Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2025"

April 23–25, 2025

Backtranslation Invariance Boosts Effectiveness of Non-English Prompts

Kurtukova A.
NTR Labs, Tomsk, Russia;
Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia
akurtukova@ntr.ai

Kozachenko A.
NTR Labs,
Tomsk, Russia
akozachenko@ntr.ai

Abstracts

We present an approach to improving non-English prompts based on backtranslation invariance (the semantics of the prompt should not change after automatic translation to English and back). It improves prompts in non-English languages for a variety of Large Language Models (LLMs), including GPT-4-o, Llama-3.1, and Mixtral8x7B. We evaluate the approach for Russian and Finnish languages. In the benchmark of removing commas from a sentence, the proposed approach achieved an accuracy increase of 42% for Russian and 54% for Finnish compared to non-invariant prompts (LLaMA). In the benchmark of counting commas, accuracy increase of 19% for Russian and 11% for Finnish (GPT).

Keywords: Prompt engineering, large language models, translation, Backtranslation Invariance
DOI: 10.28995/2075-7182-2025-23-XX-XX

1 Introduction

As the large language models (LLMs) productized and became a go-to tool in various fields including natural language processing (NLP), translation and text generation, it became apparent that the performance of these models varies across languages [1-3]. Research shows that multilingual LLMs are more effective for the English language, which is due to prevalence of English in the data used for training [4-6].

Using LLM for non-English languages may lose effectiveness due to the peculiarities of specific languages. Inflectional languages (such as Russian) and agglutinative languages (such as Finnish) have specific features that create certain difficulties when working with LLM [7]. In such languages, the morphological structure can vary significantly depending on the context [8, 9], which requires the model to have a deep understanding of syntactic and semantic relationships [10]. Unlike English, in inflectional languages the diversity of forms can lead to increased ambiguity and complicate information processing [11]. In addition, many linguistic features such as word order, case use, and agreement can be ignored or misinterpreted by models that do not have sufficient experience working with specific language groups [12]. This calls into question the universality of approaches to prompting and processing texts in languages other than English.

The hypothesis of this study was that a specific formulation of a prompt in a non-English language provides LLM results comparable in quality to those obtained using English-language prompts. To achieve this goal, we hypothesized that it is important that the original prompt has invariance to back-translation through English.

2 Prior research

The paper [2] is devoted to assessing the impact of non-English prompts on the effectiveness of a recommender system based on LLM. In the study, the authors considered both the out-of-the-box model and the T5 model retrained on multilingual prompts. The experiments were conducted for English, Turkish, and Spanish. Various prompting techniques were considered. The effectiveness of LLM was

assessed using the HitRate and NDCG metrics on open datasets: ML1M, LastFM, and Amazon-Beauty. The results obtained by the authors showed that the use of non-English prompts has a negative impact on the effectiveness of LLM as part of a recommender system. However, retraining LLM on multilingual prompts ensured uniform effectiveness of recommendations regardless of language. The efficiency measured by HitRate@10 for English decreased from 0.0679 to 0.0523, for Spanish from 0.0551 to 0.0505, and for Turkish increased from 0.0505 to 0.0523. The efficiency measured by NDCG@10 for English decreased from 0.0370 to 0.0288, for Spanish increased from 0.0297 to 0.0302, and for Turkish from 0.0269 to 0.0288.

The paper [3] presents an approach to qualitative and quantitative evaluation of LLM capabilities to work with multilingual data. The authors' approach is based on forward and backward translations and is tested for solving common sense reasoning and pun detection problems in order to determine the type of bilingualism demonstrated by LLM. To evaluate the approach, the authors used the GSM8K (primary school math problems) and CommonsenseQA (multiple-choice logical questions) and WebQuestions (question-answer pairs) datasets translated into French, Spanish, German, Japanese and Chinese. GPT-4 was used as a model. The obtained results demonstrated a significant difference in the efficiency of LLM for English and non-English languages. When solving math problems, prompting in English was on average 10% better than in other languages, and when solving logical problems - 15%. On WebQuestions, the results between European and English languages were approximately equal, while the effectiveness of LLM in Japanese and Chinese was lower by 16% and 28%, respectively.

