Асель Ермековна, Серикбаева Жанерке Алимжановна
ШЕКАРАСЫЗ ФИЗИКА: ӘР ЕЛДЕГІ ФИЗИКАЛЫҚ БІЛІМДІ САЛЫСТЫРУ **636**

Torekhan Sabrina
IS ALMATY A 15-MINUTE CITY? WALKING ACCESS TO SIX EVERYDAY PLACES FROM 170,838 BUILDINGS IN THE CITY'S EIGHT DISTRICTS **643**

Онгарбек Махабат Саятқызы
CU-PT ЖӘНЕ NI-PT ҚОРЫТПАЛАРЫНЫҢ ФИЗИКАЛЫҚ ҚАСИЕТТЕРІНЕ РЕТТЕЛУ ҮДЕРІСІНІҢ ЖӘНЕ ҚҰРЫЛЫМДЫҚ ӨЗГЕРІСІНІҢ ӘСЕРІН РЕТТЕУ **651**

Нургельдина Асель Ермековна, Әбдуәлі Мерей Ералықызы
ҮЙКЕЛІС КҮШІНІҢ ТАБИҒАТТАҒЫ РӨЛІ **659**

Ossipov Vladislav
MANY REPORTS, WEAK EVIDENCE? EVIDENCE LEVELS FOR MICROBIAL PLASTIC DEGRADATION IN THE 2,535 RECORDS OF PLASTICDB **664**

Нургельдина Асель Ермековна, Кабдулаева Жанель Батырбековна
РАДИОАКТИВТІЛІК. АЛЬФА, БЕТА, ГАММА СӘУЛЕЛЕРІНІҢ ТЕГІ МЕН ЗАҢДЫЛЫҚТАРЫ. ЯДРОЛЫҚ РЕАКЦИЯ ЖӘНЕ ЖАРТЫЛАЙ ЫДЫРАУ ПЕРИОДЫ **673**

Kashkinov Alden
DOES AN ACTIVE SUN BRING SATELLITES DOWN SOONER? SOLAR ACTIVITY AND THE RE-ENTRY OF UNCONTROLLED OBJECTS IN LOW EARTH ORBIT, 1990-2025 ..... **678**

Demin Mark
HOW MUCH LESS OXYGEN CAN THE WARMING CASPIAN SEA HOLD? THE LOSS OF OXYGEN SOLUBILITY IN THE NORTHERN, MIDDLE AND SOUTHERN CASPIAN, 1982-2022 ..... **684**