

Evaluating Large Language Models for Sentence Augmentation in Low-Resource Languages: A Case Study on Kazakh

Zhamilya Bimagambetova
KBTU
Almaty, Kazakhstan
zhami_2000@icloud.com

Dauren Rakhymzhanov
KBTU
Almaty, Kazakhstan
d_rakhymzhanov@kbtu.kz

Assel Jaxylykova
IICT
Almaty, Kazakhstan
aselya17.89@mail.ru

Alexander Pak
IICT, KBTU
Almaty, Kazakhstan
aa.pak83@gmail.com

Abstract—Large language models (LLMs) have revolutionized natural language processing (NLP) and demonstrated exceptional performance in various NLP tasks for widely spoken languages. However, their efficacy in handling low-resource languages remains an area of concern. This study investigates the performance of LLMs, particularly GPT-3, in sentence augmentation tasks for a low-resource language, Kazakh. We employ a blind peer review methodology, where five native Kazakh annotators assess the quality of LLM-generated augmentations. The results reveal that LLMs excel in popular languages like English, Chinese, and German, but face challenges with low-resource languages due to limited training data. This work sheds light on the importance of improving LLMs' adaptability and relevance to address the unique needs of low-resource languages. Further research could enhance the augmentation capabilities of LLMs in scenarios with limited data sources, ensuring their effectiveness in promoting linguistic diversity and inclusivity. Furthermore, this study underscores the significance of cross-language transfer learning and data collection efforts to empower LLMs in supporting linguistic diversity and fostering inclusivity across the global language landscape.

Index Terms—Large language models, sentence augmentation, low-resource languages, Kazakh language.

I. INTRODUCTION

In contemporary natural language processing (NLP) research, a significant focus has been placed on a small subset of the 7,168 languages spoken globally, with English playing a dominant role. This preference is primarily due to English's status as an international language and its widespread use. Consequently, many NLP models and datasets are tailored to the English language, while the vast majority of other languages, often referred to as low-resource languages (LRLs), remain relatively neglected and understudied. LRLs face several challenges, including limited linguistic resources, scarce data, and a lack of research interest [1]–[3].

The Kazakh language is an example of a LRL. Despite being the official language of Kazakhstan, Kazakh suffers from underrepresentation in the digital space, mainly due to a sparse population and a small number of native speakers. Consequently, there is a scarcity of linguistic resources, such as

corpora and datasets, hindering NLP research and applications in Kazakh.

Kazakh belongs to the Turkic group of the Altai family of languages and is an agglutinative language, that is, a language in which words are formed by adding affixes to the base. In Kazakh the accent usually falls on the last syllable of a word and can shift when adding affixes. Compared to widely spoken and studied languages such as English or Russian, very few corpora and datasets are available for Kazakh. For instance, the National Corpus of the Russian language has over 2 billion words. While the Corpus of the Kazakh language only has around 31 million words. [4] [5] Moreover, the National Corpus is not openly accessible, so researchers have to request the access or more probable create their own linguistic tools. At the proposed paper there are artificially created kazakh corpus in banking field. We made a translation of this corpus to Kazakh language which later we used to make data augmentation using ChatGPT. Thus, the existing problem lies in the issue of strengthening the linguistic tools of low-resource languages using the augmentation approach of large language models. In this study, this problem is considered on the example of the corpus of the Kazakh language of the banking sector and the well-known language model of ChatGPT.

To address the scarcity of data in LRLs like Kazakh, data augmentation techniques have been proposed. Data augmentation involves generating new texts by transforming existing data, thereby expanding the available training data for NLP models. This approach has been widely used to improve model performance in various NLP tasks. Augmentation methods can be applied at different levels of granularity, such as characters, words, sentences, and documents, to enrich the available data for low-resource languages.

Pre-trained language models (PLMs) based on transformer architectures, such as BERT [6] and GPT [7], have significantly advanced NLP research and demonstrated state-of-the-art performance on various tasks. While these models have predominantly focused on high-resource languages, the recent development of larger language models like GPT-3 [7], PaLM [8], Bloom [9], OPT [10], and the FLAN series [11], with

Ministry of Education and Sciences, Republic of Kazakhstan (Grant num. AP09260670)

hundreds of billions of parameters, has shown potential for addressing challenges in low-resource language processing.

By leveraging data augmentation techniques with large language models like GPT, it is possible to enhance linguistic resources for low-resource languages such as Kazakh. This approach holds promise for improving NLP applications in these languages and making creative works more accessible to their speakers. Emphasizing research and development in low-resource languages is critical for promoting linguistic diversity and inclusivity in the evolving landscape of NLP and artificial intelligence.

II. METHODOLOGY

The main objective of this study is to investigate the performance of large language models (LLMs), specifically GPT-3, in sentence augmentation tasks for low-resource languages, focusing on the case of Kazakh. While LLMs have shown remarkable success in various NLP tasks for widely spoken languages like English, German, Spanish, and Japanese, their effectiveness in handling tasks related to low-resource languages has not been adequately explored. We aim to highlight this limitation and draw attention to the need for improvement in LLMs' capabilities for low-resource languages.

To validate our hypothesis, we adopted a methodology similar to that used in SemEval-2015 Task 11. The blind peer review method was employed, wherein annotators were presented with both the original text and its augmented version generated by the LLM models. The annotators were then asked to assess the quality of the augmented sentences. This assessment provides valuable insights into the ability of LLMs to produce high-quality augmentations in the context of low-resource languages.

It is essential to mention that SemEval-2015 Task 11 focused on sentiment analysis using LLMs in figurative language present on Twitter. In contrast, our study aims to explore the augmentation skills of LLMs, particularly with regard to less commonly spoken languages. By comparing the results of our investigation with the Task 11 findings, we can shed light on the challenges faced by LLMs when applied to low-resource languages and offer suggestions for potential improvements in this domain.

To assess the quality of the sentence augmentations generated by LLMs for the Kazakh language, we adopted a rating scale ranging from 1 to 5:

- Very bad - There is no sense in the augmented sentence at all, it is not built according to the rules of the Kazakh language, there are obvious mistakes that the speaker would not make.
- Bad - The sentence can be understood if you know the original, but there are mistakes in the Kazakh language and / or the meaning is distorted
- Acceptable - The meaning of the sentence must be clear without knowing the original, errors in the language are allowed
- Fine - The meaning of the sentence is well understood, it is possible to use words that are not often used by the

native speaker and / or small errors in the language, a slight loss of meaning is allowed

- Great - The sentence is well composed, there are no mistakes in the Kazakh language (except for those that the speakers themselves can often make), the forward meaning is good, it does not need to be thought out

The main goal of the blind peer review method is an anonymous estimate of the accuracy of the sentence augmentation, where each annotator does not know how the other participants in the survey are assessed.

In this study, we utilized the well-known English corpus "bankin77" as our primary data source and translated it into Kazakh using a specific model/translator, which we will mention later in the paper. From this corpus, we focused on the largest class labeled "banking cards," containing 150 sentences. Through text augmentation, we applied parameter tuning in the GPT model, based on the work of a particular reference. The selection of the GPT model was based on the absence of suitable LLM models designed specifically for text augmentation in the Kazakh language. Prominent models like AugGPT and BERT do not support Kazakh, making GPT the most viable option for our investigation.

The results obtained from our experiments validated our hypothesis, demonstrating that LLMs exhibit remarkable performance in NLP tasks for widely spoken languages like English, Chinese, and German [12]. However, this efficiency does not extend to low-resource languages due to the limited availability of training data. The inadequacy of training data poses a significant challenge for LLMs when applied to languages with scant digital resources. To enhance the efficacy of LLMs in such scenarios, future research may focus on improving their adaptability and relevance with limited data sources. Moreover, additional detailed analyses could be explored to address the challenges of working with a dearth of digital data for low-resource languages.

In our current research, we address the issue of scammers potentially manipulating task results by assigning identical or random scores to all units within a given task. We utilize the standard deviation $std_u(u_i)$ of all judgments provided by annotator u_i . A low standard deviation, specifically $std_u(u_i) = 0$, can serve as an indicator that annotator u_i might be engaging in scamming behavior.

Similarly, we leverage the standard deviation $std_t(t_j)$ of all judgments provided for a statement t_j . This information allows us to identify annotation $A_{i,j}$, as given by annotator u_i for statement t_j , as an outlier if certain conditions are met.

By incorporating these measures into our annotation process, we aim to ensure the reliability and accuracy of the gathered data, mitigating the potential impact of scammers and maintaining the quality of our research outcomes:

$$\left| A_{i,j} - \text{avg}(A_{i',j}) \right| > \text{std}_t(t_j) \quad (1)$$

If 50% or more of an annotator's judgements are judged to be outliers in this way then the annotator is deemed

a scammer and dismissed from the task. Each statement was cleaned of all annotations provided by those deemed to be scammers. The code and data are provided at <https://github.com/AlexPak/OPCS2023Augmentation>.

