ПӘНДЕР АРАЛЫҚ ҒЫЛЫМДАР
МЕЖДИСЦИПЛИНАРНЫЕ НАУКИ
INTERDISCIPLINARY SCIENCES

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. No. 4 (32), Vol. 2. pp. 485–493
DOI: https://doi.org/10.5281/zenodo.24314821
Journal website: https://cajournal.kz

UDC 004.8:811.512.122

HOW MUCH WORSE DO LANGUAGE MODELS SOLVE THE SAME MATH PROBLEMS IN KAZAKH THAN IN RUSSIAN? A RE-ANALYSIS OF PUBLISHED GSM8K RESULTS

Idrissov Amirzhan
Kokshetau Bilim-Innovation Lyceum, Kokshetau, Kazakhstan

Idrissov Amirzhan — 11th-grade student, Kokshetau Bilim-Innovation Lyceum, Kokshetau, Kazakhstan.

© Idrissov Amirzhan, 2026

Abstract: This paper asks how much less often a language model solves the same mathematics problem in Kazakh than in Russian, and whether the difference shrinks for larger or stronger models. Instead of running models, the study re-analyses the accuracies that three papers published for the same 1,319 GSM8K word problems in Kazakh and Russian machine translation and compares them with the author's forecast, made before he was shown any GSM8K result, that Kazakh accuracy would be about 10% lower. All 10 general-purpose settings evaluated with worked examples scored lower in Kazakh, by 5.3 to 40.2 percentage points (median 14.5): the three strongest in Russian lost 5.3, 9.0 and 11.7 points, close to the forecast, while Qwen2.5-Omni-3B solved 11.4% of the Kazakh problems and 51.6% of the Russian ones. The gap was larger for models with lower Russian accuracy (Spearman $\rho = -0.82$), smaller for the 7- than for the 3-billion-parameter model of one line (29.4 against 40.2 points) but not steadily smaller with size in a Kazakh-adapted family, and smaller with reasoning in all five comparisons of a reasoning and a non-reasoning variant. For 9 of the 10 general settings the mathematics gap was numerically smaller than the gap in reading comprehension, which is consistent with the author's reasoning that the difficulty lies in understanding the Kazakh text, although the two benchmarks also differ in format and translation.

Keywords: Kazakh language, Russian language, language models, mathematical reasoning, GSM8K, low-resource languages

Introduction

In Kazakhstan the same school mathematics is taught in Kazakh and in Russian, and students increasingly use language models to check their homework and to prepare for olympiads. The author's earlier study in this journal found that small open models are far less reliable on olympiad geometry than on algebra [1]. A question that matters just as much for Kazakh-speaking students is whether a model that solves a problem in Russian can solve the same problem when it is written in Kazakh.

Kazakh is a low-resource language for language models. KazMMLU, a benchmark of 23,000 school and university questions from Kazakhstan, reported in its results that models perform "slightly better in Russian than in Kazakh" and in its conclusion that they "perform much better in Russian than in Kazakh" [2]. Its Kazakh and Russian questions, however, are different questions taken from different sources, so a difference between the two languages there mixes the effect of the language with a difference in the questions themselves [3, 4]. A study in this journal used KazMMLU to separate Kazakhstan-specific knowledge from universal school subjects and found that on universal subjects, including mathematics, the Kazakh and Russian scores of three small

Qwen2.5 models were within about 1.5 percentage points of each other, while Kazakh History showed a gap of 13 to 18 points [4].

A cleaner test uses the same problems in both languages. The author made this point when he chose the topic: he wrote that he wanted to check "how much the language of the problem statement affects the solution of mathematics problems" and proposed "to take the same problems in two languages, to compare the influence of the language itself and not the different difficulty of the problem sets". Such data already exist. The Institute of Smart Systems and Artificial Intelligence (ISSAI) at Nazarbayev University released Kazakh and Russian machine translations of the 1,319 test problems of GSM8K [5], a benchmark of grade-school mathematics word problems that need several steps of arithmetic [6]. Three recent papers report the accuracy of many models on both versions [3, 7, 8]. None of them, however, was written to answer the question asked here: two of them introduce Kazakh models [7, 8] and the third studies fine-tuning [3].

