IRSTI 14.15.01

N. Baizhanov[1*], A. Sabyrov[2], A. Makhmutova[3,4], S. Kadyrov[4]
[1,2]National Testing Centre, Ministry of Science and Higher Education, Astana, Kazakhstan
[3]SDU University, Kaskelen, Kazakhstan
[4]New Uzbekistan University, Tashkent, Uzbekistan
*e-mail: nurbaizhanov@gmail.com

[1]ORCID 0009-0008-5302-9858, [2]ORCID 0009-0006-1470-5433, [3]ORCID 0000-0002-8597-7667, [4]ORCID 0000-0002-8352-2597

LARGE LANGUAGE MODELS IN EDUCATIONAL MEASUREMENT OF KAZAKH LANGUAGE PROFICIENCY

This study evaluates the performance of large language models (LLMs) in assessing Kazakh language proficiency within the context of the Unified National Test (UNT) in Kazakhstan. The primary objective is to examine the accuracy, error patterns, and psychometric characteristics of five state-of-the-art LLMs—Gemini 2.5 Pro Preview, Claude 3.7 Sonnet, Deepseek R1, Qwen, and Llama 3.1-405B-Instruct—on 138 multiple-choice questions (MCQs) from the 2024 UNT Kazakh language test. The methodology involved a zero-shot evaluation with standardized prompts, ensuring no external data access, and employed statistical analyses, including Cochran's Q test, McNemar's tests, and Generalized Estimating Equations (GEE) logistic regression, to assess model performance across difficulty levels and linguistic topics. Results indicate that Gemini achieved the highest accuracy (90.6%), significantly outperforming other models, while Llama showed the lowest (37.7%). Performance varied by difficulty and topic, with Gemini excelling across all categories and others showing strengths in specific areas like complex linguistic reasoning. The study highlights the potential of LLMs for educational assessment in low-resource languages like Kazakh, while identifying gaps in model optimization, fairness, and reliability, necessitating targeted fine-tuning and culturally relevant data curation.

Keywords: Large Language Models, Educational Assessment, Kazakh Language, Unified National Test, Pedagogical Measurement, Artificial Intelligence.

Introduction

Education in Kazakhstan places strong emphasis on assessing student proficiency through the Unified National Test (UNT), which plays a central role in university admission. In recent years, increasing numbers of test takers have chosen to complete the UNT in Kazakh, reflecting both policy priorities and the broader use of the national language in education [1]. At the same time, advances in artificial intelligence, particularly large language models (LLMs), have created new opportunities for educational measurement. LLMs such as GPT-3.5, GPT-4, and other contemporary models are increasingly applied to tasks including automated grading, item generation, and diagnostic assessment [2]. These technologies promise efficiency and scalability, but their performance varies across domains, item types, and especially languages.

In language proficiency assessment, particularly for less-resourced languages, key questions remain about the validity, fairness, and reliability of LLM outputs [3]. For the Kazakh language category of the UNT, no published study has systematically evaluated LLM performance on multiple-choice questions. This gap is significant, since Kazakh proficiency is not only an academic requirement but also a cultural and policy priority. Therefore, this study aims to evaluate the performance of several state-of-the-art LLMs on the Kazakh language multiple-choice portion of the UNT. By examining their accuracy, error patterns, and psychometric characteristics, the study provides evidence on the potential and limitations of LLMs in supporting educational assessment in Kazakhstan.

Literature review

The Role of AI and LLMs in Educational Assessment and Testing. Over the past decade, advancements in artificial intelligence (AI) have dramatically transformed the educational assessment landscape. LLMs—characterized by their billions of parameters and training on vast, heterogeneous datasets—are now capable of generating human-like text and offering real-time, adaptive feedback on a variety of tasks. These systems facilitate automated test item generation, adaptive scoring, and personalized formative feedback, thereby reducing the burden on educators and allowing for scalable assessment solutions [4, 5]. For example, state-of-the-art models such as GPT-4 have demonstrated strong performance on numerous high-stakes examinations by replicating expert reasoning and delivering coherent explanations, a capability that signifies a major shift from traditional manual assessment methods [6, 7].

LLMs have been deployed to automatically generate draft test items based on textbook content or curricular themes. By extracting and synthesizing key concepts from large corpora, these models produce multiple-choice questions, true/false items, and even open-ended prompts that align with predetermined learning objectives [8]. Such automated generation not only accelerates the development cycle of test items but also enables educators to focus on qualitative review and refinement rather than on initial content creation [8]. In computerized adaptive testing environments, LLMs contribute further by assisting in real-time calibration of question difficulty based on examinee responses, thus tailoring tests dynamically to individual ability levels and ensuring that assessments are both engaging and diagnostically informative [9].

Additionally, the integration of LLMs into automated scoring systems has been transformative. Natural language processing (NLP) techniques within these models are used to evaluate essays and open-ended responses by considering not only surface features such as vocabulary usage and syntactic complexity but also deeper semantic coherence [9, 10]. The immediate feedback provided by automated systems helps students identify their learning gaps and enables teachers to design targeted interventions, thus democratizing access to high-quality, personalized assessments across diverse educational settings [9, 11]. LLMs also play a supportive role in ensuring test security through anomaly detection during online exams, flagging irregular response patterns that may indicate academic dishonesty [5, 11]. Together, these developments illustrate how AI-driven systems are gradually shifting educational assessment from traditional, labor-intensive processes to dynamic, technology-enabled paradigms that are scalable, consistent, and highly adaptable [9, 12].