III. RESULTS

The results of the assessment conducted by five annotators show variations in their evaluations of the sentence augmentations.

Annotator 1: This annotator assigned the highest share of outliers, accounting for 53% of the sentences evaluated. The mean assessment score was 4.18, indicating a generally positive view of the augmentations. The median assessment score was the highest among all annotators at 5.00, implying a tendency to provide higher ratings. Notably, this annotator had the highest number of sentences rated as "Great," comprising 106 instances, but also had a considerable number of sentences rated as "Very Bad" (14 sentences). This suggests a more optimistic outlook overall but with some instances of substantial mismatches. Due to the high value of outliers the annotator was considered as scammer and had been excluded from further analysis.

Annotator 2: This annotator had the lowest share of outliers at 19%, indicating a more consistent and critical evaluation of the augmentations. The mean assessment score was 2.94, which was the highest among all annotators. However, the median assessment score of 3.00 was not significantly higher than other annotators. Annotator 2 assigned the highest number of sentences to the "Very Bad" category (50 sentences) and the second-highest number to the "Bad / Lost Part of the Meaning" category (19 sentences). This suggests a more critical approach and a tendency to identify cases of poor alignment with the original text.

Annotator 3: This annotator had a 20% share of outliers and the lowest mean assessment score of 2.53. The median assessment score was the lowest among all annotators at 2.00, indicating a relatively pessimistic view of the augmentations. Annotator 3 assigned the highest number of sentences to both the "Very Bad" (55 sentences) and "Bad / Lost Part of the Meaning" (33 sentences) categories. This suggests a more stringent evaluation with a focus on mismatches between the augmented sentences and the original text.

Annotator 4: This annotator had a 21% share of outliers and a mean assessment score of 2.13, the second-lowest among all annotators. The median assessment score was 2.00, indicating a consistent view of the augmentations. Annotator 4 assigned the highest number of sentences to the "Very Bad" category (70 sentences) and the "Bad / Lost Part of the Meaning" category (35 sentences). This suggests a critical evaluation and a tendency to identify instances of significant divergence from the original meaning.

Annotator 5: This annotator had a 31% share of outliers and a mean assessment score of 2.95, the second-highest among all annotators. The median assessment score was 3.00, aligning with the majority of annotators. Annotator 5 assigned the highest number of sentences to the "Great" category (44

sentences) and a relatively low number to the "Very Bad" category (45 sentences). This suggests a more positive outlook overall, with a focus on successful augmentations.

Overall, the results demonstrate a range of perspectives among the annotators, with differing levels of optimism and criticality in their assessments of the sentence augmentations. Such variations in evaluations highlight the subjectivity inherent in the assessment process and underscore the importance of conducting a thorough analysis of the obtained results.

The "Total" column in the results table represents the aggregated statistics or summary of the assessments provided by the five annotators. It provides an overall perspective of the evaluations for each category of sentence augmentations.

In the "Total" row, the "Share of outliers" value is absent since it represents the percentage of sentences that were considered outliers by each individual annotator and not a cumulative value. The "Mean assessment" and "Median assessment" values in the "Total" column represent the average and median scores, respectively, across all five annotators for each sentence augmentation category.

The subsequent rows in the "Total" column list the number and percentage of sentences falling into each category:

- "Very Bad": This category comprises sentences for which the augmentation resulted in a significant mismatch with the original meaning. In the "Total" row the percentage (37%) are provided, reflecting the total count and percentage of sentences classified as "Very Bad" by all five annotators combined.

- "Bad / Lost Part of the Meaning": This category includes sentences where the augmented version partially lost the meaning of the original text. The "Total" row shows percentage (19%) of sentences falling into this category based on the cumulative assessments of all five annotators.

- "Acceptable": Sentences rated as "Acceptable" were considered to have an adequate match with the original meaning, although some minor issues may be present. The "Total" row lists percentage (11%) of sentences in this category according to the combined assessments of all five annotators.

- "Fine": This category represents sentences where the augmentation was satisfactory, with no major issues identified. In the "Total" row - percentage (13%) of sentences falling into this category are presented based on the evaluations of all five annotators.

- "Great": Sentences rated as "Great" indicate an excellent job of augmentation, fully preserving the meaning of the original text. In the "Total" row - percentage (21%) of sentences falling into this category are listed, based on the collective assessments of all five annotators.

