Can LLMs Handle Low-Resource Dialects? A Case Study on Translation and Common Sense Reasoning in Šariš

Viktória Ondrejová and **Marek Šuppa**
Department of Applied Informatics, Comenius University in Bratislava
NaiveNeuron
Cisco Systems
ondrejova18@uniba.sk, marek.suppa@fmph.uniba.sk

Abstract

While Large Language Models (LLMs) have demonstrated considerable potential in advancing natural language processing in dialect-specific contexts, their effectiveness in these settings has yet to be thoroughly assessed. This study introduces a case study on Šariš, a dialect of Slovak, which is itself a language with fewer resources, focusing on Machine Translation and Common Sense Reasoning tasks. We employ LLMs in a zero-shot configuration and for data augmentation to refine Slovak-Šariš and Šariš-Slovak translation models. The accuracy of these models is then manually verified by native speakers. Additionally, we introduce ŠarišCOPA, a new dataset for causal common sense reasoning, which, alongside SlovakCOPA, serves to evaluate LLM's performance in a zero-shot framework. Our findings highlight LLM's capabilities in processing low-resource dialects and suggest a viable approach for initiating dialect-specific translation models in such contexts.

1 Introduction

The recent explosion of development in the field of Large Language Models (LLMs) has offered an unprecedented set of capabilities in understanding, generating, translating and transforming text across a large number of contexts (Min et al., 2023). However, despite their wide-ranging applications, the effectiveness of LLMs in dialect-specific scenarios, particularly in languages with limited resources, remains a relatively unexplored domain. This gap in research presents a critical challenge, as dialects incorporate distinct linguistic traits and cultural subtleties, yet comprehensive large-scale datasets like newswire texts are not available for them.

This study aims to address this gap by focusing on Šariš, a Slovak dialect, with pronounced linguistic variety as shown in Table 1. As Slovak is a less-resourced language itself, it presents an interesting case for examining how large language models (LLMs) perform in specific dialect contexts where data is scarce. We focus on two key natural language processing (NLP) tasks: Machine Translation (MT) and Common Sense Reasoning (CSR), which we view as representative for assessing the model's ability to handle the complexities of real-world language.

English	I left the potatoes in the fridge.
Slovak	Nechal som zemiaky v chladničke.
Šariš	Ochabil som gruľe v chladňičke.
	Zochabil som bandurki v ľadňičke.

Table 1: An example of expressing a singular statement through various linguistic constructions in the Šariš dialect. Note that both of the listed examples were deemed valid and reasonable by a native speaker.

In terms of MT, we investigate how LLMs can aid in translating between Slovak and the Šaris dialect. Here the LLMs are first used in zero-shot setting, meaning that we assume that (to the best of our knowledge) the models are not directly trained with Šariš-specific data but are instead expected to apply their knowledge of Slovak to understand and translate Šariš. We use this approach both to evaluate the performance of LLMs on the Slovak → Šariš and Šariš → Slovak translation task as well as for data augmentation, which results in about 3,500 automatically translated Slovak-Šariš sentence pairs. These are then used to finetune a specific Slovak-Šariš translation model, whose performance is evaluated on a manually labelled test set.

Additionally, we further introduce a new dataset called ŠarišCOPA, designed to evaluate the model's performance in CSR tasks specifically in the Šariš dialect. This dataset is intended to complement an existing dataset for Slovak, SlovakCOPA, to compare how the models perform in understanding both the standard language and its dialect. In this case the LLM is first prompted to only output the CSR

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 130–139
June 20, 2024

classification directly while additional experiments with a prompt-specific ”translate-test” approach are also evaluated.

Our contributions can thus be summarized as follows:

- We introduce the first Slovak-Šariš translation dataset and use it to finetune a Šariš specific Machine Translation model
- We manually evaluate the quality of the translations produced by the finetuned model, as well as leading LLMs
- We introduce the ŠarišCOPA dataset and use it to evaluate the common sense reasoning performance of LLMs in Šariš
- We experiment with various LLM prompting approaches for ŠarišCOPA, including translation to English and Slovak

We release the code and data associated with our experiments in the hopes of fostering possible future research in this area at https://github.com/NaiveNeuron/saris.