The authors of the paper [4] considered the problem of transferring the capabilities of efficient generation and execution of LLM instructions to non-English languages. The study was based on the application of both out-of-box and pre-trained models of the LLaMA family (LLaMA, LLaMA2, LLaMA Chinese, etc.) to 4 standard benchmarks: C-Eval, MMLU, AGI-Eval, and GAOKAO-Bench. As part of the experiments, the authors evaluated the metrics of accuracy, fluency, informativeness, logical coherence, and harmlessness on the LLM-Eval dataset, which includes various educational tasks. The metrics were evaluated in 14 different languages: Chinese, Arabic, Vietnamese, etc. Based on the results obtained, the authors came to the following conclusions:

- expanding the vocabulary does not provide any advantages when training on data volumes of tens of billions of tokens;
- pretraining can improve the quality of responses, but does not always lead to a significant increase in the model's knowledge level;
- improving the LLM's ability to understand non-English languages is achieved at the expense of losing its initial understanding of English.

The authors of the paper [5] presented their MindMerge approach based on using the output data of one multilingual model trained for the translation task as input data for a multilingual LLM. This approach provided correct translation of not only prompts, but also user queries themselves. The efficiency of the approach was evaluated on the MGSM and MSVAMP datasets containing samples in Russian, Japanese, Spanish, and other languages. The difference in the efficiency of LLM with and without the approach reached 7.6% on average for all languages and 8% for low-resource languages.

Summarizing, several key conclusions can be drawn:

- The use of prompts in non-English languages negatively impacts the effectiveness of LLMs. Models predominantly trained on English data may distort information when processing other languages unless additional training is conducted.
- Retraining models on multilingual datasets improves their performance for non-English languages, highlighting the importance of tailoring LLM to specific languages to achieve uniform performance.
- Improvements in LLM's ability to understand non-English languages may come at a cost in English comprehension, indicating potential limitations of current approaches to training multilingual models.

3 An approach to prompt engineering based on Backtranslation invariance

We suggest an approach that is based on the hypothesis that the Backtranslation invariance of the prompt into English has a positive effect on the quality of LLM answers. We have first empirically observed this phenomenon when using compound sentences in prompts in Russian. Despite the fact that the prompt rules in Russian were formulated correctly and unambiguously, some of them were ignored by the LLM. To resolve this conflict, we decided to resort to machine translation of the prompt into English, and then from English to the original language. We observed that as a result of this manipulation some rules lost their original meaning as intended in the source language, and these were exactly the rules that were ignored by the LLM (likely, due to the impossibility of unambiguous interpretation).

The illustration of the approach to prompt engineering based on Backtranslation invariance is presented in Fig. 1. The approach includes:

- A cycle. Creating and modifying the original prompt or part of the prompt in language X to achieve invariance to backtranslation.
- Using machine translation from language X to English on the original prompt.
- Using machine translation on the resulting English prompt for backtranslation to language X.

If the prompt obtained after steps 2 and 3 matches the prompt from step 1, the cycle exits. Otherwise, the cycle repeats again, starting from step 1.

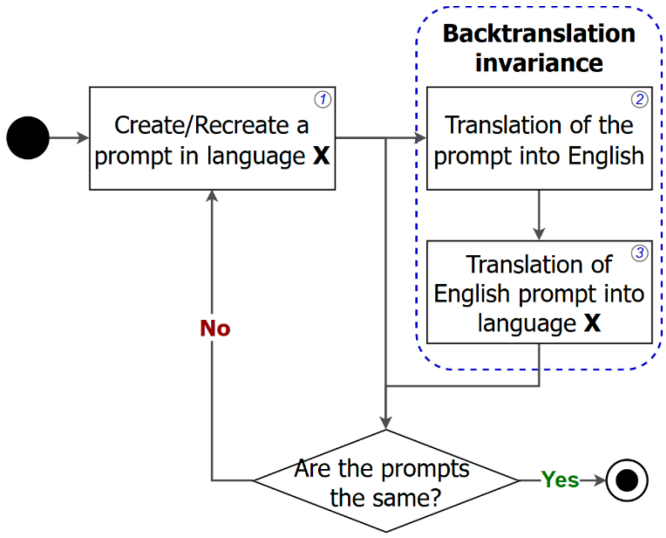

Figure 1: UML diagram of the approach

4 Dataset and sets of prompts

4.1 Test task

We created a small benchmark dataset to check the rule-following in different languages with ambiguous wording of the prompt. We created a small benchmark dataset to check the rule-following in different languages with ambiguous wording of the prompt. The first simple task was to remove commas from the text of a sentence if the word "I" was present. The second, more difficult task was to count the number of commas in the sentence.