This paper therefore re-analyses the published results and asks: on the same mathematics problems, how much lower is a model's accuracy in Kazakh than in Russian, does the gap shrink as models become larger or stronger, and does it look like a problem of language or of calculation?

The author's forecast

The direction was chosen by the author on 29 October 2026 from two directions proposed by his mentor. When he was offered it, the mentor's message included three numbers from KazMMLU [2]: the best models reached 76.9% in Kazakh and 81.4% in Russian on high-school subjects (GPT-4o and DeepSeek), and Qwen2.5-7B-Instruct averaged 47.8% over all levels without worked examples. The author had been shown no GSM8K result in Kazakh or Russian. On 2 November 2026 he wrote his forecast:

"I think that in Kazakh the share of correct answers will be about 10 percent lower than in Russian. I think that during training small models saw more texts and mathematics problems in Russian, so the problem statements and terms in Kazakh will be harder for them to understand. But the formulas and calculations themselves are the same, so I do not expect a very large difference."

The forecast contains a number and a reason. "10 percent lower" can be read as 10 percentage points or as 10% of the Russian accuracy, so both are reported below. The reason predicts that the difficulty lies in understanding the Kazakh text rather than in the arithmetic, which can be checked by comparing the gap in mathematics with the gap in a pure reading task for the same models.

Materials and methods

Problems. All results in this paper refer to the same set of problems: the GSM8K test set [6] in the Kazakh and Russian versions released by ISSAI [5]. The dataset card describes them as "the machine-translated Kazakh and Russian versions" of the GSM8K test set; the dataset viewer shows 1,319 problems in each language, and the first problem is the same in both (the problem about Janet's ducks, answer 18). The Qolda paper, by the ISSAI group, states that its authors made the translations with "GPT-4o-guided machine translation" [7]; the Qolda-AVL paper uses "adapted versions from our previous work", that is, from [7] [8]; and the third paper describes the benchmark as "parallel Kazakh/Russian machine-translated versions of the original test sets released by ISSAI" with 1,319 problems [3]. No paper gives a version number, so it is assumed, as these descriptions indicate, that all three used the same translated set.

Published results. Three sources report accuracy on both versions for the same models (Table 1 lists the settings).

– Arystanbekov et al. [8], Table 4, with ten settings: six of general-purpose models (Qwen2.5-Omni 3B and 7B, and Qwen3-VL-4B and Qwen3-Omni-30B-A3B, each released as separate Instruct and Thinking versions) and four of three Kazakh-adapted models (KazLLM-1.0-8B, Qolda with reasoning on and off, and Qolda-AVL-5B). GSM8K was evaluated with "a three-shot CoT approach", that is, with three worked examples that show step-by-step solutions.

– Arystanbekov et al. [7], Table 4: the same evaluation design (three-shot, step-by-step) for Qwen3-4B and InternVL3.5-4B, each with reasoning ("thinking") mode on and off, and two Kazakh-adapted models of about 8 billion parameters, Sherkala-8B-Chat and HordeVision. Three

rows of this table (KazLLM-1.0-8B and the final Qolda model with reasoning on and off) are identical to rows in [8] and were counted once; the earlier training stages of Qolda that the table also lists are not included. This table is published as an image; its values were copied from the image and checked twice.

– Mussa et al. [3], Table 5: eleven released checkpoints of 7 to 14 billion parameters of different kinds, including base, instruction-tuned and reasoning-distilled models, evaluated "zero-shot", that is, without worked examples, with greedy decoding, at most 512 generated tokens, and the last number in the answer taken as the model's answer. Only the "Base" columns, before the authors' own fine-tuning, are used.

The first two sources come from the same research group and use the same pipeline; they are treated as one evaluation design ("three-shot") and the third as another ("zero-shot"). The two designs are never mixed in one average. Five Kazakh-adapted models (KazLLM-1.0-8B, Sherkala-8B-Chat, HordeVision, Qolda and Qolda-AVL-5B) are reported separately from general-purpose models, because training on Kazakh text is expected to change the gap. The Qolda-AVL model card [9] reports the same benchmark for Qolda-AVL models of 5, 9 and 34 billion parameters; it is used only for the size comparison within that family, because the card does not describe its evaluation settings and its 5B values (85.75% in Kazakh and 90.90% in Russian) differ slightly from those in [8] (87.79% and 92.65%).