2 Slovak and its Dialects

Despite being a relatively small language in terms of the number of native speakers (roughly 5 million native speakers), Slovak has multiple dialects. In this work, we focus on the eastern part of Slovakia, where the majority of population speak in a multiple dialects from the Šariš, Spiš, Zemplín regions. Even though we categorize these dialects to distinct groups, their historical, phonetic and lexical features are intertwined. A substantial overlap exists in lexical terms between dialects, with minimal variance observed (Pavlíková, 2016). Additionally, instances occur where native speakers interchange words from different dialects within the same discourse. Given these linguistic dynamics, in pursuit of maximizing corpus size, we considered amalgamation of all 3 of the dialects eligible for extraction.

The **Šariš** dialect holds notable significance within the family circle of the Prešov region, where a substantial portion of the population consistently employs it in their daily interactions. Specifically, statistics published in (Vodičková, 2009) reveal that approximately 22.5% of the population, amounting to roughly 180 thousand speakers, within the Prešov region utilize the Šariš dialect as their primary mode of communication. From the broader perspective Šariš, as an Eastern Slovak dialect, is classified as ”Vulnerable” by the UNESCO Atlas of the World’s Languages in Danger (Moseley, 2010).

2.1 Šariš-Specific Challenges

The dialect lacks a formal codification, leading to an absence of definitive linguistic rules governing their usage in speech and writing. Consequently, dialectal variations manifest across different areas which can be as small as villages, resulting in multiple potential translations for a single word within the same dialect. An example of this phenomenon can be seen in Table 1.

Conversely, Eastern Slovakian dialects exhibit distinct features. Unlike standard Slovak, these dialects lack long vowels. The Slovak ”ď” ([d^{j}] in IPA) and ”ť” ([t^{j}] in IPA) are replaced by ”c” ([ts] in IPA) and ”dz” ([dz] in IPA), respectively. Most importantly, however, a majority of Eastern Slovakian dialects, including those of the Šariš region, do not include the vowel ”y”.

Another challenge arises from the fact that certain highly specific terms either cannot be adequately translated into Slovak or risk losing their intended meaning. Additionally, the Šariš dialect incorporates numerous archaic expressions that have fallen out of common usage, making them potentially incomprehensible to some speakers.

3 The Translation Task

Our aim with the translation task is to validate to what extent are the findings of (Gu et al., 2018) still relevant, which found that less than 13k sentence pairs are not enough to train a neural machine translation model to reasonable quality. To this end, we introduce the ŠarišSet corpora with the help of a LLM.

3.1 ŠarišSet

Creating a corpus for a new language presents substantial challenges. The ŠarišSet dataset, containing over 4,000 sentences in the Šariš dialect from Eastern Slovakia, was compiled from various online sources. To ensure a solid benchmark, a subset of 500 sentences received manual translation by three native speakers[1]. The bulk of the dataset was translated through a hybrid method combining prompt engineering with manual review of outputs

[1] Here, ”native speaker” refers to someone fluent in the Šariš dialect with extensive exposure from childhood.

	Šariš	Slovak
vocabulary size	3560	3647
Q1	11	11
Median	16	17
Q3	23	23
Mean	18.98	19.23
SD	11.46	11.77

Table 2: The table shows the quantitative statistics of the ŠarišSet dataset as vocabulary size for the source and target languages, as well as the Q1, Q2, Median and Mean of the number of words per sentence. In addition, the standard deviation is displayed in the SD row.

from GPT-3.5-Turbo and GPT-4 (Achiam et al., 2023).

The Table 2 shows the aggregated statistics related to the dataset, such as its vocabulary size and quantitative statistics for the introduced dataset.

Extraction In order to gather data for a dialect of a low-resource language spoken by only a few tens of thousands of individuals, the conventional automated methodology proved unfeasible. With scarce online resources beyond traditional folk songs, the absence of suitable web pages for scraping presented a challenge. Šariš texts are predominantly confined to a handful of niche blogs and sporadic Facebook posts. To avoid the complexities of the Facebook interface, our focus was directed solely towards the identified blogs outlined in Appendix A, discovered through extensive online searches (mainly by searching a very specific word in the dialect), alongside the aforementioned folk songs which could be systematically scraped using the `scrapy` library in Python[2].