The dataset was based on the prose of Russian literary classics (Tolstoy, Chekhov, etc.). The total number of works amounted to 297, with a total of 182,476 non-empty lines. Text segments (chunks) were filtered based on several criteria:

- A length of at least 10 words;
- Absence of Latin characters and punctuation marks indicating direct or indirect speech;
- Presence of at least one comma in the chunk – the number of such chunks was limited by 90% of the total dataset size;

- Absence of commas in the line – the number of such chunks was limited by 10% of the total dataset size;
- Presence of at least one pronoun “I” – the number of such chunks was limited by 70% of the total dataset size;
- Absence of the pronoun “I” – the number of such chunks was limited by 30% of the total dataset size.

The total number of chunks extracted was 15,436. To reduce computational load, 248 chunks were selected conforming to ratio requirements and taking into account that some chunks met multiple conditions simultaneously.

4.2 Prompt sets for English language

The English prompts were developed as a benchmark to show that the least ambiguous non-English prompts should achieve quality metrics comparable to the English prompts.

The first prompt for removing all commas from the text when the condition is met (En1):
“I will provide you with a sentence. Please rewrite it. Please answer in Russian only.
If the word ‘I’ is in the sentence, remove all commas from the text.
There should be nothing in the answer except for the rewritten sentence.
*Sentence: **{sample}**”*

The second prompt is for counting the number of commas in the text (hereinafter – En2):
“I will provide you with a sentence.
Count the number of commas in the text of the sentence and write only the number of commas.
*Sentence: **{sample}**”.*

4.3 Prompt sets for non-English languages

In this study, an elementary prompt was developed to perform the task of counting commas in a sentence. The elementary prompt (Ru1-1, Fi1-1), whose wording is transparent, was intentionally made difficult for the LLM to understand.

This study tested the invariance of back-translation of English cues. The potential effectiveness of this approach was hypothesized based on the fact that a large amount of English-language data is used to train multilingual LLMs, while the proportion of data in some non-English languages may be insignificant.

Table 1 presents the prompts for two languages other than English (Russian, Finnish). The wording of the first prompt with a conditional construction is presented with the prefix 1-, and the second prompt without a conditional construction is presented with the prefix 2- in the column **“**Substring **“**{prompt}**””**.

The second prompt meaning changes significantly when backtranslated through English, but for a native Russian speaker the differences between the formulations are not significant.

We calculate the word error rate (WER) metric for translation from the original language into English using the reference English text. The WER was also estimated for backtranslation through English to the original language. The obtained WER values were summed up for further analysis. The results obtained are presented in Table 2. The hypothesis of invariance to back translation suggests that if the final absolute value of WER between different prompts exceeds a certain threshold, then the probability of high-quality text processing by a LLM with such a prompt will be reduced.

Language	Original test prompt	Substring “{prompt}”	Notation
Russian	Я напишу тебе предложение. Перепиши его. Пожалуйста, отвечай на русском языке. {prompt} В ответе должно быть только переписанное предложение. Предложение: {sample}	Если внутри предложения есть слово "я", то удали все запятые из текста.	Ru1-1
		Обнаружив внутри текста предложения "я" в качестве слова, нельзя переписывать запятые из данного текста.	Ru1-2
	Я напишу тебе предложение. Перепиши его. Пожалуйста, отвечай на русском языке. {prompt} Предложение: {sample}	Ты обязан подсчитать количество запятых в тексте предложения и написать только количество запятых.	Ru2-1
		Посчитай запятых объём в тексте, написанном далее, предоставив для меня одну эту цифру только.	Ru2-2
Finnish	Kirjoitan sinulle ehdotuksen. Kirjoita se uudelleen. Vastaus venäjäksi. {prompt} Vastaus ei saa sisältää muuta kuin uudelleenkirjoitetun lauseen. Tarjous: {sample}	Jos lauseen sisällä on sana "mina", poista tekstistä kaikki pilkut.	Fi1-1
		Jos löydät sanan "I" lauseen tekstistä, et voi poistaa pilkkuja kyseisestä tekstistä.	Fi1-2
	Kirjoitan sinulle ehdotuksen. {prompt} Tarjous: {question}	Arvioi lauseessa olevien erottimien määrä ja merkitse vain pilkkujen määrä.	Fi2-1
		Laske, kuinka monta erotinmerkki koko tekstissä on, ja ilmoita pilkkua lukumäärä - tämä on pyyntö.	Fi2-2