The "Total" column provides an aggregate view of the evaluations across all annotators, offering a comprehensive understanding of the quality and distribution of the sentence augmentations. It serves as a valuable summary of the overall assessment results and aids in drawing conclusions about the performance of the large language model in handling the Kazakh language in the context of sentence augmentation.

TABLE I
RESULTS OF THE EXPERIMENT - MANUAL ASSESSMENT

-	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Annotator 5	Total
Share of outliers	53%	19%	20%	21%	31%	-
Mean assessment	4,18	2,94	2,53	2,13	2,95	-
Median assessment	5,00	3,00	2,00	2,00	3,00	-
Very bad	14	50	55	70	45	37%
Bad	10	19	33	35	24	19%
Acceptable	17	15	17	15	18	11%
Fine	4	20	20	18	18	13%
Great	106	45	26	13	44	21%

IV. DISCUSSION

The variations in the annotators' assessments highlight the subjectivity inherent in the evaluation process for sentence augmentations. These differences could be attributed to factors such as individual language proficiency, linguistic nuances, and personal interpretations of sentence meaning. To address this issue, future studies could explore methods to reduce annotation subjectivity and enhance the reliability of assessments.

Additionally, the small size of the available Kazakh corpus presents a limitation in this study. To mitigate this, efforts should be made to expand and curate larger corpora specifically for low-resource languages, allowing for more comprehensive evaluations of LLMs' performance. Collaborative efforts with linguists and native speakers could be instrumental in creating such valuable linguistic resources.

Furthermore, considering the scarcity of LLMs tailored to low-resource languages, researchers should explore the possibility of developing language-specific pre-training models that cater to the unique linguistic characteristics and data limitations of these languages. Fine-tuning LLMs on domain-specific tasks using augmented data may also improve their performance for low-resource languages.

Overall, this study sheds light on the challenges and opportunities in leveraging LLMs for low-resource language processing. By addressing these challenges and investing in research to enhance NLP capabilities for these languages, we can foster inclusivity, preserve linguistic diversity, and unlock the full potential of NLP technology for all communities worldwide.

V. CONCLUSION

In this study, we explored the performance of large language models (LLMs), particularly GPT-3, in sentence augmentation tasks for the low-resource language of Kazakh. Our findings indicate that while LLMs have demonstrated exceptional capabilities in popular languages such as English, Chinese, and German, they face significant challenges when applied to low-resource languages due to limited training data availability. The blind peer review methodology, involving five native Kazakh annotators, provided valuable insights into the quality of the LLM-generated augmentations.

The results emphasize the need for further research and development to improve the effectiveness of LLMs in low-

resource language processing. Enhancing their adaptability and relevance in scenarios with limited data sources will be crucial for making significant progress in NLP applications for such languages. By addressing these challenges, we can unlock the potential of LLMs in promoting linguistic diversity and ensuring that no language is left behind in the advancements of NLP and artificial intelligence.

ACKNOWLEDGMENT

We gratefully acknowledge the financial support of the Ministry of Education and Sciences, Republic of Kazakhstan (Grant num. AP09260670 "Development of methods and algorithms for augmentation of input data for modifying vector embeddings of words")

REFERENCES

- [1] C. Cieri, M. Maxwell, S. Strassel, and J. Tracey, "Selection criteria for low resource language programs," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 4543–4549.
- [2] D. Jurgens, Y. Tsvetkov, and D. Jurafsky, "Incorporating dialectal variability for socially equitable language identification," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 51–57.
- [3] L. Campbell, "Ethnologue: Languages of the world," 2008.
- [4] O. Makhambetov, A. Makazhanov, Z. Yessenbayev, B. Matkarimov, I. Sabyrgaliyev, and A. Sharafudinov, "Assembling the kazakh language corpus," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1022–1031.
- [5] V. Plungyan, "Why do we need the national corpus of the russian language?" *Informal introduction/Website of the National Corpus of the Russian Language*. URL: www.ruscorpora.ru, 2005.
- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [8] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.
- [9] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.
- [10] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.
- [11] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

- [12] H. Dai, Z. Liu, W. Liao, X. Huang, Z. Wu, L. Zhao, W. Liu, N. Liu, S. Li, D. Zhu *et al.*, “Chataug: Leveraging chatgpt for text data augmentation,” *arXiv preprint arXiv:2302.13007*, 2023.