Throughout the scraping process, filtering criteria were implemented. The native speaker visually inspected the texts, reviewing the initial and final two sentences. If the sentences appeared plausible, with words in their proper positions and the structure intact, the text was kept and saved. The acquired data subsequently underwent a cleaning process via a script designed to remove duplicates, highly offensive language, extraneous characters, and segment the text into coherent sentences.

The final sentences originate from 133 various longer texts obtained from multiple blogs, together with more than 170 folk songs.

[2] https://scrapy.org/

Automatic Translation Given the laborious nature of manual translation, we opted to employ the GPT-3.5-Turbo and GPT-4 models for translating the remaining sentences, comparing their performance using various prompt engineering techniques.

Initially, we focused on the GPT-3.5-Turbo model, experimenting with three distinct prompts. The first prompt, applied to both models, was straightforward as we can see in Figure 1.

```
translate to Slovak
```

Figure 1: The first simplest prompt used for translation.

We further tested a more nuanced prompt, encouraging the model by stating that even an inaccurate translation would be beneficial (see Figure 2).

```
Please, try to translate this into Slovak, even an inaccurate version would help a ton
```

Figure 2: The second prompt used for translation.

Finally, we utilized a persona-based approach, directing the model to take on the role of a bilingual eastern Slovak youth proficient in translating dialects into Slovak. The prompt, visible in Figure 3, presented a scenario where the model was a native Šariš dialect speaker conversing with someone unfamiliar with it.

```
You're an eastern Slovak young man who has lived in one village his entire life. Though you are proficient in Slovak due to schooling, at home with your family, you speak in the eastern Slovak dialect known as Šariština. You've introduced a girl from central Slovakia, fluent in Slovak but unfamiliar with Šariština, to your family and need to provide the most accurate translation of this sentence into Slovak
```

Figure 3: The third prompt used for translation that employed the persona-based approach.

A selection of model responses is illustrated in Table 8.

We manually evaluated these results across 50 sentences, selecting the most suitable translation from the three generated ones. Surprisingly, the translations from Prompt 2 proved to be highly comparable to those from Prompt 3, despite the added narrative context. Ultimately, we chose the second prompt due to fewer instances of extraneous words in the final outputs.

When it finally came to the translating the reminder of ”ŠarišSet”, it was necessary to decide between using GPT-3.5-Turbo and GPT-4. Utilizing a similar approach as above, we evaluate the results of each on 50 sentences and concluded that GPT-4 is a better fit for this sort of a translation task and was used to translate the remaining 3,500 sentences using the Prompt 2 chosen before. The same prompt was then used for translation of the test set as well. Additional details on how these models were accessed can be found in Section B.

3.2 NLLB-Based Model

In the very first iteration we experimented with the mBART model (Tang et al., 2020), specifically the mBART-50 version that was created by multilingual fine-tuning. Perhaps owing to the fact that Slovak was not included in the languages it was pretrained on, the model tended to collapse to outputting a single word and not being useful at all.

As an alternative to the mBART model, we also experimented with the NLLB-200 model which was created as part of the No Language Left Behind project (Costa-jussà et al., 2022). The aim of this project is to provide open-source models ”capable of delivering evaluated, high-quality translations directly between 200+ languages – including low-resource languages”[3]. The list of 204 languages does not include Šariš but as opposed to mBART, it does include Slovak (which (Costa-jussà et al., 2022) even lists as being high resourced on page 15 in Table 1) and hence we opted to experiment with using it as the basis for the Šariš → Slovak and Slovak → Šariš translation models. We did so by adding a new ”pseudo language” tag `sar_Latn` to the model and finetuning it on the dataset introduced in Section 3.1. We finetuned the model, specifically its `nllb-200-distilled-600M` version[4], with the batch size of 16, 500 warm up steps and 20 000 training steps. Additionally, the maximum output length was set to 128.

[3] https://ai.meta.com/research/no-language-left-behind/

[4] https://huggingface.co/facebook/nllb-200-distilled-600M

	Š → S		S → Š	
	F	**A**	**F**	**A**
GPT-3.5-Turbo	2.96	3.15	1.02	1.23
GPT-4	**3.45**	**3.51**	1.17	1.57
NLLB	3.09	3.00	**3.16**	**3.80**

Table 3: The average fluency (F) and adequacy (A) obtained during evaluation of various models and translation directions. Š represents Šariš and S represents Slovak. The best result per each metric and language pair is boldfaced.

3.3 Evaluation

In our experimental framework, we utilize adequacy and fluency metrics (Chatzikoumi, 2020) to manually evaluate the outputs generated by the machine translation models. Each output, corresponding to a given source text, underwent assessment by an annotator on a graded scale ranging from 1 to 5, where the higher numbers represent better adequacy and fluency.

In terms of adequacy, we are primarily concerned with whether the output effectively conveys the same meaning as the input sentence. We evaluate whether any part of the original message is lost, added, or distorted during the translation process. Therefore, the rating of 5 signifies preservation of all semantic aspects from the source text, whereas a score of 1 indicates complete loss of meaning.

Regarding fluency, our focus lies in assessing whether the output exhibits fluent expression in the target language. This entails considerations of grammatical correctness and the use of idiomatic word choices to ensure that the translated text reads naturally and smoothly. Likewise, a fluency score of 5 indicates seamless language coherence in alignment with the intended target output, whereas a score of 1 suggests incomprehensibility.

During evaluation, we conducted comparisons between the translated sentences. If a text contained 1-2 errors (untranslated words, mismatched case ending and so on), it would receive a score of 4. Conversely, if the translated sentence exhibited only 1-2 accurately translated words and rest was implausible, it would be awarded a score of 1, and so forth.

The evaluation, conducted by a native speaker and detailed in Table 3, indicates that GPT-4 excelled in translating from Šariš to Slovak, while the NLLB model reported the best performance in the opposite direction. Notably, both GPT-3.5-Turbo

	PREMISE		CHOICE 1	CHOICE 2
sk en	*Vonku sa zotmelo.* It got dark outside.	R	*Z oblohy začali padať snehové vločky.* Snowflakes began to fall from the sky.	*Na oblohe sa objavil mesiac.* The moon became visible in the sky.
šr en	*Šľisknul som še na žemi.* I slipped on the floor.	C	*Kachľička bula prasknuta.* The tile was cracked.	*Kachľička bula morka.* The tile was wet.

Table 4: Examples of forward (Result [R]) and backward reasoning (Cause [C]) in the COPA, SlovakCOPA and ŠarišCOPA validation sets. Note that Šariš is denoted as šr in the list of languages.

and GPT-4 showed poor performance in translating from Slovak to Šariš, indicating a challenge in producing coherent Šariš output. Conversely, GPT-4's superior performance in translating to Slovak, surpassing even the fine-tuned NLLB model, underscores the importance of language-specific proficiency in LLM-based translation.

4 The Common Sense Reasoning Task

To gauge the effectiveness of natural language processing (NLP) systems in understanding different languages, it is crucial to employ various testing methods. Common sense reasoning evaluation is particularly significant, as it is a fundamental aspect of these systems, underscored by previous research (Davis and Marcus, 2015). The Choice Of Plausible Alternatives (COPA) serves as a notable benchmark, testing systems' ability to decipher cause-and-effect relationships in English sentence pairs (Roemmele et al., 2011). Due to its acclaim, COPA has been expanded into multiple languages through the XCOPA benchmark (Ponti et al., 2020) and adapted for Slavic languages such as Slovenian (Ljubešić et al., 2022a), Serbian (Ljubešić et al., 2022b), and Croatian (Ljubešić, 2021). Our study introduces the ŠarišCOPA dataset, focusing on the Šariš dialect.

4.1 ŠarišCOPA

The COPA framework is generally implemented as a binary classification challenge. Models must choose the more plausible scenario from two options, based on a given premise and question. This assessment distinguishes between cause and effect in scenarios: "cause" questions ask for the reason behind an event while "effect" questions seek the consequence of an event.

The ŠarišCOPA dataset, designed to test LLMs' common sense reasoning in Šariš, consists of 500 test and 100 validation triplets, each with a premise and two choices. It adapts the original English COPA, following XCOPA's translation methodology (Ponti et al., 2020), with the translation work carried out by native speakers from the ŠarišSet project. Additionally, we compare results with the SlovakCOPA dataset, created by a professional translator using a similar method. The format and examples of these datasets are displayed in Table 4.