Table 1: Prompts in non-English languages

Notation	Translation into English	Eng Tr. - Eng Orig. WER	Backtranslation	Src. Orig. - Src. Backtr. WER	WER Sum.
Ru1-1	If the word "I" is inside the sentence, then remove all commas from the text	0.21	Если слово "я" находится внутри предложения, то уберите все запятые из текста	0.5	0.71
Ru1-2	If you find "I" as a word inside the text of a sentence, you cannot rewrite commas from this text	1.0	Если вы встретите "I" в качестве слова в тексте предложения, вы не сможете переписать запятые из этого текста	1.00	2.0
Ru2-1	You must count the number of commas in the text of the sentence and write only the number of commas	0.22	Вы должны подсчитать количество запятых в тексте предложения и написать только их количество	0.31	0.53
Ru2-2	Count the volume of commas in the text written below, providing me with this one figure only	0.61	Подсчитайте количество запятых в тексте, написанном ниже, и получите только эту цифру	0.71	1.32
Fi1-1	Estimate the number of separators in the sentence and mark only the number of commas.	0.33	Arvioi lauseessa olevien erottimien määrä ja merkitse vain pilkkujen määrä.	0.0	0.33
Fi1-2	If you find the word "I" in the text of a sentence, you cannot remove commas from that text	0.79	Jos löydät sanan "I" lauseen tekstistä, et voi poistaa pilkkuja kyseisestä tekstistä	0.0	0.79
Fi2-1	Estimate the number of separators in the sentence and mark only the number of commas	0.00	Arvioi lauseessa olevien erottimien määrä ja merkitse vain pilkkujen määrä	0.0	0.00
Fi2-2	Calculate how many separator characters there are in the whole text, and indicate the number of comma - this is a request	1.06	Laske, kuinka monta erotinmerkkiä koko tekstissä on, ja ilmoita pilkun määrä - Tämä on pyyntö	0.27	1.33

Table 2: Results of WER metrics calculation

5 Experimental setup and Results

We consider 3 models: GPT-4, Mixtral and LLaMA. Information on the parameters for each LLM is presented in Table 3. The selection of parameters was based on the experience of using these models in RAG systems.

LLM	Configuration
gpt-4o-mini [12]	temperature: 0.14, top_p: 0.95, max_tokens: 4000
Mixtral-8x7B-Instruct-v0.1 [13]	temperature: 0.15, n_predict: 6000, top_p: 0.95, min_p: 0.05, repeat_penalty: 1.2, presence_penalty: 1
Meta-Llama-3.1-70B-Instruct [14]	temperature: 0.14, n_predict: 4000, top_p: 0.95, min_p: 0.04, repeat_penalty: 1.095, frequency_penalty: 0.01, presence_penalty: 1.3

Table 3: LLM Configuration

The LLM performance metrics were evaluated on the dataset described in detail in Section 3. For Russian and Finnish, experiments were performed with 2 equal-meaning prompts, the first of which (Ru1-1 and Fi1-1) was obtained using the Backtranslation Invariance-based approach described in Section 2, and the second by replacing words with synonyms or polysemous words.

To calculate the accuracy metric, the reference dataset was modified by applying strict rules im-plemented through regular expressions. For the first rule, described in prompts 1-1, 1-2, and En1, respectively, commas were removed only if the sample contained the letter “I” as a separate word. For the second rule, described in prompts 2-1 through 2-2 for non-English languages and En1 for English, all commas in the samples were counted.

The result of the LLM's work on the first proposal was considered correct if:

- commas were removed or counted according to the rule described in the prompt;
- the answer contained Cyrillic or only a number, as required by the prompt;
- the words contained in the answer were not glued together by removing space characters.

The result of the LLM's work on the second proposal was considered correct if:

- the first number that came up matched the number of commas in the sentence;
- the answer contained less than 50 words.