Reading comprehension. To compare the gap in mathematics with a gap that is purely about language, the same tables [7, 8] report Belebele, a reading comprehension benchmark of 900 multiple-choice questions on short passages [10]. Belebele is "fully parallel" across 122 language variants [10]: the same passages and questions were translated by human translators, so it is the same test in Kazakh and Russian.

Measures. For each model the gap is Russian accuracy minus Kazakh accuracy, in percentage points (pp). The relative gap is the gap divided by Russian accuracy. A 95% interval for the gap was computed from the published accuracies and the number of problems (1,319; 900 for Belebele), treating the two languages as independent samples. Because the problems are the same in both languages, and a problem that is hard in one language is likely to be hard in the other, an interval that used this pairing would be narrower, so the interval given errs on the side of caution. The forecast was compared with the results by asking how many gaps lie between 5 and 15 pp, that is, within 5 points of the forecast 10. Rank correlations are Spearman's ρ. None of the accuracies in Tables 1–4 comes from a model run by the author or his mentor: each is copied from a published table, and every other number is calculated from them.

Results

The gap in every model

Table 1 shows all 27 settings. Every one of the 16 settings evaluated with worked examples scored lower in Kazakh than in Russian, and in every case the 95% interval lies above zero. Among the 10 general-purpose settings the gap ranged from 5.3 to 40.2 pp, with a median of 14.5 pp; in relative terms Kazakh accuracy was 5.5% to 77.9% lower than Russian (median 15.9%). Five of the ten gaps lie between 5 and 15 pp, close to the forecast, and the other five are larger than 15 pp. Read as a relative loss, four of the ten lie between 5% and 15%. No general model had a gap below 5 pp.

Table 1 – Accuracy (%) on the same 1,319 GSM8K problems in Kazakh and Russian, and the gap (Russian minus Kazakh, pp, with a conservative 95% interval); rows in each block are ordered by gap

Model (parameters)	Source	Kazakh	Russian	Gap, pp
General models, three-shot				
Qwen2.5-Omni-3B (3B)	[8]	11.38	51.59	40.2 (37.0 to 43.4)
InternVL3.5-4B, no reasoning (4B)	[7]	60.52	89.97	29.5 (26.4 to 32.5)
Qwen2.5-Omni-7B (7B)	[8]	30.96	60.39	29.4 (25.8 to 33.1)
Qwen3-4B, no reasoning (4B)	[7]	66.26	89.06	22.8 (19.7 to 25.9)
InternVL3.5-4B, reasoning (4B)	[7]	77.91	93.05	15.1 (12.5 to 17.8)
Qwen3-VL-4B-Instruct (4B)	[8]	75.59	89.53	13.9 (11.1 to 16.8)