LLM Performance on Multiple-Choice Question Answering in High-Stakes Examinations. A significant stream of research has focused on the performance of LLMs in multiple-choice question (MCQ) settings, especially on high-stakes examinations used in professional licensure and academic selection. Early iterations of generative models such as GPT-3.5 exhibited inconsistent performance on domain-specific MCQs, largely due to limitations in their reasoning abilities and the necessity for extensive prompt engineering [6]. However, subsequent models, notably GPT-4, have achieved remarkable improvements, attaining scores that place them within the top decile on examinations such as the Uniform Bar Exam and the United States Medical Licensing Examination (USMLE) [7, 13].

Researchers have investigated not only raw accuracy but also the calibration of confidence estimates in LLM outputs when answering medical and legal MCQs. Calibration analyses for GPT-4, for instance, have indicated that the model's probability assignments closely track actual correctness, a feature considered critical in high-stakes testing environments where score margins have significant implications [14]. In parallel, chain-of-thought prompting techniques—which encourage the model to articulate intermediate reasoning steps—have been shown to boost accuracy by making the decision process more transparent and interpretable [11, 15].

Another focus of research has been on the sensitivity of LLMs to the ordering of answer options. Positional bias, where models may disproportionately favor options presented in certain positions, has been documented, and techniques such as majority voting over multiple reordered presentations have been developed to mitigate these biases [16, 17]. These ensemble methods not only help in refining the final answer but also improve reliability by confirming that the chosen answer is robust across different prompt formulations [18, 19]. They further underscore the necessity of continuous refinements in prompting and calibration to ensure that LLMs' performance is both optimal and ethically reliable for high-stakes decision-making [20, 21].

Research on LLMs in Low-Resource Languages. While breakthroughs in LLM performance have predominantly emerged from high-resource language contexts, there is a growing recognition that these models must be extended to accommodate low-resource languages. Kazakh, a language characterized by agglutinative morphology, vowel harmony, and complex syntactic structures, epitomizes the challenges encountered in this area [22]. Existing multilingual LLMs, although designed to handle several languages simultaneously, often exhibit diminished performance when processing Kazakh due to limited training data and tokenization schemes optimized for Indo-European languages [23].

Recent efforts have sought to address these challenges by developing culturally and linguistically tailored benchmarks such as the KazMMLU dataset. This benchmark comprises thousands of multiple-choice questions designed specifically to assess competencies in Kazakh and Russian, thereby reflecting the bilingual nature of the region's educational system [23]. Evaluations using KazMMLU have revealed that while proprietary models like GPT-4o and DeepSeek V3 perform robustly on Russian test items—often achieving accuracies well above 75%—their performance on Kazakh items lags significantly, highlighting a critical imbalance in model efficacy across languages [22, 23].

The performance gap is not solely a function of data scarcity but also arises from the inherent linguistic complexities present in Kazakh. Standard subword segmentation techniques, which work effectively for languages with relatively simple morphology, struggle to capture the nuances of Kazakh grammar and syntax, resulting in poorer understanding and lower quality outputs [5, 24]. Moreover, research has indicated that prompts written in English often lead to higher accuracy than those written entirely in Kazakh, suggesting that current models are biased toward high-resource language representations that were more abundantly available during training [21, 22]. To mitigate these limitations, research in transfer learning from closely related languages such as Russian and specialized prompt engineering techniques is being explored, though these measures have only partially closed the performance gap [11].

Furthermore, the evaluation benchmarks developed so far for low-resource languages have primarily focused on narrow NLP tasks such as sentiment analysis or named entity recognition, with limited attention to the comprehensive reasoning abilities required in educational assessments [22, 24]. This gap underscores the necessity for developing comprehensive datasets and evaluation protocols that not only assess factual recall but also complex problem-solving and critical thinking in the context of educational measurement for languages like Kazakh [23].

Challenges Related to Fairness, Validity, and Reliability in AI-Driven Educational Assessment. A significant area of concern in the deployment of LLMs in educational assessment relates directly to the issues of fairness, validity, and reliability. These concerns are particularly acute when the training data for LLMs is derived from sources that may not represent the full diversity of student populations, potentially leading to biased outcomes that disadvantage underrepresented groups [25]. The opaqueness that characterizes many deep learning models makes it challenging to diagnose the specific sources of bias, which can range from demographic inequities to cultural misinterpretations, consequently affecting the fairness and reliability of automated scoring processes [26, 27].

Fairness concerns are not limited to demographic biases but extend to the methodological biases that occur in automated scoring systems. Frequently, these systems rely on proxy measures such as essay length, vocabulary complexity, or punctuation frequency, which might correlate with writing proficiency in a biased manner [8, 10]. Such methods often fail to capture more abstract elements of quality, including critical thinking, creativity, and coherence, thereby potentially reducing construct validity in high-stakes exams [9, 11]. These challenges are compounded when high-stakes decisions such as licensure or university admissions are based on scores generated by AI systems, where even small biases can have disproportionate impacts on individual educational trajectories [8].

Reliability also emerges as a critical challenge due to the “black-box” nature of many LLM architectures. The uncertainty inherent in deep neural network algorithms raises questions about the consistency of scores produced by AI systems over time. Various studies have attempted to address this by using ensemble methods, such as majority voting or calibration techniques, to ensure that LLM outputs remain stable across multiple runs and different prompt formulations [11, 16]. Although such techniques can improve the overall reproducibility of the results, concerns persist regarding the interpretability of the decision-making processes used by these models [12, 21].

Moreover, academic integrity issues further complicate the adoption of LLM-based assessments. The ease with which students can leverage LLMs to generate entire responses or essays—sometimes without proper attribution—has sparked debate regarding whether such practices constitute plagiarism or represent a legitimate aid in learning [23, 25]. These issues necessitate a balanced approach wherein the benefits of rapid, automated feedback are weighed against the risk of undermining authentic learning and critical thinking skills [8]. Finally, ongoing interdisciplinary collaborations are required to develop robust ethical frameworks and standardized guidelines that ensure transparent, fair, and reliable use of LLMs in educational assessment [26, 28].