The results obtained for the Russian language are presented in Fig. 2 and Fig. 3. The graph in Fig. 2 shows the accuracy metrics for the case requiring the removal of commas when the sample contains the independent word “I”. Prompt En1 is the reference result, Ru1-1 has a formulation obtained by the developed approach, prompt Ru1-2 has a complicated formulation of the condition. Similar to the case without a condition, GPT and LLaMA show a result identical to that obtained for the English-language prompt - 61%. The difference in the performance of prompts Ru1-1 and Ru1-2 is 47% for GPT, 42% for LLaMA and 34% for Mixtral.

The graph in Fig. 3 shows the accuracy metrics for the case requiring counting the number of commas in a sample. Prompt En2 is the reference result, Ru2-1 has the formulation obtained by the developed approach, and prompt Ru2-2 has a complicated formulation of the problem. The best result using the developed approach reaches 53% using GPT, which is 19% higher than the result obtained using an ambiguously formulated prompt.

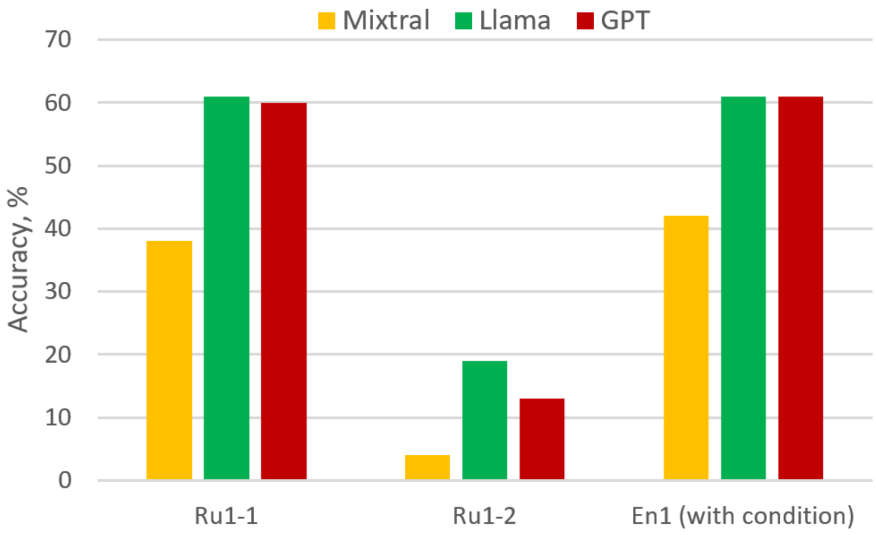

Figure 2: Accuracy metrics for Russian language (with condition)

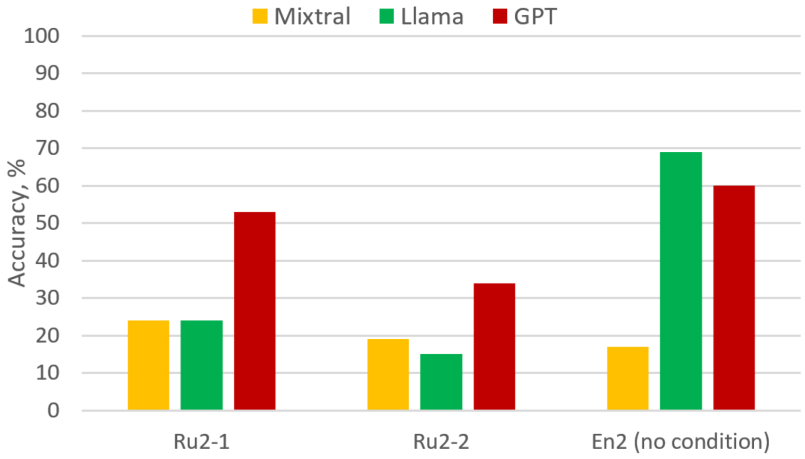

Figure 3: Accuracy metrics for Russian language (no condition)

The results obtained for Finnish are presented in Fig. 4 and 5. For the case with the condition (Fig. 4), GPT and LLaMA show the best result of 61% for both the reference prompt and the prompt obtained by the Backtranslation Invariance approach. The difference in accuracy between Fi1-1 and Fi1-2 reaches 54%. The graph in Fig. 5 shows the accuracy metrics for the case requiring counting the number of commas in a sample. Prompt Fi2-1 has the formulation obtained by the developed approach, and prompt Fi2-2 has a complicated formulation of the problem. The best result using the developed approach reaches 36% using GPT, which is 11% higher than the result obtained using an ambiguously formulated prompt.