4.2 Evaluation

Our evaluation of the SlovakCOPA and ŠarišCOPA datasets began with comparing native speaker labels to those from the original COPA dataset, revealing a 100% match in both cases.

Subsequently, we tested GPT-3.5-Turbo and GPT-4 on these datasets using specific prompts for the "cause" as well as the "efect" scenario. These prompts were inspired by the prompts used by "BENCHić - the benchmark for Bosnian, Croatian, Montenegrin, Serbian (and friends)"[5]. They were designed to minimize the amount of noise in the responses of LLMs and their full text can be seen below:

COPA Prompt: Cause

Given the premise "premise", and that we are looking for the cause of this premise, which hypothesis seems more plausible?
Hypothesis 1: "hypothesis1".
Hypothesis 2: "hypothesis2".
Please answer only with "1" or "2".

COPA Prompt: Effect

Given the premise "premise", and that we are wondering what happened as a result of this premise, which hypothesis seems more plausible?
Hypothesis 1: "hypothesis1".
Hypothesis 2: "hypothesis2".
Please answer only with "1" or "2".

[5] This benchmark can be found at https://github.com/clarinsi/benchich/tree/main/copa

As Table 5 shows, GPT-3.5-Turbo performed well on SlovakCOPA (76.6% accuracy) but struggled with ŠarišCOPA (55.4% accuracy, near random chance). GPT-4 showed remarkable performance on SlovakCOPA (96.6% accuracy) and significantly outperformed GPT-3.5-Turbo on ŠarišCOPA (79.8% accuracy), albeit with a 4.8% rate of unparseable responses, such as ”*As an AI language model, I'm unable to understand the premise and hypotheses because they are not in a recognizable language or a standard linguistic structure. Therefore, I can't determine which hypothesis is more plausible.*”.

We also tested a method where the model first translates the input into a more resource-rich language before making a prediction. This approach, inspired by the performance of GPT-4 in Šariš to Slovak translation and recent research on multilinguality in LLMs (Liu et al., 2024) and cross-lingual transfer (Ebing and Glavas, 2023), involved slightly modified prompts for translation into English and Slovak, which can be found below.

COPA Prompt: Cause with translation

Given the premise ”premise”, and that we are looking for the cause of this premise, which hypothesis seems more plausible?
Hypothesis 1: ”hypothesis1”.
Hypothesis 2: ”hypothesis2”.

First translate the premise and the hypotheses to English, then answer only with ”Prediction: 1” or ”Prediction: 2”.

COPA Prompt: Effect with translation

Given the premise ”premise”, and that we are wondering what happened as a result of this premise, which hypothesis seems more plausible?
Hypothesis 1: ”hypothesis1”.
Hypothesis 2: ”hypothesis2”.

First translate the premise and the hypotheses to English, then answer only with ”Prediction: 1” or ”Prediction: 2”.

The results, labeled ”+ translate en” and ”+ translate sk” in Table 5, showed that translating to English improved GPT-3.5-Turbo's performance on SlovakCOPA (from 76.6 to 88.0) and ŠarišCOPA (from 55.4 to 71.0), with a slight increase for GPT-4 on ŠarišCOPA (from 79.8 to 82.0). Translating to Slovak yielded less pronounced improvements. Interestingly, the number of unparseable responses increased, including ”*The text provided is not in a recognizable language, therefore it cannot be translated or used to make a prediction.*” in English and ”*The premise and hypotheses are already in Slovak, but they are written in a dialect or with many spelling mistakes, making them difficult to understand. Therefore, it is impossible to make a prediction.*” in Slovak, hinting at GPT-4's ability to recognize Šariš as a Slovak dialect.

Model	SlovakCOPA	ŠarišCOPA
GPT-3.5-Turbo	76.6 (0.0)	55.4 (0.0)
+ translate en	**88.0** (0.2)	**71.0** (0.4)
+ translate sk		70.0 (0.4)
GPT-4	**96.6** (0.0)	79.8 (4.8)
+ translate en	**96.6** (0.0)	**82.0** (8.6)
+ translate sk		81.6 (8.8)

Table 5: The accuracy of GPT 3.5 Turbo and GPT 4 on the SlovakCOPA and ŠarišCOPA datasets. The number in parentheses denotes the number of responses that we were unable to parse. The best performing model in a specific model family on a particular dataset is boldfaced.

5 Discussion

This study aimed to assess the proficiency of large language models (LLMs) in processing the Šariš dialect, a low-resource variant of Slovak. Our investigation, detailed in Section 3, showcased GPT-4's ability to translate between Slovak and Šariš, albeit with varying success, particularly in Šariš-targeted translations. Enhancing the NLLB model with GPT-4's Šariš translations significantly improved its performance, outstripping GPT-3.5-Turbo in Slovak to Šariš translation accuracy and surpassing both GPT iterations in the opposite direction. This indicates that leveraging LLMs for initial translations can create a solid foundation for building effective translation tools for underrepresented dialects, as demonstrated by our results with just 3,500 sentences.

Furthermore, as detailed in Section 4, translation plays a crucial role in the Common Sense Reasoning Task. Having models translate inputs to English or Slovak before making inferences improved the outcomes for both GPT-3.5-Turbo and GPT-4, with English translations being marginally more effective. Intriguingly, GPT-4 occasionally declined to make predictions, identifying inputs as specific Slovak dialects or variants, indicating its potential in dialect recognition, despite limitations in dialect generation.

In summary, our experiments illustrate that LLMs have the potential to be instrumental in handling dialects with scarce resources. By integrating strategic prompting with LLMs, we cannot only enhance model performance but also empower subsequent models trained on the data produced, setting a promising direction for future research in NLP for low-resource dialects.

6 Related Work

Machine translation for low-resource languages and dialects has been an active area, often leveraging transfer learning from high-resource languages (Tars et al., 2021; Maimaiti et al., 2019). Dialect translation has been studied for Arabic (Harrat et al., 2019), German (Honnet et al., 2018), Portuguese (Costa-jussà et al., 2018) and French, Croatian, Serbian and Malay (Lakew et al., 2018) dialects, finding substantial data in the dialect language is beneficial.

The Choice of Plausible Alternatives (COPA) dataset (Roemmele et al., 2011) has been widely used to evaluate commonsense causal reasoning in English, and has further been translated into 11 languages, including resource-poor languages like Haitian Creole as part of XCOPA (Ponti et al., 2020) and separately into Slavic languages as well (Ljubešić, 2021; Ljubešić et al., 2022a,b). Analysis has found translate-test approaches can boost performance over zero-shot cross-lingual transfer (Artetxe et al., 2023), aligning with our findings. Our ŠarišCOPA dataset provides a new test for reasoning in a low-resource dialect context.

While Slovak is considered a lower-resource language compared to major world languages, there has been some prior work on developing NLP tools and resources. This includes machine translation systems focused on European languages (Popel, 2018), pre-trained language models like SlovakBERT (Pikuliak et al., 2022) and annotated datasets for tasks like named entity recognition (Suba et al., 2023) and question answering (Hládek et al., 2023). However, work specifically targeting Slovak dialects like Šariš has been very limited. Perhaps the closest work to ours would be (Darjaa et al., 2018) in which the authors conduct a preliminary analysis on the distinguishability of Slovak dialects in spoken language and introduce the Sound Archive of Slovak Dialects – roughly 150 hours of recordings which include all basic Slovak dialects. To the best of our knowledge, our work is the first to investigate the use of Natural Language Processing specifically on texts in Slovak dialects.

7 Conclusion

This study assesses LLMs' abilities in translating and understanding the Šariš dialect through machine translation and common sense reasoning tasks, introducing the ŠarišCOPA dataset. While LLMs show proficiency in translating from Šariš to Slovak, reverse translations pose challenges. The inclusion of translation as a preprocessing step improved common sense reasoning performance, particularly notable when comparing results on ŠarišCOPA with SlovakCOPA. These findings highlight the potential and limitations of LLMs in processing and reasoning in low-resource dialects. The code and data associated with our experiments can be found at `https://github.com/NaiveNeuron/saris`.

Limitations

Data Scarcity Despite our efforts, the amount of Šariš data we could obtain remains very limited compared to standard benchmarks for high-resource languages. The ŠarišSet corpus contains only around 4,000 sentences, and ŠarišCOPA has just 600 examples. This scarcity makes it difficult to fully assess LLM capabilities and prevents training extremely high-performing dialect-specific models from scratch. Obtaining more in-domain data would strengthen future analyses.

Human Evaluation Our human evaluations of translation quality and the ŠarišCOPA dataset drew upon a limited number of native Šariš speakers. While we took care to involve highly proficient speakers, from multiple parts of the Šariš region, inherent subjectivity in such evaluations means the ratings may not fully generalize. A larger evaluation involving more speakers would increase confidence. Additionally, no standard evaluation datasets exist for Šariš, preventing benchmarking against prior work.

Model Limitations The prominent LLMs like GPT-3, GPT-4, and NLLB that we evaluated are large models trained primarily on text from high-resource languages. While their pretraining data likely contained little-to-no examples of low-resource dialects like Šariš, it is difficult to claim that with certainty – particularly for models that are not publicly released, which further hinders the reproducibility of our experiments.

Reasoning About Dialect While our ŠarišCOPA probe provides a window into LLM's commonsense reasoning abilities for the dialect, the examples come from a single constructed dataset. Drawing broader conclusions about general language understanding of Šariš from this limited test would be an overreach. More comprehensive benchmarks probing other core language skills are needed.

Acknowledgements

This research was partially supported by grant APVV-21-0114. Last but not least, we would like to thank our annotators, in particular Pavol Hudák, Daniel Haraksim, Matej Novotný and Matúš Kobyl'an from Technical University of Košice (TUKE) for their help with the Šariš to Slovak and English to Šariš translation and Terézia Šuppová for her help with English to Slovak translation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Marta R. Costa-jussà, Marcos Zampieri, and Santanu Pal. 2018. A neural approach to language variety translation. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 275–282, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sakhia Darjaa, Róbert Sabo, Marián Trnka, Milan Rusko, and Gabriela Múcsková. 2018. Automatic recognition of slovak regional dialects. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 305–308. IEEE.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Benedikt Ebing and Goran Glavas. 2023. To translate or not to translate: A systematic investigation of translation-based cross-lingual transfer to low-resource languages. *ArXiv*, abs/2311.09404.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine translation for arabic dialects (survey). *Information Processing & Management*, 56(2):262–273.

Daniel Hládek, Ján Staš, Jozef Juhár, and Tomáš Koctúr. 2023. Slovak dataset for multilingual question answering. *IEEE Access*, 11:32869–32881.

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine translation of low-resource spoken dialects: Strategies for normalizing Swiss German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Surafel Melaku Lakew, Aliia Erofeeva, and Marcello Federico. 2018. Neural machine translation into language varieties. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 156–164, Brussels, Belgium. Association for Computational Linguistics.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models.

Nikola Ljubešić. 2021. Choice of plausible alternatives dataset in croatian COPA-HR. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Boshko Koloski, Kristina Zdravkovska, and Taja Kuzman. 2022a. Choice of plausible alternatives dataset in macedonian COPA-MK. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Mirjana Starović, Taja Kuzman, and Tanja Samardžić. 2022b. Choice of plausible alternatives dataset in serbian COPA-SR. Slovenian language resource repository CLARIN.SI.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-round transfer learning for low-resource nmt using multiple high-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(4).

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2).

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Michaela Pavlíková. 2016. Východoslovenské nářečí v psaném textu. SUPERVISOR: prof. PhDr. Marie Krčmová, CSc.

Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marian Simko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. 2022. SlovakBERT: Slovak masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7156–7168, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Martin Popel. 2018. CUNI transformer neural MT system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 482–487, Belgium, Brussels. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

David Suba, Marek Suppa, Jozef Kubik, Endre Hamerlik, and Martin Takac. 2023. WikiGoldSK: Annotated dataset, baselines and few-shot learning experiments for Slovak named entity recognition. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 138–145, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Maali Tars, Andre Tättar, and Mark Fishel. 2021. Extremely low-resource machine translation for closely related languages. *CoRR*, abs/2105.13065.

Zuzana Vodičková. 2009. Šarišská nárečová lexika.

A Data Sources

The list below outlines the domains which were ultimately used for extraction of the ŠarisSet. The majority of the sentences were obtained from various local newspapers, blogs and folk tales found on the following internet pages:

- https://www.obeckrivany.sk/
- https://bandzone.cz/
- https://www.ilonas.net/valal/
- https://prerag.sk/
- https://blog.pravda.sk/

Similarly, for obtaining the texts of various Šariš folk songs, the following domains were scraped:

- https://narecie.sk/
- https://www.videorohal.com/

B Details on Accessing GPT-3.5-Turbo and GPT-4

All models were accessed via the AzureOpenAI endpoints[6], with the API version being set to `2023-07-01-preview` and the temperature=0 to aid reproducibility.

C GPT-4 Translations

The examples of good and bad translations from Šariš to Slovak obtained from the GPT-4 model oaired with the Prompt 2 can be seen in Table 6 and Table 7, respectively.

D Translation Prompts

In Table 8 we can see the original sentence in Šariš and its Slovak manual translation; together with the translations obtained from GPT-3.5-Turbo model using Prompts 1, 2 and 3.

[6] `https://learn.microsoft.com/en-us/azure/ai-services/openai/reference`

Šariš dialect	**Slovak**
”Ta ňeznam,” - hvarim ja jej, ”ja ňemam kedi, hibaľ na večar?”	”Ja neviem,” hovorím jej, ”ja nemám kedy, možno večer?”
Dok sme buľi malki, naša mama nam veľo času pošvecovala.	Keď sme boli malí, naša mama nám venovala veľa času.
Počali sme medži sobu bešedovac.	Začali sme medzi sebou rozhovor.
Dovidzeňa i ščešľivo - skričal som jej.	”Dovidenia a šťastlivo,” zakričal som na ňu.
Šofer po nas vriskal, naj śe sciśime, bo ho to ruši.	Vodič na nás kričal, aby sme prestali vrieskať, pretože ho to rušilo.

Table 6: Examples of a bad translation from Šariš dialect to Slovak obtained from the GPT-4 model.

Šariš dialect	**Slovak**
Ona vichovala parobkom frajirku.	Ona obdaruvala chlapca drahým darčekom.
Taže me ňemaľi aňi dze śedzec, aňi dze spac.	Teda ma nemajú ani na seďať, ani na spať.
Jak me prišľi na chižu, ta me muśeľi stolok podkladac.	Keď ma priviezli na pohotovosť, museli mi podkladať stôl.

Table 7: Examples of a good translation from Šariš dialect to Slovak obtained from the GPT-4 model.

	Translations
Šaris	Heslo jej babi bulo: ”ked na dakoho ňemožeš patric zos očmi, popatri na ňho zos šercom”
Slovak	Heslo jej babky bolo: ”Keď na niekoho nemôžeš pozrieť s očami, pozri naňho so srdcom”.
Prompt 1	Heslo jej babičky bolo: ”Keď na niekoho nemôžeš hľadieť očami, pozri naňho srdcom.”
Prompt 2	Heslo jej babičky bolo: ”Keď na niekoho nemôžeš pozerať očami, pozri naňho srdcom.”
Prompt 3	Heslo jej babičky bolo: ”Keď na niekoho nemôžeš pozerat z očí, popozri na neho z srdcom”.
Šariš	Ponežе še M. ľubel chvaľic, i ponosovac, zos veľima stvarami, vše ked prichodzel.
Slovak	Pretože sa M. rád chválil aj povyšoval s veľkými svárami všade kde prišiel.
Prompt 1	Áno, napríklad, oni sa hnevali na M., keď im niečo prekladal, lebo, áno, všetko rozumeli.
Prompt 2	Pretože sa M. ľúbil chváliť a pýšiť sa mnohými vecami, keďže prichádzal.
Prompt 3	Tak, napríklad, oni sa hnevali na M., keď im niečo prekladal, pretože, áno, všetko rozumeli.
Šariš	”Ta ňeznam,”- hvarim ja jej, ”ja ňemam kedi, hibaľ na večar?”
Slovak	”Tak neviem” - hovorím jej ja, ”ja nemám kedy, možno večer?”
Prompt 1	”Ta neviem,”- pýtam sa jej, ”ja nemám plán, ideš večerať?”
Prompt 2	”Ta neviem,” - hvarím jej, ”ja nemám mačku, alebo ísť večer?”
Prompt 3	”Neznamená to,” hovorím jej, ”ja nemám práve teraz čas, ideme večerať?”

Table 8: Comparison of Šariš sentences with Slovak translations and translations from Prompts 1, 2, and 3 obtained from GPT-3.5-Turbo.