Model (parameters)	Source	Kazakh	Russian	Gap, pp
Qwen3-4B, reasoning (4B)	[7]	82.56	94.22	11.7 (9.3 to 14.1)
Qwen3-Omni-30B-A3B-Instruct (30B)	[8]	85.43	94.39	9.0 (6.7 to 11.2)
Qwen3-VL-4B-Thinking (4B)	[8]	85.43	92.56	7.1 (4.8 to 9.5)
Qwen3-Omni-30B-A3B-Thinking (30B)	[8]	90.59	95.90	5.3 (3.4 to 7.2)
Kazakh-adapted models, three-shot				
KazLLM-1.0-8B (8B)	[8]	59.14	75.99	16.9 (13.3 to 20.4)
HordeVision (8B)	[7]	73.25	89.97	16.7 (13.8 to 19.6)
Qolda, no reasoning (4.3B)	[8]	77.66	90.27	12.6 (9.9 to 15.4)
Sherkala-8B-Chat (8B)	[7]	64.56	77.06	12.5 (9.1 to 15.9)
Qolda, reasoning (4.3B)	[8]	84.60	93.62	9.0 (6.7 to 11.4)
Qolda-AVL-5B (5B)	[8]	87.79	92.65	4.9 (2.6 to 7.1)
Released checkpoints, zero-shot				
Phi-4 (14B)	[3]	3.8	49.2	45.4 (42.5 to 48.3)
Qwen2.5-7B (7B)	[3]	19.4	57.8	38.4 (35.0 to 41.8)
Ministral-8B-Instruct (8B)	[3]	30.2	57.9	27.7 (24.1 to 31.3)
DeepSeek-R1-Distill-Qwen-7B (7B)	[3]	7.5	30.1	22.6 (19.7 to 25.5)
Aya-23-8B (8B)	[3]	1.4	22.3	20.9 (18.6 to 23.2)
Hunyuan-MT-Chimera-7B (7B)	[3]	3.3	21.9	18.6 (16.2 to 21.0)
Qwen3-8B (8B)	[3]	40.4	52.7	12.3 (8.5 to 16.1)
Llama-3-8B (8B)	[3]	8.4	17.5	9.1 (6.6 to 11.6)
Mistral-7B-v0.3 (7B)	[3]	2.0	8.7	6.7 (5.0 to 8.4)
Gemma-7B (7B)	[3]	6.2	11.3	5.1 (3.0 to 7.2)
KazLLM-1.0-8B, Kazakh-adapted (8B)	[3]	22.9	23.8	0.9 (−2.3 to 4.1)

Figure 1 plots Kazakh against Russian accuracy for every setting. The solid line marks equal accuracy and the dashed line marks Kazakh 10 points lower; filled circles are general models and open circles Kazakh-adapted models evaluated with three worked examples [7, 8], and crosses are released checkpoints evaluated without examples [3]. If the forecast were exactly right, the points would lie on the dashed line. The strong models sit close to it; the weak ones fall far below it.

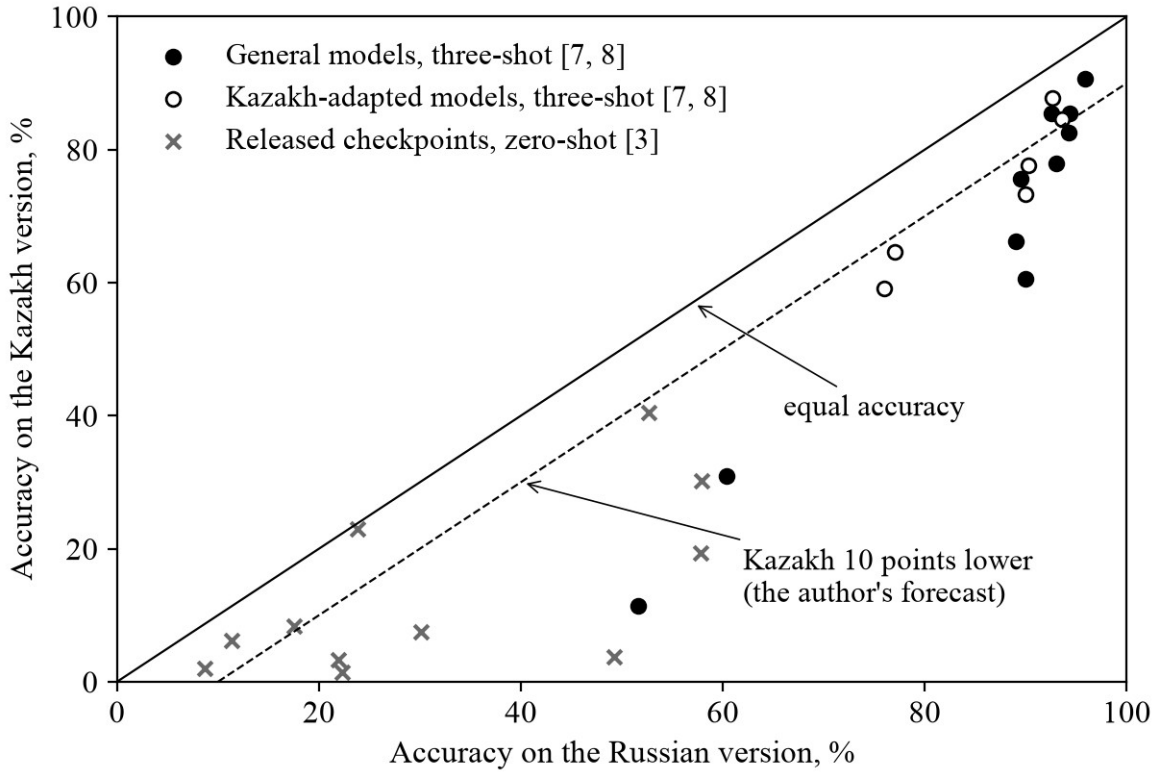


Figure 1 - Kazakh against Russian accuracy on the same GSM8K problems

Weaker models lose more

The size of the gap followed the strength of the model. Among the 10 general settings, the lower a model's Russian accuracy, the larger its Kazakh gap (ρ = −0.82). The three settings with Russian accuracy above 94% lost 5.3, 9.0 and 11.7 pp, while the two Qwen2.5-Omni models, the weakest in Russian, lost 29.4 and 40.2 pp. Part of this association is built into the measure: a model near 95% in Russian cannot lose as many points as a model near 50%. Qwen2.5-Omni-3B solved 51.6% of the Russian problems but only 11.4% of the same problems in Kazakh, a Kazakh accuracy about one fifth of its Russian accuracy.

The released checkpoints evaluated without examples (Table 1, bottom) show the same problem in a sharper form. For the ten general checkpoints Russian accuracy was low (8.7% to 57.9%), and in Kazakh seven of them solved fewer than 10% of the problems, so their relative gaps were very large (median 70.8%). Their gaps in percentage points are smaller for the weakest checkpoints, which had little accuracy left to lose: Mistral-7B solved 8.7% in Russian and 2.0% in Kazakh. Among these ten checkpoints the rank correlation between Russian accuracy and the gap is positive (ρ = 0.79), the opposite sign to the three-shot setting, which shows that percentage-point gaps between models at the floor and models near the ceiling should not be read as one scale. Mussa et al. inspected the answers and found that Kazakh outputs were "more often truncated before a final answer" and that correct Kazakh reasoning was more often scored wrong by the answer extraction, "so part of the gap reflects response form, not reasoning" [3]. For Phi-4, 89.5% of the Kazakh answers were empty, against 23.7% of the Russian answers [3]. The only zero-shot checkpoint with no clear gap was KazLLM-1.0-8B, which was trained on Kazakh text, and it scored low in both languages (22.9% and 23.8%).

Model size and reasoning

Table 2 compares smaller and larger models from the same paper. Going from 3 to 7 billion parameters in the Qwen2.5-Omni line raised Kazakh accuracy from 11.4% to 31.0% and Russian from 51.6% to 60.4%, and the gap fell from 40.2 to 29.4 pp. Between Qwen3-VL-4B and the 30-billion-parameter Qwen3-Omni the gap fell from 13.9 to 9.0 pp without reasoning and from 7.1 to 5.3 pp with it. These two are different model lines, so size is not the only difference between them; Qwen3-Omni-30B-A3B is also a mixture-of-experts model, which, as "A3B" in its name indicates, uses about 3 billion parameters for each token.

Table 2 – The gap (pp) for smaller and larger models in one paper [8]: one model line (Qwen2.5-Omni) and two different lines (Qwen3-VL-4B and Qwen3-Omni-30B-A3B)

Model line	Smaller model	Gap	Larger model	Gap
Qwen2.5-Omni	3B	40.2	7B	29.4
Qwen3-VL-4B / Qwen3-Omni-30B, no reasoning	4B	13.9	30B	9.0
Qwen3-VL-4B / Qwen3-Omni-30B, reasoning	4B	7.1	30B	5.3

In all five comparisons of a reasoning and a non-reasoning variant, the gap was smaller with reasoning, because reasoning raised Kazakh accuracy much more than Russian accuracy (Table 3). In three of them (InternVL3.5-4B, Qwen3-4B and Qolda) the same model was evaluated with the mode switched on and off, in [7] with different sampling settings for the two modes; the other two are separately released Instruct and Thinking versions of Qwen3-VL-4B and Qwen3-Omni-30B-A3B. For Qwen3-4B, for example, reasoning added 16.3 points in Kazakh and 5.2 in Russian, and the gap halved from 22.8 to 11.7 pp.