Gaps in the Literature and Justification for Investigating LLM Performance in Kazakh Language Testing (UNT). Despite significant progress in applying LLMs to educational assessments in high-resource language settings, several critical gaps justify a closer investigation of model performance for low-resource languages such as Kazakh, especially in the context of the Unified National Testing (UNT) system. First, while numerous studies have detailed the impressive achievements of models like GPT-4 in handling MCQ tasks and adaptive testing in English, there is a marked paucity of research addressing the performance of these systems in languages that exhibit complex morphological features and limited digital resources [23].

Benchmark evaluations such as those based on the KazMMLU dataset have provided early insights into this gap. These evaluations demonstrate that while proprietary models (e.g., GPT-4o, DeepSeek V3) achieve commendable performance in Russian—a language for which relatively abundant training data exist—their accuracies drop significantly when processing Kazakh test items [23]. Furthermore, studies have indicated that English-language prompts lead to superior performance relative to Kazakh prompts, suggesting that the underlying training paradigms and tokenization methods are not sufficiently optimized for languages with complex agglutinative morphology [22]. Such disparities highlight the urgent need for dedicated research on prompt engineering, data curation, and model adaptation that better capture the linguistic nuances of Kazakh [5, 24].

Another notable gap is the lack of research on the longitudinal impacts of integrating LLM-driven assessments in low-resource settings. Although many studies have focused on short-term performance metrics, the long-term educational outcomes—such as changes in student learning trajectories, retention of critical thinking skills, and overall academic growth—remain poorly understood, particularly in the context of culturally-specific and linguistically diverse environments like Kazakhstan [5,11]. There is also minimal insight into how LLM-driven scoring systems interact with locally defined educational standards and curricula, an issue that becomes critical when assessments are used for high-stakes decisions such as university admissions and licensure [23].

Moreover, existing benchmarks for low-resource languages are often limited to basic NLP tasks such as sentiment analysis or named entity recognition, without addressing the comprehensive reasoning and domain-specific knowledge required for educational assessments [22, 24]. The current literature thus falls short of providing a complete picture of LLM performance on intricate tasks such as MCQ answering, essay grading, and cognitive diagnostics in a low-resource language context [21, 22]. This situation is exacerbated by the lack of culturally relevant training data and evaluation protocols that reflect the authentic educational experiences of Kazakh-speaking students [23].

The need for comprehensive, bilingual benchmarks that cover a wide range of subjects—from STEM to social sciences—is therefore imperative. Such benchmarks would not only enable a detailed assessment of LLM performance but also inform the development of tailored model architectures and fine-tuning strategies that bridge the performance gap between high- and low-resource languages [23]. Importantly, these efforts should also address issues of fairness and reliability by incorporating mechanisms for differential item functioning analysis and cross-cultural validation [26, 28]. Ultimately, investigating LLM performance on Kazakh language testing is not merely a technical challenge but also an ethical imperative to ensure that the benefits of AI-driven assessments are distributed equitably across diverse linguistic communities [23].

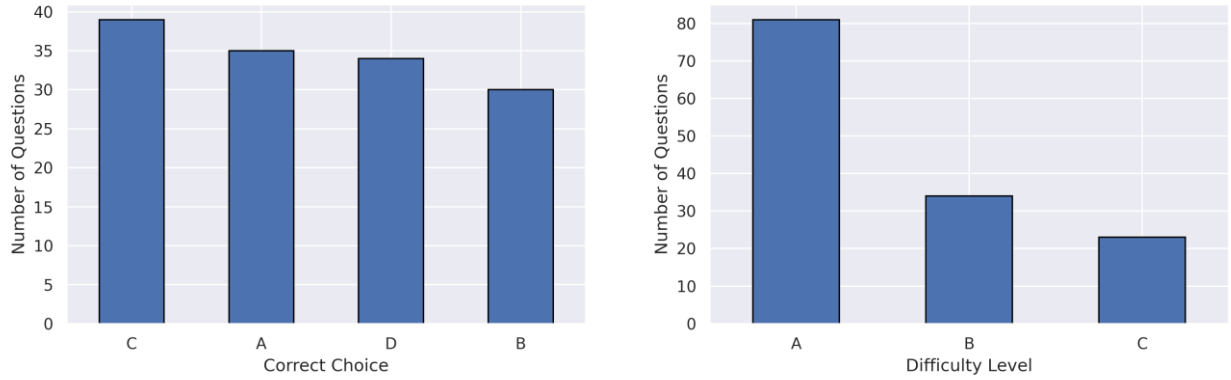
In summary, the integration of AI and large language models into educational assessment systems marks a transformative shift in how learning is measured and evaluated. LLMs have demonstrated exceptional prowess in generating content, calibrating adaptive tests, and scoring high-stakes examinations with performance metrics that are increasingly competitive with those of human experts [4, 7]. However, while high-resource language contexts have benefited immensely from these advancements, significant disparities persist when these systems are applied to low-resource languages, particularly Kazakh.

The literature reviewed here underscores not only the remarkable progress that has been achieved with MCQ answering and adaptive testing in high-stakes examinations but also identifies critical shortcomings in model performance, especially in handling complex morphological nuances and culturally specific content [22, 23]. Moreover, concerns regarding fairness, validity, and reliability in automated assessment systems remain paramount, necessitating continued research into ethical frameworks, prompt engineering, and culturally tailored data curation [26, 28].