It is important to note that the accuracy of Mixtral for Fi2-1 is slightly higher than that shown in the graph, this is due to the fact that, unlike other models that present results in digital format, its answer was presented in numerals.

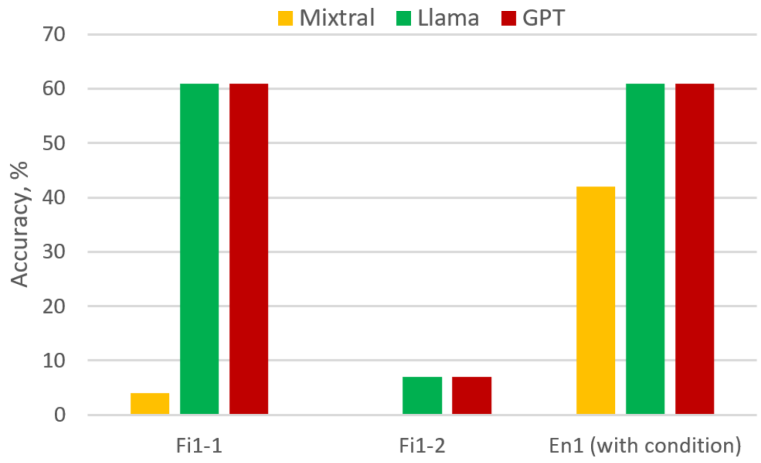

Figure 4: Accuracy metrics for Finnish (with condition)

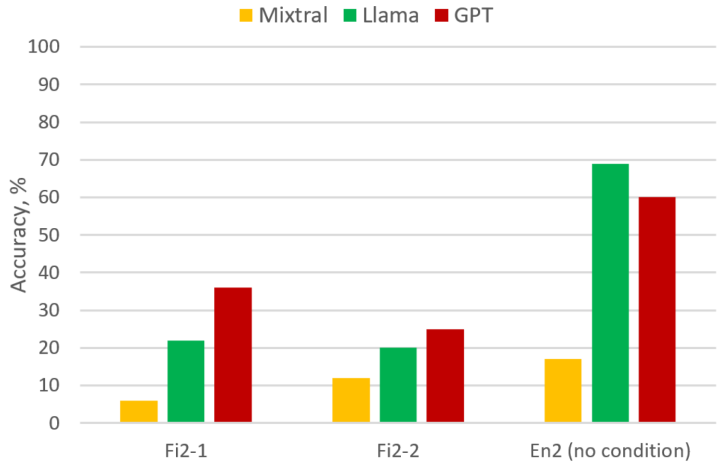

Figure 5: Accuracy metrics for Finnish (no condition)

According to the results of a series of experiments, an increase in the number of polysemous or synonymous words that may lose their original meaning when machine translated into English leads to a decrease in the ability of LLM to understand the context of the prompt and the tasks it contains. The obtained metrics also confirm that the approach based on backtranslation invariance allows achieving high efficiency of LLM, in some cases equal to that obtained when using an English-language prompt.

The key experimental results and comparison of the performance of LLM using the backtranslation invariance-based approach and without it are presented in Table 4.

Language	**LLM**	**Condition**	**Invariant prompt, %**	**Non-invariant prompt, %**	**Delta, %**
Russian	GPT	–	53	34	19
	LLaMA	+	61	19	**42**
Finnish	GPT	–	36	25	11
	LLaMA	+	61	7	**54**

Table 4: Summary of experimental results

6 Discussion

The results showed that the use of prompts with full back-translation invariance can positively affect the performance of language models for a given task. It is worth noting that only limited formulations of prompts were tested in this study. Although improvements in the performance of the language model were observed on specific examples, further research is needed to evaluate the scalability of the approach on more diverse datasets and tasks. It is important to study the applicability of the method to other languages, especially those where grammar and vocabulary features can affect the results of back-translation.

The current implementation of this approach requires significant manual effort. At the stage of prompt preparation, it is necessary to manually check their invariance using back-translation tools, such as Google and Yandex translators. This process can be labor-intensive and limit the scalability of the method. In the future, the approach can be automated by integrating with translator APIs. This will significantly speed up the process of checking the invariance of prompts. In addition, using synonym dictionaries to automatically substitute alternative formulations can simplify the creation of texts that are resistant to back translation. Automation of these processes will allow this method to be applied to a large number of diverse tasks.