© Central Asian Scientific Journal, 2026

Table 3 – Gap with and without reasoning ("thinking"): three models with the mode switched on and off and two Instruct/Thinking pairs [7, 8]

Model	Gap without reasoning	Gap with reasoning	Kazakh gain	Russian gain
InternVL3.5-4B	29.5	15.1	+17.4	+3.1
Qwen3-4B	22.8	11.7	+16.3	+5.2
Qwen3-VL-4B	13.9	7.1	+9.8	+3.0
Qwen3-Omni-30B-A3B	9.0	5.3	+5.2	+1.5
Qolda (Kazakh-adapted)	12.6	9.0	+6.9	+3.4

Models adapted to Kazakh also showed a gap, from 4.9 to 16.9 pp (Table 1), a range that overlaps that of the general models; because none of them is compared here with the general model it was built from, these data do not show how much adaptation changed the gap. The smallest, 4.9 pp for Qolda-AVL-5B, is about as small as that of the 30-billion-parameter general model with reasoning. Within the Qolda-AVL family the model card reports gaps of 5.2, 8.9 and 3.8 pp for the 5, 9 and 34 billion parameter models [9], so in this family a larger model did not steadily reduce the gap.

Mathematics gap compared with reading gap

Russian accuracy was within 3 points of English accuracy for seven of the ten general settings and at most 9.1 points lower (Qwen2.5-Omni-3B). Kazakh accuracy was 6.7 to 49.3 points below English [7, 8]. The loss is therefore specific to Kazakh and is not shared by Russian.

Table 4 compares the gap in mathematics with the gap in reading comprehension on Belebele for the same 16 settings. For 9 of the 10 general settings the mathematics gap was numerically smaller than the reading gap. In some of them the difference is within the error of measurement, for example InternVL3.5-4B without reasoning (29.5 against 30.3 pp), and for Qwen2.5-Omni-3B the mathematics gap was slightly larger (40.2 against 39.0 pp). Across all 16 settings the two gaps went together ($\rho = 0.87$): a model that read Kazakh poorly also solved Kazakh mathematics poorly. For the Kazakh-adapted models the picture was mixed: in three of the six settings (KazLLM-1.0-8B, Qolda without reasoning and Qolda-AVL-5B) the mathematics gap was the larger one.

Table 4 – Gap (pp) in mathematics (GSM8K, machine-translated) and in reading comprehension (Belebele, human-translated) for the same settings [7, 8]

Model	GSM8K gap	Belebele gap
Qwen2.5-Omni-3B	40.2	39.0
Qwen2.5-Omni-7B	29.4	34.4
InternVL3.5-4B, no reasoning	29.5	30.3
InternVL3.5-4B, reasoning	15.1	32.4
Qwen3-4B, no reasoning	22.8	28.7
Qwen3-4B, reasoning	11.7	17.6
Qwen3-VL-4B-Instruct	13.9	20.9
Qwen3-VL-4B-Thinking	7.1	13.6
Qwen3-Omni-30B-A3B-Instruct	9.0	11.8
Qwen3-Omni-30B-A3B-Thinking	5.3	8.9
KazLLM-1.0-8B	16.9	13.8
HordeVision	16.7	23.6
Sherkala-8B-Chat	12.5	12.6
Qolda, no reasoning	12.6	10.0
Qolda, reasoning	9.0	9.5
Qolda-AVL-5B	4.9	4.2

One published measurement points to a possible mechanical reason why Kazakh is harder for at least one of these models. On a text sample of about 21,000 words in each language, the tokenizer of Qwen3-VL, which splits text into the pieces a model reads and writes, needed on average 4.78 pieces per Kazakh word, against 2.46 per Russian word and 1.26 per English word [8]. This was not measured on mathematics problems; if it holds for them, a Kazakh problem and its solution would be roughly twice as long for this model as the Russian ones.

© Central Asian Scientific Journal, 2026

Discussion

The forecast was right about the direction and about the strongest models, and wrong about the weak ones. Every general model solved fewer problems in Kazakh than in Russian, and for the three strongest settings the gap was close to the forecast 10 points (5.3, 9.0 and 11.7 pp). For weaker models the gap was much larger, up to 40 points, and for the weakest released checkpoints, including the 14-billion-parameter Phi-4, Kazakh accuracy fell almost to zero, partly because answers were empty or cut off [3]. Small models are the ones a student can run on an ordinary computer, and they had some of the largest gaps.

The author's reason is consistent with the results, though they do not prove it. He expected the difficulty to lie in understanding the Kazakh problem statement rather than in the arithmetic because, in his words, "the formulas and calculations themselves are the same". In 9 of 10 general settings the gap in mathematics was numerically smaller than the gap in reading the same language, and the two gaps rose and fell together. The effect of reasoning mode points the same way: when a model first writes out its reasoning, its Kazakh accuracy rises much more than its Russian accuracy, as if the extra steps help it to work out what the Kazakh text asks. The published audit of zero-shot answers adds that part of the Kazakh loss is in the form of the answer: Kazakh answers were cut off or not recognised more often [3]. The tokenizer measurement offers one more possible reason, if Kazakh solutions are also about twice as long in tokens [8]. These explanations are consistent with the data but were not tested here.

The results differ from those of the earlier study in this journal, where Kazakh and Russian scores on universal school subjects in KazMMLU were within about 1.5 points [4]. The difference may reflect that the Kazakh and Russian KazMMLU questions are different questions, that KazMMLU is a multiple-choice test, and that [4] tested other models; the present data cannot separate these.

For a student the practical conclusion is simple. With a small model, asking a mathematics question in Kazakh can cost many more wrong answers than asking the same question in Russian; turning on the model's reasoning mode or using a stronger model was associated with a smaller loss, but did not remove it.

Limitations

The Kazakh and Russian problems are machine translations (GPT-4o-guided [7]); the dataset card reports no human check and no error rate, so a translation error can make a problem harder in one language. The author read the Kazakh versions of five problems that a 7-billion-parameter model had solved in Russian but not in Kazakh in an exploratory run (see Author contributions) and found wording that could be misread or made clearer in three of them, for example "rabbit" in the singular where the problem compares with the number of rabbits; five problems are too few to estimate an error rate. If the three papers used different versions of the translations, comparisons between papers are affected. Belebele, used for comparison, was translated by people, so the comparison of the two gaps also compares two kinds of translation. The results come from three papers, two of them by the same group, and the three-shot results used the generation settings recommended in each model card [8] or fixed settings for each mode [7], not one setting for all models, so single-point differences between models should not be over-interpreted. The 95% intervals are conservative because they ignore that the problems are the same in both languages. No published table gives Kazakh and Russian results on the same mathematics problems for models below 3 billion parameters, including the 1.5B and 3B Qwen2.5 instruction models of the author's earlier study [1], so the smallest models could not be studied from published results. Two size comparisons within one model line are available: Qwen2.5-Omni 3B and 7B [8] and the Qolda-AVL family [9], whose evaluation settings are not described; the comparisons between 4B and 30B models mix size with differences between model lines. Finally, only final answers were scored by the original authors, and the paper relies on their scoring.

Conclusion

On the same 1,319 grade-school mathematics problems, every general-purpose language model in the published results scored lower in Kazakh than in Russian. The gap was about 5 to 12

percentage points for the three strongest models, close to the author's forecast of 10, and up to 40 points for weaker ones. It was smaller with reasoning mode in all five comparisons and smaller for the 7B than for the 3B model of one line, but it did not fall steadily with size in the one Kazakh-adapted family with three sizes, and Kazakh-adapted models also showed gaps of 4.9 to 16.9 points. The gap in mathematics was numerically smaller than the gap in reading comprehension for 9 of the 10 general models, which is consistent with the author's view that the main difficulty for these models is understanding Kazakh, not doing the calculation, although the two benchmarks also differ in format and translation.

Note: original quotations

The author's words were written in Russian and are translated in the text by meaning, with nothing added or removed. Topic choice, 29 October 2026: «Мне интересно проверить насколько язык условия влияет на решение математических задач. Я предлагаю взять одни и те же задачи на двух языках чтобы сравнить именно влияние языка а не разную сложность подборок». Forecast, 2 November 2026: «Я думаю на казахском доля правильных ответов будет примерно на 10 процентов ниже чем на русском. Я думаю что при обучении маленькие модели видели больше текстов и математических задач на русском, поэтому условия и термины на казахском им будет сложнее понять. Но сами формулы и вычисления одинаковые, поэтому слишком большой разницы я не ожидаю».

Data availability

Every value used here is openly published in [2–5, 7–9]. A table that lists each copied value with its source table and page, and the program that recomputes Tables 1–4 and Figure 1 from that table, can be obtained from the author.

Author contributions and use of AI tools

The author chose this direction from two proposed by his mentor, proposed comparing the same problems in two languages (29 October 2026), wrote the forecast (2 November 2026) and, on 5 November 2026, checked the Kazakh translations of five problems (see Limitations). Between 2 and 4 November 2026, after the forecast, the mentor ran three small open models of 1.5, 3 and 7 billion parameters on 250 problems of the same GSM8K test set, and on 4 November shared the results with the author; they also showed much lower accuracy in Kazakh, and the five problems the author checked came from that run, but its results are not used here, because the paper was designed to rest only on published results. His mentor in an educational mentoring programme, working with a generative AI assistant (Claude, Anthropic), found the published results, copied the numbers, wrote the program that computes the tables and the figure, translated the author's words and drafted the English text. On 15 November 2026, after reading the full draft of 13 November, the author confirmed that his words and their translation are rendered correctly; he takes responsibility for the content of the paper.

Conflicts of interest

The study received no external funding. The author attends a paid educational mentoring programme; his mentor's contribution is described above. The author declares no other conflicts of interest.

References:

1. Idrissov A. Where do small open language models break on olympiad mathematics? Domain, difficulty and model size on Omni-MATH // Central Asian Scientific Journal. – 2026. – No. 4(32), Vol. 1. – Pp. 410–414. – DOI: 10.5281/zenodo.23758386.

2. 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 (Volume 1: Long Papers). – 2025. – Pp. 14403–14416. – DOI: 10.18653/v1/2025.acl-long.701.

3. Mussa A., Tuimebayev Z., Mansurova M. et al. 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. – DOI: 10.3390/info17090921.

4. Zhumagali Y. Does a bigger model close the gap? Kazakh-language accuracy on Kazakhstan-specific versus universal school subjects in small open-weight language models // Central Asian Scientific Journal. – 2026. – No. 4(32), Vol. 1. – Pp. 593–596. – DOI: 10.5281/zenodo.23758467.

5. Institute of Smart Systems and Artificial Intelligence (ISSAI). GSM8k_Kazakh_Russian: dataset card [Electronic resource]. – URL: https://huggingface.co/datasets/issai/GSM8k_Kazakh_Russian (accessed: 07.11.2026).

6. Cobbe K., Kosaraju V., Bavarian M. et al. Training verifiers to solve math word problems [Electronic resource] // arXiv preprint. – 2021. – arXiv:2110.14168. – URL: https://arxiv.org/abs/2110.14168 (accessed: 15.11.2026).

7. Arystanbekov B., Nurimanov A., Maxutov A. et al. Qolda: A small vision-language model for the Kazakh language // IEEE Access. – 2026. – Vol. 14. – Pp. 46392–46414. – DOI: 10.1109/ACCESS.2026.3676919.

8. Arystanbekov B., Maxutov A., Nurimanov A. et al. Extending a vision-language model with audio understanding: Introducing Qolda-AVL for the Kazakh language // Big Data and Cognitive Computing. – 2026. – Vol. 10, No. 6. – Art. 192. – DOI: 10.3390/bdcc10060192.

9. Institute of Smart Systems and Artificial Intelligence (ISSAI). Qolda-AVL-5B: model card [Electronic resource]. – URL: https://huggingface.co/issai/Qolda-AVL-5B (accessed: 07.11.2026).

10. Bandarkar L., Liang D., Muller B. et al. The Belebele benchmark: A parallel reading comprehension dataset in 122 language variants // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2024. – Pp. 749–775. – DOI: 10.18653/v1/2024.acl-long.44.

© Central Asian Scientific Journal, 2026