Materials and methods

In UNT, the Kazakh Language component consists of 40 MCQs designed to assess reading comprehension, grammar, vocabulary, and linguistic reasoning across multiple subdomains. The methodology encompasses data preparation, model selection, experimental design, and evaluation procedures to address the research questions on LLMs' ability to handle low-resource language assessments, with a focus on improving preparation for the UNT in Kazakhstan's educational context. All LLMs relied on internal reasoning via prompt engineering, with no access to online searches or external corpora, ensuring that their performance reflected inherent model capabilities rather than external retrieval.

To construct the evaluation dataset, we collected 138 UNT Kazakh Language MCQs from the 2024 test cycle, originally presented in Kazakh. These questions were used in their original language to ensure fidelity to the source material and to assess the multilingual and low-resource language capabilities of the selected LLMs. This approach avoided potential distortions from translation and allowed for a more authentic evaluation of model performance in a real-world national testing setting. Due to confidentiality constraints, sample questions are not included here, but representative UNT Kazakh language test items are available from the National Test Center (2024) [29].



a) Distribution of Correct Answer Choices b) Distribution of Questions by Difficulty Level

Figure 1 - Descriptive Statistics of the UNT Kazakh Language Dataset

Questions were categorized into three difficulty levels—A (easy), B (moderate), and C (hard)—based on historical UNT performance data, expert judgment, and expected percent-correct thresholds (Figure 1a). The dataset contained 81 level A questions, 34 level B questions, and 23 level C questions. This distribution ensured a balanced representation across difficulty levels and aligned with national testing standards. The distribution of correct answers was relatively balanced across alternatives, indicating no systematic bias in item construction (Figure 1b).

The dataset further covered a wide range of linguistic domains, distributed across 22 topics. The most represented topics included punctuation (n = 22), main and secondary sentence parts (n = 18), synonyms, antonyms, and homonyms (n = 10), orthography of compound and hyphenated words (n = 9), and assimilation rules (n = 9). Less frequent but still relevant categories included verb forms (n = 6), sentence phrases (n = 5), adjectives, interjections, and nouns (n = 4 each), with specialized categories such as archaic words, neologisms, and professional terms represented by one item each. This broad coverage ensured that the dataset reflected the curricular diversity of the Kazakh language test.

Five state-of-the-art LLMs were selected for evaluation based on their multilingual performance, reasoning capabilities, and accessibility through the OpenRouter platform. The models represent a diverse set of providers and architectures. Gemini 2.5 Pro Preview (Google) offers a million-token context window and is particularly strong in mathematical reasoning. Qwen (Alibaba) uses a Mixture of Experts architecture with 235 billion parameters in total, of which 22 billion are active during inference. Deepseek R1 (DeepSeek) is specialized in technical reasoning tasks, while Claude 3.7 Sonnet (Anthropic) emphasizes high performance in natural language proficiency. Finally, Llama 3.1-405B-Instruct (Meta AI) represents an open-source, large-scale model suitable for broad research applications. For performance evaluation, each LLM was provided with a standardized prompt directing it to select the single best answer to a UNT multiple-choice question (e.g., “Answer: C”). The evaluation was conducted in a zero-shot setting, ensuring that the models had no prior exposure to the test items. Each prompt contained the full question text and four answer choices, with explicit instructions for the model to reply using only the corresponding letter (A, B, C, or D). Model outputs were processed using regular expressions to extract the selected letter, with non-standard responses resolved by interpreting the first character or marking them as invalid.

To analyze model performance, we applied a combination of nonparametric and regression-based statistical techniques suitable for binary outcomes. Cochran’s Q test was used to evaluate whether accuracy differences across the five LLMs were statistically significant when answering the same set of questions. For pairwise comparisons, McNemar’s tests with Bonferroni correction were conducted to identify which models differed significantly from each other. In addition, a

Generalized Estimating Equations (GEE) logistic regression model with an exchangeable correlation structure was employed to account for within-question clustering and to estimate the relative performance of each model compared to a baseline. Together, these methods provided both omnibus and pairwise insights into LLM performance, while controlling for dependence across repeated measures on the same test items.

Results and discussion

Figure 2 presents the overall accuracy of the five evaluated LLMs on the UNT Kazakh Language dataset. Gemini achieved the highest score, correctly answering 90.6% of the questions, which indicates near-human level performance in this low-resource language setting. Claude followed with 61.6%, demonstrating moderate proficiency but with noticeable gaps compared to Gemini. Deepseek and Qwen performed similarly, with accuracies of 53.6% and 49.3%, respectively, reflecting only partial capability in handling the linguistic and reasoning demands of the assessment. Llama obtained the lowest score at 37.7%, suggesting that even large-scale open-source architectures may struggle without targeted optimization for low-resource languages. These differences highlight both the potential and the limitations of current LLMs in supporting high-stakes national testing contexts.

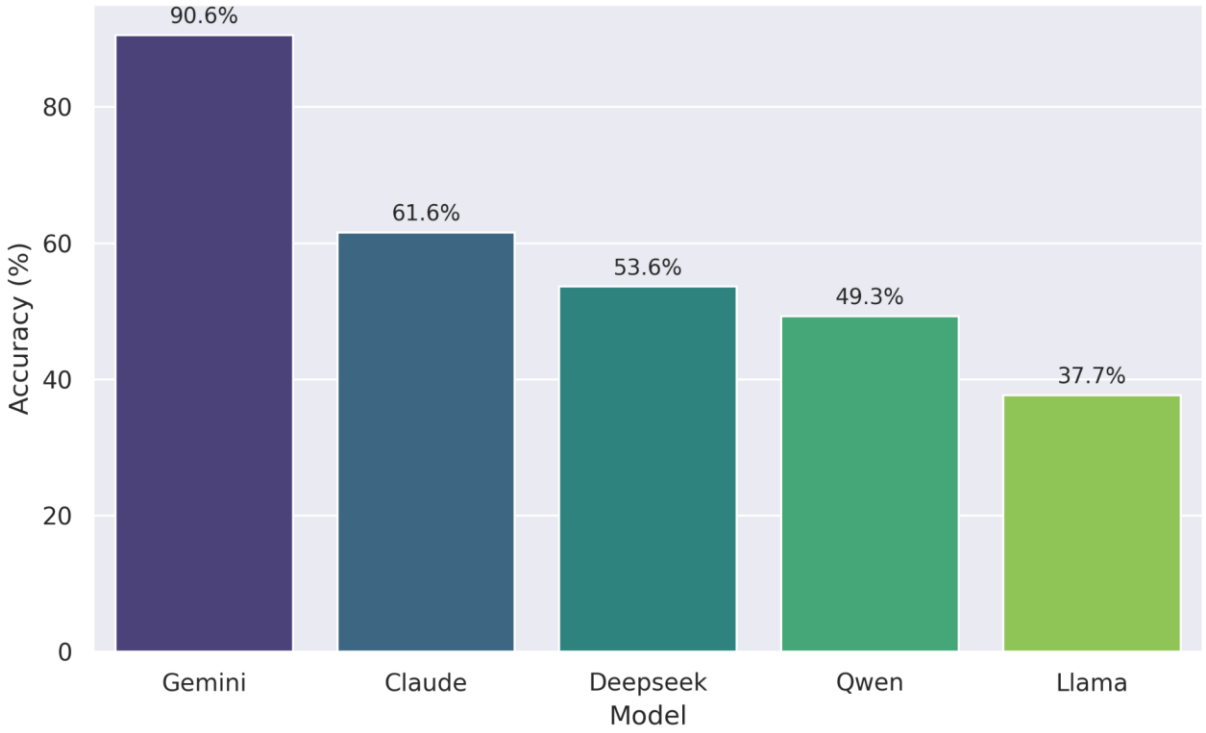


Figure 2 - Performance of LLM's on Kazakh language test in UNT

Figure 3 illustrates model accuracy across different difficulty levels. Gemini consistently achieved the highest accuracy, maintaining strong performance on easy (91.4%) and moderate (91.2%) items, with only a slight decline on hard items (87.0%). Claude displayed a different trend, performing moderately on easy questions (58.0%) but improving with increasing difficulty, reaching 69.6% on the hardest items. Deepseek followed a similar pattern, starting lower on easy items (51.9%) but showing its best performance on difficult questions (73.9%). Qwen demonstrated relatively stable results, hovering around 43–50% on easy and moderate items before a marked increase to 69.6% on hard items. Llama, in contrast, showed weaker performance overall, with

accuracy declining from 42.0% on easy items to 29.4% on moderate items, before partially recovering to 34.8% on hard items. These results suggest that while Gemini excels across all levels, other models—particularly Claude, Deepseek, and Qwen—may be more resilient on challenging items than on simpler ones, highlighting distinct strengths in handling complex linguistic reasoning.

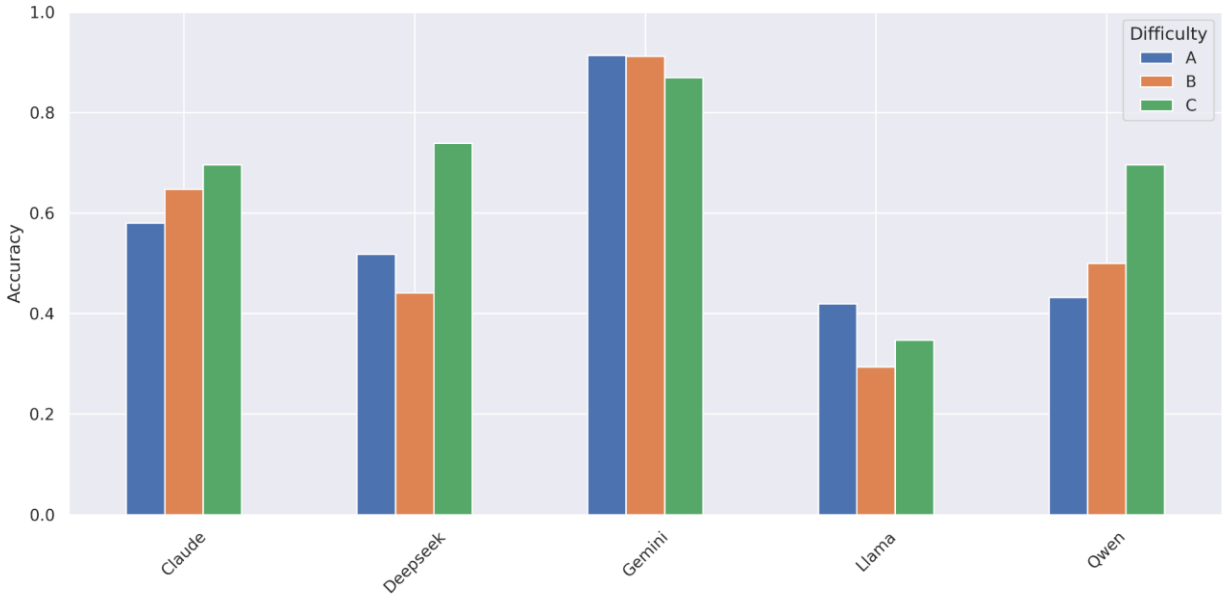


Figure 3 - Accuracy of LLMs on the Kazakh language test by difficulty level

The statistical analysis revealed substantial variation in LLM performance across the 138 UNT Kazakh language questions. Cochran's Q test confirmed that the models' accuracies were not equal ($Q = 108.30$, $df = 4$, $p < .001$), highlighting meaningful differences in their ability to handle low-resource language items. Post-hoc McNemar tests with Bonferroni correction showed that Gemini consistently outperformed Claude, Deepseek, Llama, and Qwen in pairwise comparisons, while some contrasts such as Claude vs Deepseek and Llama vs Qwen were not significant. The GEE logistic regression reinforced these findings, with Gemini achieving significantly higher accuracy than the baseline ($\beta = 1.79$, $p < .001$), corresponding to an estimated improvement of around 15–20 percentage points in correct responses. By contrast, Llama ($\beta = –0.98$, $p < .001$) and Qwen ($\beta = –0.50$, $p = .012$) performed significantly worse, suggesting they struggled with the nuances of Kazakh grammar and vocabulary. Deepseek showed a marginal trend toward lower performance ($p = .099$), positioning it between Claude and the weaker-performing models.

Table 1 summarizes accuracy across major linguistic topics. Gemini consistently achieved perfect accuracy across all categories, underscoring its strong coverage of Kazakh grammar, vocabulary, and syntax. Claude also performed robustly, particularly on professional terms, numerals, pronouns, and sentence-level constructions, though its accuracy was somewhat lower on polysemous words and adjective-related items. Deepseek demonstrated uneven performance, excelling on numerals and phrases but failing on pronouns and struggling with core grammatical categories such as nouns and adjectives. Qwen showed mixed results, performing well on compound sentences and pronouns but less consistently on other areas. Llama displayed the lowest overall accuracy, with performance dropping sharply on complex and sentence-level tasks. These findings highlight that while Gemini shows broad mastery, other models reveal topic-specific strengths and weaknesses, which could guide targeted applications in exam preparation and language learning.

Table 1 - Accuracy of LLMs on Kazakh language test by topic area (%)

Topic (Simplified English)	Claude	Deepseek	Gemini	Llama	Qwen
Professional / taboo terms	100.0	100.0	100.0	100.0	100.0
Numerals and their functions	100.0	100.0	100.0	33.3	66.7
Pronouns and their functions	100.0	0.0	100.0	100.0	100.0
Polysemous words (multiple meanings)	85.7	57.1	100.0	57.1	42.9
Complex sentences (subordinate clauses)	83.3	66.7	100.0	16.7	66.7
Phrases (noun and verb phrases)	80.0	80.0	100.0	80.0	40.0
Compound sentences (coordination)	75.0	50.0	100.0	37.5	75.0
Interjections and mimetic words	75.0	50.0	100.0	25.0	25.0
Nouns and their functions	75.0	25.0	100.0	50.0	50.0
Adjectives and degrees of comparison	75.0	0.0	100.0	25.0	50.0

From a practical perspective, Gemini's relative advantage translates into a substantially higher likelihood of selecting the correct option on UNT-style items, which could make it a more reliable tool for educational support in Kazakhstan's context. Meanwhile, the weaker performance of Llama and Qwen emphasizes the limitations of even large-scale models when applied to low-resource languages without fine-tuning. These findings illustrate both the promise and the current gaps of LLMs for national language assessments: while state-of-the-art models like Gemini can approach usable levels of performance, significant disparities remain, reinforcing the need for targeted adaptation and training on Kazakh data to ensure fairness and effectiveness in real-world applications.

Conclusion

The evaluation of five state-of-the-art LLMs on the UNT Kazakh language MCQs reveals significant disparities in their performance, with Gemini demonstrating near-human accuracy (90.6%) and robustness across difficulty levels and linguistic topics, while models like Llama and Qwen struggled, particularly without targeted optimization for Kazakh's complex morphology. These findings underscore the promise of LLMs in supporting scalable, efficient educational assessments for low-resource languages but highlight critical limitations in fairness, validity, and reliability. To ensure equitable application in high-stakes contexts like the UNT, future work should focus on fine-tuning models with Kazakh-specific datasets, developing culturally tailored benchmarks, and addressing biases through advanced prompt engineering and ethical frameworks. Such efforts are essential to bridge performance gaps and enhance the role of AI in Kazakhstan's educational landscape.

Acknowledgments

This research was funded by the National Testing Centre under the Ministry of Science and Higher Education of the Republic of Kazakhstan.

References

1. Nurseitova, K., Kaliyeva, A., Denst-Garcia, E., & Kussainova, R. (2017) Impact Of Language Of Instruction On Progress In Kazakhstan. *European Proceedings of Social and Behavioural Sciences*.

2. Teckwani, S. H., Wong, A. H. P., Luke, N. V., & Low, I. C. C. (2024). Accuracy and reliability of large language models in assessing learning outcomes achievement across cognitive domains. *Advances in Physiology Education*, *48*(4), 904-914.

3. Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., ... & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90-112.

4. Abd-Alrazaq, A., AlSaad, R., Alhuwail, D., Ahmed, A., Healy, P. M., Latifi, S., ... & Sheikh, J. (2023). Large language models in medical education: opportunities, challenges, and future directions. *JMIR medical education*, *9*(1), e48291.

5. Alekseev, A., & Turatali, T. (2024, October). KyrgyzNLP: challenges, progress, and future. In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 3-39). Cham: Springer Nature Switzerland.

6. Bommarito II, M., & Katz, D. M. (2022). GPT takes the bar exam. *arXiv preprint arXiv:2212.14402*.

7. Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, *382*(2270), 20230254.

8. Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia journal of mathematics, science and technology education*, *19*(8), em2307.

9. Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., Gao, Y., Andersen, O., ... & Buttery, P. (2023). On the application of large language models for language teaching and assessment technology. *arXiv preprint arXiv:2307.08393*.

10. Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education sciences*, *13*(7), 692.

11. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, *103*, 102274.

12. Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring in educational measurement. *Advancing natural language processing in educational assessment*, *142*.

13. Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health*, *2*(2), e0000198.

14. Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

15. Law, A. K., So, J., Lui, C. T., Choi, Y. F., Cheung, K. H., Kei-ching Hung, K., & Graham, C. A. (2025). AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Medical Education*, *25*(1), 208.

16. Pezeshkpour, P., & Hruschka, E. (2024, June). Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2006-2017).

17. Rouzegar, H., & Makrehchi, M. (2024). Generative AI for enhancing active learning in education: A comparative study of GPT-3.5 and GPT-4 in crafting customized test questions. *arXiv preprint arXiv:2406.13903*.

18. Sari, A. N. (2024). Exploring the Potential of Using AI Language Models in Democratising Global Language Test Preparation. *International Journal of TESOL & Education*, *4*(4), 111-126.

19. Stribling, D., Xia, Y., Amer, M. K., Graim, K. S., Mulligan, C. J., & Renne, R. (2024). The model student: GPT-4 performance on graduate biomedical science exams. *Scientific Reports*, *14*(1), 5670.

20. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, *15*(3), 1-45.

21. Maeda, H. (2025). Field-testing multiple-choice questions with AI examinees: English grammar items. *Educational and Psychological Measurement*, *85*(2), 221-244.

22. Maxutov, A., Myrzakhmet, A., & Braslavski, P. (2024, August). Do LLMs speak Kazakh? A pilot evaluation of seven models. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)* (pp. 81-91).

23. Togmanov, M., Mukhituly, N., Turmakhan, D., Mansurov, J., Goloburda, M., Sakip, A., ... & Koto, F. (2025). KazMMLU: Evaluating Language Models on Kazakh, Russian, and Regional Knowledge of Kazakhstan. *arXiv preprint arXiv:2502.12829*.

24. Karibayeva, A., Karyukin, V., Abduali, B., & Amirova, D. (2025). Speech Recognition and Synthesis Models and Platforms for the Kazakh.

25. Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice*, *20*(2), 1-24.

26. Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., ... & Morilova, P. (2024). The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges. *Chinese/English Journal of Educational Measurement and Evaluation*, *5*(3), 3.

27. Walsh, A. (2024). Cusco Quechua and the world of AI: a case study on low resource languages and large language models.

28. Li, S. (2025, April). Trustworthy AI Meets Educational Assessment: Challenges and Opportunities. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 27, pp. 28637-28642).

29. National Test Center. (2024). Sample UNT Kazakh language test questions. Available online at: https://testcenter.kz/upload/iblock/ba5/_-_.pdf (Accessed September 17, 2025).

Baizhanov N., Sabyrov A., Makhmutova A., Kadyrov S.

ІРІ ТІЛ МОДЕЛЬДЕРІНІҢ ҚАЗАҚ ТІЛІНІҢ БІЛІМ ДЕҢГЕЙІН БАҒАЛАУДАҒЫ РӨЛІ

Бұл зерттеу Қазақстандағы Ұлттық бірыңғай тестілеу (ҰБТ) контекстінде қазақ тілін меңгеруді бағалаудағы ірі тіл модельдерінің (LLM) өнімділігін бағалайды. Негізгі мақсат — бес заманауи LLM-нің: Gemini 2.5 Pro Preview, Claude 3.7 Sonnet, Deepseek R1, Qwen және Llama 3.1-405B-Instruct — 2024 жылғы БҰТ қазақ тілі тестінен алынған 138 таңдаулы сұрағындағы дәлдігін, қателіктер сипатын және психометриялық сипаттамаларын зерттеу. Әдістемеге сыртқы деректерге қол жеткізусіз, стандартталған нұсқаулар негізінде нөлдік оқыту режиміндегі бағалауды, сондай-ақ Кохран Q тесті, МакНемар тесттері және модельдердің күрделілік деңгейлері мен лингвистикалық тақырыптар бойынша өнімділігін талдау үшін жалпыланған бағалау теңдеулерімен (GEE) логистикалық регрессияны қоса алғандағы статистикалық талдау кірді. Нәтижелер Gemini-дің ең жоғары дәлдікке (90,6%) қол жеткізгенін және басқа модельдерден айтарлықтай жоғары көрсетеді, ал Llama ең төменгі нәтиже көрсетті (37,7%). Өнімділік күрделілік пен тақырыпқа байланысты өзгеріп отырды, бұл ретте Gemini барлық санаттарда басымдылық танытты, ал басқа модельдер күрделі лингвистикалық ойлау сияқты нақты белгілі бір салаларда өздерінің мықты жақтарын көрсетті. Зерттеу қазақ тілі сияқты ресурстары шектеулі тілдерде білімді бағалау үшін LLM -нің әлеуетін атап көрсетеді, алайда модельдерді оңтайландырудағы, әділдік пен сенімділіктегі олқылықтарды анықтайды, бұл мәдени бейімделген деректерді мақсатты түрде нақтылауды және пайдалануды талап етеді.

Түйін сөздер: Ірі тіл модельдері, Білім беру бағалауы, Қазақ тілі, Бірыңғай ұлттық тестілеу, Педагогикалық өлшеу, Жасанды интеллект.

Baizhanov N., Sabyrov A., Makhmutova A., Kadyrov S.

РОЛЬ КРУПНЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ В ОЦЕНКЕ ВЛАДЕНИЯ КАЗАХСКИМ ЯЗЫКОМ

Данное исследование оценивает производительность крупных языковых моделей (LLM) в оценке владения казахским языком в контексте Единого национального тестирования (ЕНТ) в Казахстане. Основная цель – изучить точность, характер ошибок и психометрические характеристики пяти современных LLM: Gemini 2.5 Pro Preview, Claude 3.7 Sonnet, Deepseek R1, Qwen и Llama 3.1-405B-Instruct – на 138 вопросах с множественным выбором (MCQ) из теста по казахскому языку ЕНТ 2024 года. Методология включала оценку в режиме нулевого обучения с использованием стандартизированных подсказок без доступа к внешним данным, а также статистический анализ, включая тест Кохрана Q, тесты МакНемара и логистическую регрессию с обобщенными оценочными уравнениями (GEE) для анализа производительности моделей по уровням сложности и лингвистическим темам. Результаты показывают, что Gemini достигла наивысшей точности (90,6%), значительно превосходя другие модели, в то время как Llama показала наименьший результат (37,7%). Производительность варьировалась в зависимости от сложности и тематики, при этом Gemini демонстрировала превосходство во всех категориях, а другие модели показывали сильные стороны в специфических областях, таких как сложное лингвистическое мышление. Исследование подчеркивает потенциал LLM для образовательной оценки в языках с ограниченными ресурсами, таких как казахский, но выявляет пробелы в оптимизации моделей, справедливости и надежности, что требует целенаправленной доработки и использования культурно адаптированных данных.

Ключевые слова: Крупные языковые модели, Образовательная оценка, Казахский язык, Единое национальное тестирование, Педагогическое измерение, Искусственный интеллект.

Сведения об авторах:

Байжанов Нурсейт Абсаттарович - РГП на ПХВ «Национальный центр тестирования» МНВО РК, первый заместитель директора, ул. Родниковая, 1/1, 010008, Астана, Казахстан, e-mail: nurbaizhanov@gmail.com.

Сабыров Аслан Асхатович - РГП на ПХВ «Национальный центр тестирования» МНВО РК, старший эксперт-аналитик отдела аналитики и искусственного интеллекта, ул. Родниковая, 1/1, 010008, Астана, Казахстан, e-mail: aslansabyrov@gmail.com.

Махмутова Альфира Абдукалыковна - Университет SDU, старший преподаватель, факультет образования и гуманитарных наук, ул. Абылайхана, 1/1, Каскелен, 040900, Казахстан. e-mail: alfira.makhmutova@sdu.edu.kz.

Кадыров Ширали Маратжанович - Новый Узбекский университет, доцент, кафедра общих дисциплин, Мовараунахр, 1, 100000, Ташкент, Узбекистан. e-mail: sh.kadyrov@newuu.uz.

Авторлар туралы мәлімет:

Байжанов Нұрсейт Абсаттарұлы - ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, директордың бірінші орынбасары, Родниковая көш., 1/1, 010008, Астана, Қазақстан, e-mail: nurbaizhanov@gmail.com.

Сабыров Аслан Асхатұлы - ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, Талдау және жасанды интеллект бөлімінің аға сарапшы-талдаушысы, Родниковая к-сі, 1/1, Астана, 010008, Қазақстан, e-mail: aslansabyrov@gmail.com.

Махмутова Альфира Абдукалыковна - SDU университеті, аға оқытушы, Білім беру және гуманитарлық ғылымдар факультеті, Абылайхан көшесі, 1/1, Қаскелең, 040900, Қазақстан. e-mail: alfira.makhmutova@sdu.edu.kz.

Кадыров Ширали Маратжанович - Жаңа Өзбекстан университеті, доцент, Жалпы білім беру кафедрасы, Мовараунахр, 1, 100000, Ташкент, Өзбекстан. e-mail: sh.kadyrov@newuu.uz.

Information about authors:

Baizhanov Nurseit Absattarovich - Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, First Deputy Director, 1/1 Rodnikovaya St., 010008, Astana, Kazakhstan, e-mail: nurbaizhanov@gmail.com.

Sabyrov Aslan - Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Senior Data Analyst, 1/1 Rodnikovaya St., Astana, 010008, Kazakhstan, e-mail: aslansabyrov@gmail.com.

Makhmutova Alfira - SDU University, Senior Lecturer, Faculty of Education and Humanities, 1/1 Abylaykhan st, Kaskelen, 040900, Kazakhstan, e-mail: alfira.makhmutova@sdu.edu.kz.

Kadyrov Shirali - New Uzbekistan University, Associate professor, Department of General Education, Movarounnahr, 1, 100000, Tashkent, Uzbekistan. e-mail: sh.kadyrov@newuu.uz.

Сведения о документе

Тип документа	Внутренний документ
Номер и дата документа	15-10-5-2-04/4072 от 11.11.2025
Ссылка на документ	https://uto.workspace.kz/storage/document_attachments/m6QFroAQkWUEAWvOE54aHnkW78ebWbqmsgAVGCpQ.pdf
Отправитель	РГП на ПХВ "НАЦИОНАЛЬНЫЙ ЦЕНТР ТЕСТИРОВАНИЯ"
Автор	Туралбаева К. Т., Заведующий (тел: +7 701 843 31 69, email: kengebahit@mail.ru

Лист согласования

ФИО	Дата и время	Результат	ЭЦП
Туралбаева Кенжекул Турганбаевна	2025-11-11 13:18:57	Согласован	Нет

Лист подписания

Туралбаева Кенжекул Турганбаевна	2025-11-11 13:19:17	Подписан	Нет

Лист регистрации

Лист отправки

Лист корреспондентов