7 Conclusion

In the course of the study, the particular effectiveness of Backtranslation Invariance for LLM prompting was established. The results of a series of experiments conducted for Russian and Finnish using three different LLMs showed that an unambiguous and transparent formulation of a non-English prompt allows achieving results comparable to those obtained with an English-language prompt. For each language, we evaluated the LLM with different formulations. The difference between the most confusing of them and the one obtained by the developed approach was 42% for Russian and 54% for Finnish. The best ability to understand the languages considered was demonstrated by the LLaMA-3.1 and GPT-4-o-mini models.

In future research, we plan to investigate Backtranslation Invariance in relation to languages other than English.

Acknowledgements

The authors would like to thank the anonymous reviewers who provided valuable feedback that significantly improved the article.

References

[1] Gabriel Nicholas, Aliya Bhatia. Lost in Translation: Large Language Models in Non-English Content Analysis // Computation and Language. – 2023. – Vol. arXiv:2306.07377. – Access mode: https://arxiv.org/abs/2306.07377.

[2] Makbule Gulcin Ozsoy. Multilingual Prompts in LLM-Based Recommenders: Performance Across Languages // Information Retrieval. – 2024. – Vol. arXiv: arXiv:2409.07604. – Access mode: https://arxiv.org/abs/2409.07604.

[3] Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, Grzegorz Kondrak. Don't Trust ChatGPT when Your Question is not in English: A Study of Multilingual Abilities and Types of LLMs // Computation and Language. – 2023. – Vol. arXiv:2305.16339. – Access mode: https://arxiv.org/abs/2305.16339.

[4] Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, Xuanjing Huang. LLaMA Beyond English: An Empirical Study on Language Capability Transfer // Computation and Language. – 2024. – Vol. arXiv:2401.01055. – Access mode: https://arxiv.org/abs/2401.01055.

[5] Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, Fei Yuan. MindMerger: Efficient Boosting LLM Reasoning in non-English Languages // Computation and Language. – 2024. – Vol. arXiv:2405.17386. – Access mode: https://arxiv.org/abs/2405.17386.

[6] Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, Sadao Kurohashi. Beyond English-Centric LLMs: What Language Do Multilingual Language Models Think in? // Computation and Language. – 2025. – Vol. arXiv:2408.10811. – Access mode: https://arxiv.org/abs/2408.10811.

[7] Marion Di Marco, Alexander Fraser. Subword Segmentation in LLMs: Looking at Inflection and Consistency // Association for Computational Linguistics. – 2024. – Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. – P. 12050–12060. – Access mode: https://aclanthology.org/2024.emnlp-main.672.
[8] Garn Rimma. Interactive Russian Grammar: The Case System // Journal of the National Council of Less Commonly Taught Languages. – 2009. – Vol. 6. – P. 37–58.
[9] Timberlake Alan. A Reference Grammar of Russian // Cambridge University Press: Reference Grammars. – 510 p.
[10] Valentin Hofmann, Leonie Weissweiler, David Mortensen, Hinrich Schütze, Janet Pierrehumbert. Derivational Morphology Reveals Analogical Generalization in Large Language Models // Computation and Language. – 2024. – Vol. arXiv:2411.07990. – Access mode: https://arxiv.org/abs/2411.07990.
[11] Laurie Bauer. The function of word-formation and the inflection-derivation distinction // Words in their Places. A Festschrift for J. Lachlan Mackenzie. – 2004. – Access mode: https://www.wgtn.ac.nz/lals/about/staff/publications/Bauer-Infl-Deriv.pdf.
[12] Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, Anders Søgaard. Word Order Does Matter and Shuffled Language Models Know It // Association for Computational Linguistics. – 2022. – Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – Access mode: https://aclanthology.org/2022.acl-long.476.
[13] GPT-4o mini: advancing cost-efficient intelligence. – Access mode: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.
[14] Albert Q. Jiang et al. Mixtral of Experts // Machine Learning. – Vol. arXiv:2401.04088. – Access mode: https://arxiv.org/abs/2401.04088.
[15] Llama-3.1-70B-Instruct. – Access mode: https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct.