JSI and WüNLP at the DIALECT-COPA Shared Task: In-Context Learning From Just a Few Dialectal Examples Gets You Quite Far

Nikola Ljubešić[1,2], **Taja Kuzman**[1], **Peter Rupnik**[1],
Ivan Vulić[3], **Fabian David Schmidt**[4], and **Goran Glavaš**[4]
[1]Dept. of Knowledge Technologies, Jožef Stefan Institute
[2]University of Ljubljana
[3]Language Technology Lab, University of Cambridge
[4]Center for AI and Data Science (CAIDAS), University of Würzburg
{nikola.ljubesic, taja.kuzman, peter.rupnik}@ijs.si
{fabian.schmidt, goran.glavas}@uni-wuerzburg.de, iv250@cam.ac.uk

Abstract

The paper presents the JSI and WüNLP systems submitted to the DIALECT-COPA shared task on causal commonsense reasoning in dialectal texts. Jointly, we compare LLM-based zero-shot and few-shot in-context inference (JSI team), and task-specific few-shot fine-tuning, in English and respective standard language, with zero-shot cross-lingual transfer (ZS-XLT) to the test dialects (WüNLP team). Given the very strong zero-shot and especially few-shot in-context learning (ICL) performance, we further investigate whether task semantics, or language/dialect semantics explain the strong performance, showing that a significant part of the improvement indeed stems from learning the language or dialect semantics from the in-context examples, with only a minor contribution from understanding the nature of the task. The higher importance of the dialect semantics to the task semantics is further shown by the finding that the in-context learning with only a few dialectal instances achieves comparable results to the supervised fine-tuning approach on hundreds of instances in standard language.

1 Introduction

Causal commonsense reasoning is an important aspect of natural language understanding (NLU) abilities of the large language models (LLMs); their performance on such tasks probes the extent to which the LLMs have acquired commonsense and world knowledge. Choice Of Plausible Alternatives (COPA) dataset (Roemmele et al., 2011) has de facto been the standard evaluation benchmark for causal commonsense reasoning for over a decade.[1] Like on most other NLU tasks, state-of-the-art LLMs exhibit impressive performance on the English COPA dataset (Chowdhery et al., 2023; Zhong et al., 2022). LLMs, unlike their smaller encoder-based predecessors (e.g., BERT, RoBERTa), also offer spectacular COPA performance for other languages (Ponti et al., 2020; Žagar and Robnik-Šikonja, 2022; Shi et al., 2023), including South Slavic languages, both with Latin and Cyrillic scripts, reaching accuracy levels between 94% and 97%[2]. Though LLMs excel on high-resource and moderately resourced standard languages, their utility for commonsense reasoning in truly low-resource languages (Senel et al., 2024) and especially dialects (Joshi et al., 2024) has been much less scrutinized. In the DIALECT-COPA shared task of the VarDial Evaluation Campaign 2024 (Chifu et al., 2024), COPA is extended to geographically very localized dialects (i.e., micro- or nano-dialects) of South Slavic languages that are very rarely present in texts online, and thus could not have been (except perhaps in minimal traces) present in the pretraining corpora of LLMs.

In this work, we focus on benchmarking decoder-style LLMs in the DIALECT-COPA task, covering a variety of closed-source and open-source LLMs in zero-shot and few-shot in-context learning (ICL) inference setups. Subsequently, we select the best-performing open-source model during in-context learning (Mixtral Instruct) and fine-tune it for the task in the standard supervised fashion – assuming a somewhat larger training dataset – with training instances either in English or in the respective standard language of the target dialect (e.g., Slovenian for the Cerkno dialect).

[1] Inter alia, the COPA dataset is included to the selection of tasks in the well-known benchmark for general-purpose natural language understanding SuperGLUE (Wang et al., 2019).

[2] https://github.com/clarinsi/benchich/tree/main/copa

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 209–219
June 20, 2024 ©2024 Association for Computational Linguistics

We make use of all the development data provided inside the shared task, namely the translations of the COPA dataset (Roemmele et al., 2011) into the standard Slovenian, Croatian, Serbian and Macedonian languages, as well as the translations available for two out of three dialects, namely the Cerkno and the Torlak dialects (Ljubešić et al., 2024). While we have access to both training and development portions of COPA datasets for other languages and dialects, the Chakavian dialect is a surprise dialect: findings from the other two dialects thus steered decisions for Chakavian too.

To sum up, we evaluate the LLMs in the following scenarios: **1)** *zero-shot inference* where the model is presented with the task description in English and needs to provide an answer to the COPA instances in the South Slavic dialects; **2)** *few-shot in-context learning* (ICL) where the prompt is extended with additional examples from the respective COPA dataset; and **3)** *fine-tuning zero-shot cross-lingual transfer* (ZS-XLT) (Lauscher et al., 2020; Schmidt et al., 2022), in which an LLM is (in a parameter-efficient manner) fine-tuned on training data in English or a standard South Slavic language (Slovenian, Serbian, and Croatian, respectively) and then used to make predictions in the corresponding target dialect (Cerkno, Torlak, and Chakavian, respectively).

The ICL variants in general, with few target dialect instances in the context, exhibit a significantly improved performance in comparison to zero-shot performance. Comparing ICL to fune-tuning zero-shot cross-lingual transfer, we observe a comparable performance.

Following the finding of significant improvements through just a few target dialect examples, we investigate the source of these few-shot ICL performance gains. We find that the exposure to the dialect itself through the few in-context instances is key, as opposed to exposure to the COPA task itself.

2 Multi-Parallel COPA Datasets

Our work focuses on the Choice Of Plausible Alternatives (COPA) dataset, originally published in English (Roemmele et al., 2011), and its translation-based derivatives in a selection of South Slavic languages and dialects. All COPA datasets have the same set of instances, and they differ only in the language variety in which the instances are written. The COPA dataset consists of 1,000 examples, split into 400 training, 100 development and 500 test instances. Each instance consists of three sentences: a statement (*premise*) and two possible *effects* or *causes* (alternatives) for the statement, e.g., a premise *All my socks were in the laundry* is coupled with two *effect* choices: *I wore sandals* (correct/plausible) and *I wore boots* (incorrect).

We evaluate the models on 'standard language' and dialectal versions of a selection of South Slavic languages. More precisely, we use the following COPA datasets for three South Slavic dialects – the Slovenian Cerkno dialect (COPA-SL-CER), the Croatian Chakavian dialect (COPA-HR-CKM), and the Torlak dialect of Serbian (COPA-SR-TOR) (Ljubešić et al., 2024). The models' performance on the dialectal datasets is compared with their performance on the datasets in the standard South Slavic language that is closest to them, namely Slovenian (COPA-SL) (Žagar et al., 2020), Croatian (COPA-HR) (Ljubešić, 2021), Serbian (COPA-SR) (Ljubešić et al., 2022b) and Macedonian (COPA-MK) (Ljubešić et al., 2022a). All the datasets were translated from the English COPA dataset (Roemmele et al., 2011) following the XCOPA translation and adaptation methodology (Ponti et al., 2020), except for Slovenian which was translated as part of the Slovenian SuperGLUE benchmark (Žagar and Robnik-Šikonja, 2022). Torlak, Serbian and Macedonian datasets are written in Cyrillic and all other in Latin script.

The COPA datasets for the standard South Slavic languages and English are openly available, whereas the dialectal COPA datasets have been introduced in the DIALECT-COPA shared task, part of the VarDial Evaluation Campaign 2024 (Chifu et al., 2024) and are currently only partly available: as part of the shared task, the training and development portions were made publicly available for all languages (Ljubešić et al., 2024);[3] the test splits have been made available only to the shared task participants. Inside the shared task, no training and development data were given for the Chakavian dialect, to enable estimation and analysis of models' performance "in the wild" for a new (truly low-resource) dialect.

3 Models in Evaluation

In this work, we extend the prior experiments that focused on the use of LLMs for the task (Wi-

[3] The training and development splits can be accessed at the CLARIN.SI repository: `http://hdl.handle.net/11356/1766`.

bowo et al., 2023; Ljubešić et al., 2024) by (i) evaluating a larger number of open- and closed-source instruction-tuned generative LLMs, and by (ii) widening investigation from the basic zero-shot scenarios to few-shot in-context learning and cross-lingual transfer of supervised fine-tuning. In this section, we outline all the models, with links to the models provided in Appendix A.

GPT-3.5 Turbo and GPT-4 are closed-source models provided by OpenAI through their payable API (OpenAI, 2023a,b). We use the versions `gpt-3.5-turbo-0125` and `gpt-4-0125-preview` through the chat completion endpoint, with temperature set to 0. The models are said to be trained on massive multilingual web text collections; however, the details on pretraining data, as well as the details of the training procedure and model architecture are not publicly known.

Mistral 7B Instruct is an open-source model provided by Mistral AI (Jiang et al., 2023). We experiment with two 7B model variants, `Mistral-7B-Instruct-v0.1` and `Mistral-7B-Instruct-v0.2`, where the main difference is that v0.2 extends the context size from 8k to 32k input tokens. The details on the pretraining data have not been made available.

Mixtral 8×7B Instruct is another open-source model from Mistral AI (Jiang et al., 2024). We use the `Mixtral-8x7B-Instruct-v0.1` variant. The main difference between Mistral and Mixtral is the introduction of a sparse mixture-of-experts network in Mixtral, where 8 feed-forward blocks are added to each layer. For each token, two blocks are selected to process it. As a consequence, despite having 47B parameters in total, only 13B active parameters are used for each token. Furthermore, it is stated that Mixtral was pretrained on much larger quantities of multilingual data than Mistral. The context size is 32K tokens.

mT0-XXL is an open-source model developed by the BigScience academic initiative (Muennighoff et al., 2023). We use the `mT0-XXL` variant which has 13 billion parameters. The model is a fine-tuned version of the multilingual mT5 model (Xue et al., 2021), which was pretrained on a sample from the mC4 dataset covering 101 languages.

Aya 101 is an open-source model developed by Cohere For AI (Üstün et al., 2024). We use the `aya-101` variant with 13B parameters. As mT0 above, it is an instruction-tuned version of mT5 (Xue et al., 2021), relying on a multilingual dataset that covers 101 languages.

Gemma 7B It is an open-source model provided by Google (Mesnard et al., 2024). It is a lightweight 7B version of Google's closed-source Gemini model family (Anil et al., 2023), and it was trained primarily on English data.

Falcon-7B-Instruct is an open-source 7B model developed by the Technology Innovation Institute (Almazrouei et al., 2023). It is an instruction-tuned version of the Falcon-7B language model which was pretrained on English and French data.

Llama-2-7B-Chat is a 7B open-source model from Meta (Touvron et al., 2023), with the context size of 4000 tokens, intended primarily for English.

In sum, the coverage of evaluated models is extensive, where the models vary in their availability (open-sourced versus 'black-box' commercial models), size, as well as their pretraining data. For instance, while mT0 and Aya 101 were pretrained on massively multilingual datasets, other models are primarily built for English only, such as Gemma and Llama-2-Chat. Further, while most models have 7B parameters, Mixtral 8×7B Instruct, mT0 and Aya 101 have 13B parameters.

To maximize the comparability between the results of the models, we provide them all with identical prompts (available in Appendix B). We ran all our experiments on a single A100 40GB.[4]

4 Results and Discussion

We now delve into the main experiments, covering zero-shot and different 10-shot ICL scenarios, followed by ablations on the importance of learning 'language/dialect semantics' versus 'task semantics/structure' in ICL. Finally, we report experiments with supervised fine-tuning.[5]

4.1 Zero-Shot Inference

Table 1 summarizes the results of zero-shot inference with LLMs on the training portions of the datasets (400 examples), with models listed in decreasing order of performance on standard language datasets (column STD), that is, Slovenian (sl), Croatian (hr), Serbian (sr) and Macedonian

[4] Due to this, we relied on an 8-bit quantization for Mistral models and a 4-bit quantization for Mixtral models.

[5] While the data for Serbian, Macedonian and Torlak are available both in the Latin and in the Cyrillic script, we report only the results on Serbian and Torlak Latin data and Macedonian Cyrillic data; these options yielded higher absolute scores across the models.

(mk). The ranking of the models based on the dialectal performance (column DIA), i.e., on Cerkno (sl-cer) and Torlak (sr-tor), is similar.

While the model ranking is relatively similar relative on both standard and dialect varieties, all models expectedly perform substantially worse on dialectal datasets. For instance, the best-performing system, GPT-4, drops 36.5 accuracy points (from 96 to 59.5) between Slovenian (sl) and its Cerkno dialect (sl-cer), and from 95.8 to 76 on average. Such drops are observed for all the other models as well (e.g., mT0 as the best-performing open-source model has 14 points lower average accuracy on DIA compared to STD).

Overall, GPT-4 outperforms the open-model competition by a wide margin, with mT0 as the closest follower (10 accuracy points difference on DIA). Expectedly, Mixtral performs much better than its smaller Mistral 7B Instruct counterparts. Two systems that perform worse than expected are Aya 101, which closely follows the design of an earlier mT0 model, and Gemma 7B It. Finally, Falcon-7B-Instruct and Llama-2-7B-Chat perform worse than the random baseline of 50% due to their inability to follow instructions, frequently providing answers in which neither of the two alternatives is chosen. This might stem from their limited multilingual capabilities, as outlined in Section 3.

Other Observations. It is worth noting that models generally tend to exhibit similar performance across the standard language variety: there are no large or consistent differences in performance on Slovenian, Croatian, Serbian and Macedonian, despite the fact that these languages are not equally resourced (e.g., Slovenian is by far the most resourced of the four, whereas Macedonian is the least resourced (Terčon and Ljubešić, 2023)).

In contrast, models' performance across the two dialects is vastly different. The Cerkno dialect seems to be much more challenging for all models than the Torlak dialect. This, we believe, stems from the fact that Torlak is significantly closer to the standard Serbian and Macedonian than Cerkno is to standard Slovenian (Ljubešić et al., 2024).

4.2 Few-Shot In-Context Learning

We next perform in-context learning (ICL) only over the models that performed above the random baseline in the zero-shot evaluation. First, we note that mT0 and Aya 101, both based on mT5, actually experienced performance decrease when moving from zero-shot to few-shot ICL scenarios. We speculate that this might be a consequence of limited context size and encoder capacity, which might be incapable of encoding a longer prompt. We thus present only the results where models show gains moving from zero-shot to ICL scenarios.

In our preliminary experiments, we varied the number of few-shot examples from the development set provided to the models. The results show consistent improvements as the number of shots increases up to 10, followed by minor and negligible gains with 20 instead of 10 shots. For that reason, we report the results in the 10-shot scenario. An example of a prompt is provided in Appendix B. An overview of results with zero-shot (Section 4.1) versus 10-shot prompting scenarios is provided in Table 2.

The main finding is that ICL, for the models with sufficient context sizes where ICL works as expected, offers substantial performance benefits both for the standard languages (column STD) and for the target dialects (column DIA). Interestingly, the largest absolute gains from 10-shot ICL are observed for the most difficult, Cerkno dialect: performance of GPT-4 rises from 60% to 74% in accuracy.

The observed gains with ICL thus open up the following question – where do the gains come from? Is it the adaptation to the task and its structure, or is it rather the adaptation to the target language and dialect and a better understanding of it? We discuss this next in the prompt ablation tests.

Prompt Ablation Experiments. We aim to discriminate between the contributions of learning the 'task semantics' versus learning the 'language/dialect semantics' by performing two experiments: **1)** in the `list` experiment we add to the initial zero-shot prompt only lists of sentences of the target language, and **2)** in the `task` experiment the structure of the task is added to the initial prompt by providing instances from the COPA dataset but without any answer. As before, we use the development dataset instances for few-shot prompts. With `list` we ablate the task definition, while with `task` we ablate the information on the answer, but still provide information on task itself.

The results are given in Table 3. The main finding is that the substantial part of the total improvement comes from the language/dialect semantics, represented by the `list` results. An answer to the task, missing in the `task` scenario, but included in

Model	STD	DIA	sl	sl-cer	hr	sr	sr-tor	mk
gpt-4-0125-preview	0.958	0.760	0.960	0.595	0.960	0.968	0.925	0.943
mT0-xxl	0.798	0.660	0.787	0.540	0.738	0.765	0.713	0.838
gpt-3.5-turbo-0125	0.799	0.646	0.802	0.547	0.820	0.830	0.745	0.745
aya-101	0.710	0.610	0.728	0.530	0.645	0.665	0.623	0.720
Mixtral-8x7B-Instruct-v0.1	0.691	0.521	0.682	0.405	0.705	0.713	0.637	0.665
gemma-7b-it	0.599	0.546	0.593	0.522	0.570	0.618	0.552	0.605
Mistral-7B-Instruct-v0.2	0.524	0.396	0.515	0.285	0.542	0.537	0.487	0.497
Mistral-7B-Instruct-v0.1	0.510	0.495	0.507	0.487	0.507	0.515	0.500	0.502
falcon-7b-instruct	0.432	0.442	0.500	0.485	0.463	0.458	0.510	0.357
llama-2-7b-chat	0.114	0.032	0.175	0.020	0.152	0.145	0.090	0.035

Table 1: Zero-shot results, with additional averages reported over the three standard languages (STD) and the three dialects (DIA). Results are reported in accuracy scores.

Model	# shots	STD	DIA	sl	sl-cer	hr	sr	sr-tor	mk
Mistral-7B-Instruct-v0.2	0	0.524	0.396	0.515	0.285	0.542	0.537	0.487	0.497
Mistral-7B-Instruct-v0.2	10	0.734	0.570	0.718	0.507	0.757	0.752	0.632	0.708
Mixtral-8x7B-Instruct-v0.1	0	0.691	0.521	0.682	0.405	0.705	0.713	0.637	0.665
Mixtral-8x7B-Instruct-v0.1	10	0.780	0.624	0.802	0.5	0.818	0.795	0.748	0.703
gpt-3.5-turbo-0125	0	0.799	0.646	0.802	0.547	0.820	0.830	0.745	0.745
gpt-3.5-turbo-0125	10	0.828	0.666	0.845	0.53	0.84	0.858	0.802	0.77
gpt-4-0125-preview	0	0.958	0.760	0.960	0.595	0.960	0.968	0.925	0.943
gpt-4-0125-preview	10	0.984	0.853	0.98	0.738	0.988	0.99	0.968	0.978

Table 2: Zero- and ten-shot results in terms of accuracy across models that improve with few-shot prompting. Averages for datasets in standard languages (STD), i.e., Slovenian, Croatian, Serbian and Macedonian, and dialectal datasets (DIA), i.e., Cerkno and Torlak, are given.

Variant	STD	DIA	sl	sl-cer	hr	sr	sr-tor	mk
zero-shot	0.691	0.521	0.682	0.405	0.705	0.713	0.637	0.665
10-shot	0.780	0.624	0.802	0.5	0.818	0.795	0.748	0.703
list	0.745	0.607	0.74	0.515	0.775	0.757	0.698	0.708
task	0.786	0.619	0.818	0.492	0.805	0.802	0.745	0.72

Table 3: Results over the ablated 10-shot examples on the Mixtral 8×7B Instruct model, either to the level of a list of sentences (list) or tasks without any answer given (task), compared to the previous results of zero-shot and 10-shot experiments. We additionally provide averages over standard languages (STD) and dialects (DIA).

the 10-shot scenario, seems to be almost irrelevant for ICL. However, the remaining gap between the list and the task rows in Table 3 indicates that providing examples of the task, although without the answer, is still beneficial.

These results shed important light on *why* in-context learning offers substantial gains both on standard languages and on dialects. However, there is another angle, specific to this shared task, that these results open up. Namely, both the list- and the task- transformed prompts do not require the correct answer to be known as part of in-context examples; they can therefore be run even on the Chakavian dialect, for which no training and development data were available in the shared task. Interestingly, omitting an answer even yields minor gains on the datasets in standard languages, and just a minor drop in performance on the dialectal datasets.

4.3 Fine-Tuning and ZS-XLT

The WüNLP team next investigates zero-shot cross-lingual transfer (ZS-XLT) with an LLM fine-tuned on English training data or the training data in the corresponding standard language (e.g., for Chakavian as target, we train on the instances from the training portion of Croatian COPA). Following the JSI team's zero-shot inference and few-shot ICL results, we opt to tune Mixtral 8×7B Instruct as the best-performing open-source LLM in their ICL experiments. We fine-tune the model generatively, using the prompt below, and constraining the output vocabulary to "1", "2" (we minimize the standard negative log likelihood loss):

```
'Premise: "{PREMISE}"
 Question: "{QUESTION}"
 Choice 1: "{CHOICE1}"
 Choice 2: "{CHOICE2}"
 Answer: '
```

Since we are running supervised fine-tuning, we chose to prepend the task description to the prompt.[6] We carry out fine-tuning in a parameter-efficient manner, using quantized (4-bit) low-rank adaptation (Q-LoRA) (Hu et al., 2021; Dettmers et al., 2024), optimizing the LoRA matrices with AdamW (Loshchilov and Hutter, 2018) (learning rate 10^{-5} with linear decay, no warmup). We train on the whole training set (400 instances) in batches of 32 instances, for 10 epochs, checkpointing the model after every update.

Although the DIALECT-COPA shared task offers validation portions in target languages/dialects, one should note that, following Schmidt et al. (2022, 2023b), using target language development set for model selection violates true zero-shot cross-lingual transfer: the labeled target language validation instances would, in fact, be better used as training data (Schmidt et al., 2023b). Because of this, we report results for two model variants: (1) training on English instances (en) vs. instances of the corresponding standard language (x) × (2) selecting the last checkpoint (last) of the training run (true ZS-XLT) vs. selecting the model checkpoint that has the best performance on the target language validation set (val, violates true ZS-XLT). These four variants are named as: MixtralLoRA-{en,x}-{last,val}. Table 4 summarizes the performance for all four variants on the validation data of standard South Slavic languages as well as target dialects. The final official shared task results for all four variants (runs), on the test portions of target dialects, are reported in the next section.

4.4 Results on Test Data

We present the official test data results of both teams in Figure 1. The runs from WüNLP comprise fine-tuning Mixtral 8×7B Instruct either on the English or the standard data across two model selection scenarios, as described in Section 4.3. Similar to the results during the development phase (Table 4), there is no strong difference between the variants: the averages are almost identical. However, comparing this set of results to the zero-shot approach with Mixtral 8×7B Instruct, we observe positive impact of fine-tuning, even if fine-tuning was conducted on English or standard language data.

The list and the task approaches in the 10-shot ICL scenario with Mixtral 8×7B Instruct, conducted by JSI, improve over the zero-shot scenario, arriving roughly to the level of the WüNLP fine-tuning results.

The best results of the two teams, as in the shared task overall, are obtained, not surprisingly, with the GPT-4-based take on zero-shot inference, and even more on the two approaches to 10-shot ICL without having the correct answers at hand. While zero-shot prompting already improves over any of Mixtral results on each of the three dialects, achieving an average result of 75% accuracy, the model excels further once 10 examples of the language are provided for ICL, even only as examples of the dialect in question, with the average result rising to 83%. Describing the nature of the task combined with the 10 shots, but without the correct answer, yields an additional gain, resulting in an average accuracy of 87%.

Interestingly, the 'harder' the dialect, the more is gained by just submitting exemplary sentences of the dialect during in-context learning, with a much more significant jump from zero-shot scenario (gpt4-zero) to the scenario with a list of sentences in the dialect added (gpt4-list) on the Cerkno dialect (considered a 'hard dialect') than on the Torlak dialect (considered an 'easy dialect'). We see similar further gains moving from the scenario with the list of sentences in the dialect (gpt4-list) to the scenario where examples of the task are added (gpt4-task).

5 Conclusion

In this work, we benchmark three mainstream approaches for using LLMs for causal commonsense reasoning in three South Slavic dialects: (1) zero-shot inference with LLMs, (2) few-shot in-context learning, and (3) supervised fine-tuning and zero-shot cross-lingual transfer. We find that, for the same LLM, both few-shot ICL and cross-lingual transfer with supervised fine-tuning (with training instances in English or in the standard language of the target dialect) expectedly outperform zero-shot inference with LLMs. Somewhat surprisingly, few-shot ICL with as few as 10 in-dialect instances tends to perform comparably to fine-tuning based

[6] Recent work indeed suggests that, unlike in zero-shot inference and few-shot ICL, task description prompts have limited effect on performance in supervised fine-tuning (Li et al., 2023).

Variant	STD	DIA	sl	sl-cer	hr	sr	sr-tor	mk
MixtralLoRA-en-last	0.815	0.615	0.82	0.52	0.82	0.87	0.71	0.75
MixtralLoRA-en-val	0.82	0.645	0.82	0.57	0.82	0.87	0.72	0.77
MixtralLoRA-x-last	0.825	0.675	0.80	0.57	0.83	0.89	0.78	0.78
MixtralLoRA-x-val	0.833	0.69	0.82	0.60	0.84	0.89	0.78	0.78

Table 4: Fine-tuning zero-shot cross-lingual transfer results (ZS-XLT) on the *validation* data: fine-tuning Mixtral-Instruct 8x7B with Q-LoRA, either on English training data (en) or the training portion of the standard language corresponding to the target dialect (x); For each of the two models (*en* vs. *x*) we report the performance of the `last` checkpoint as well as the checkpoint that yields the best validation performance. We additionally provide averages over standard datasets (STD) and dialectal datasets (DIA).

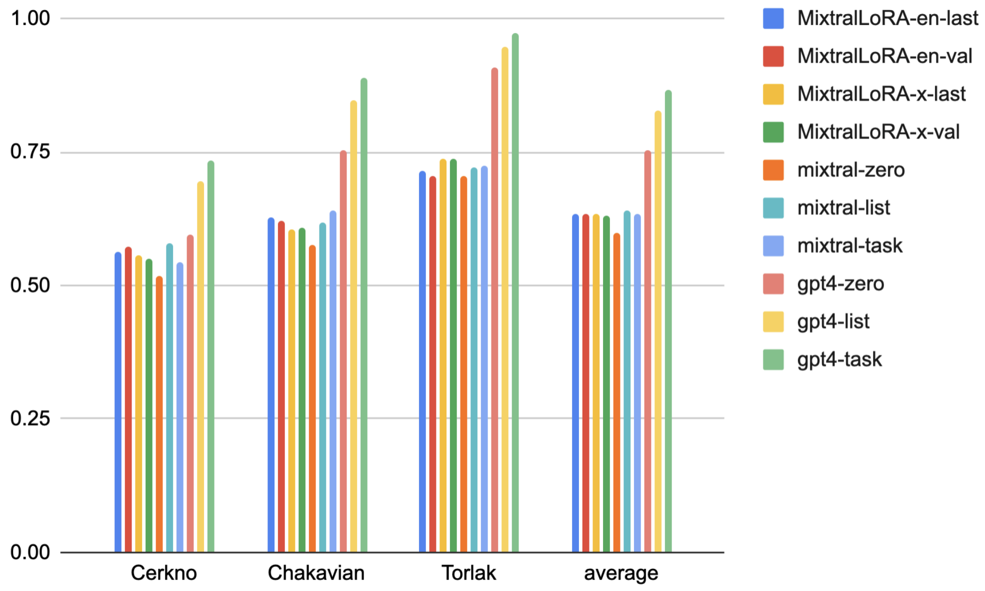

Figure 1: Test data results

on 400 instances in standard languages that are related to the corresponding dialect. Further inspection reveals that the LLMs leverage the few provided in-dialect instances to improve their understanding of the target dialect, rather than to learn the task and its structure. Future work will investigate further recent strategies for improving performance of LLMs for low-resource languages and in cross-lingual transfer, including, *inter alia*, checkpoint averaging in fine-tuning (Schmidt et al., 2023a) and supervised in-context learning (Li et al., 2023).

Acknowledgements

The research presented in this paper was conducted within the research project titled Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language (J7-4642) funded by the Slovenian Research and Innovation Agency, and withing the research programme “Language resources and technologies for Slovene” (P6-0411), both funded by the Slovenian Research and Innovation Agency (ARIS).

6 Limitations

One of the limitations of the presented paper is the use of closed-source models. While we decided to include them in the analyses to be able to obtain an insight into how well the open-source models perform in comparison to the closed-source models, we should note that we have limited insights to the architecture of these models and that the reproducibility of these results might be hindered by updates to the models that might not be communicated openly.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Adrian Chifu, Goran Glavaš, Radu Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *arXiv preprint arXiv:2401.05632*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Chengzu Li, Han Zhou, Goran Glavaš, Anna Korhonen, and Ivan Vulić. 2023. On task performance and model calibration with supervised and self-ensembled in-context learning. *arXiv preprint arXiv:2312.13772*.

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.

Nikola Ljubešić, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024. DIALECT-COPA: Extending the standard translations of the COPA causal commonsense reasoning dataset to south slavic dialects. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, and et al. 2024. Gemma.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

OpenAI. 2023a. ChatGPT General FAQ. `https://help.openai.com/en/articles/6783457-chatgpt-general-faq`. Accessed: March 3, 2023.

OpenAI. 2023b. Gpt-4 technical report.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

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*.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2023a. Free lunch: Robust cross-lingual transfer via model checkpoint averaging. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5712–5730, Toronto, Canada. Association for Computational Linguistics.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2023b. One for all & all for one: Bypassing hyperparameter tuning with model averaging for cross-lingual transfer. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with big brother's help – a benchmark and evaluation for Turkic languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688, St. Julian's, Malta. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Luka Terčon and Nikola Ljubešić. 2023. CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2023. COPAL-ID: Indonesian language reasoning with local culture and nuances. *arXiv preprint arXiv:2311.01012*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Aleš Žagar and Marko Robnik-Šikonja. 2022. Slovene SuperGLUE Benchmark: Translation and Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2058–2065.

Aleš Žagar, Marko Robnik-Šikonja, Teja Goli, and Špela Arhar Holdt. 2020. Slovene translation of SuperGLUE. Slovenian language resource repository CLARIN.SI.

Qihuang Zhong, Liang Ding, Yibing Zhan, Yu Qiao, Yonggang Wen, Li Shen, Juhua Liu, Baosheng Yu, Bo Du, Yixin Chen, et al. 2022. Toward efficient language model pretraining and downstream adaptation via self-evolution: A case study on superglue. *arXiv preprint arXiv:2212.01853*.

A Overview of Models

The models in evaluation (see Section 3) along with the links to access them are available in Table 5.

B Overview of Prompts

Zero-shot prompt An example from the Slovenian Cerkno dataset.

You will be given a task. The task definition is in English, but the task itself is in another language. Here is the task!

Given the premise "Muoje telu je metalu sinca na traua.", and that we are looking for the cause of this premise, which hypothesis is more plausible?

Hypothesis 1: "Sunce je šlu guor.".

Hypothesis 2: "Traua je bla pakuošena.".

Answer only with "1" or "2".

Answer:

Ten-shot prompt An example from the Croatian Chakavian dataset.

You will be given a task. The task definition is in English, but the task itself is in another language. You are to choose the more likely hypothesis given a premise. Take into account that we are either looking for a cause or an effect of the premise. Answer only with "1" or "2". Here are some examples of the task:

Model	Link
gpt-3.5-turbo-0125	–
gpt-4-0125-preview	–
Mistral-7B-Instruct-v0.1	https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1
Mistral-7B-Instruct-v0.2	https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
Mixtral-8x7B-Instruct-v0.1	https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
mT0-xxl	https://huggingface.co/bigscience/mt0-xxl
aya-101	https://huggingface.co/CohereForAI/aya-101
gemma-7b-it	https://huggingface.co/google/gemma-7b-it
falcon-7b-instruct	https://huggingface.co/tiiuae/falcon-7b-instruct
llama-2-7b-chat	https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

Table 5: Models in evaluation along with their huggingface.co links.

Example 1:

Premise: "Muški je otpra špino."

Question: "effect"

Hypothesis 1: "Školjka ot zahoda se je napunila z oduon."

Hypothesis 2: "Oda je počela teć z mlaznici."

Answer: "2"

Example 2:

Premise: "Mlada je našla neko blago va žitaricah."

Question: "effect"

Hypothesis 1: "Nalila je mlieko va škudelico."

Hypothesis 2: "Je zgubila tiek."

Answer: "2"

Example 3:

...

Example 10:

Premise: "Šlovek je čuda popi na fešte."

Question: "effect"

Hypothesis 1: "Ta drugi dan ga je bolela glava."

Hypothesis 2: "Ta drugi dan mu je kapa nuos."

Answer: "1"

Now to your task!

Premise: "Moje tielo je hitalo hlat na travo."

Question: "cause"

Hypothesis 1: "Sunce je hodilo van."

Hypothesis 2: "Trava je bila pokošena."

Answer:

List prompt The ten-shot prompt, but omitting the structure of the task in the examples, and rather giving just samples of the language the task will be in.

You will be given a task. The task definition is in English, but the task itself is in another language. Here are some samples of the language the task is in:

Sample 1:

"Muški je otpra špino."

"Školjka ot zahoda se je napunila z oduon."

"Oda je počela teć z mlaznici."

Sample 2:

"Mlada je našla neko blago va žitaricah."

"Nalila je mlieko va škudelico."

"Je zgubila tiek."

Sample 3:

...

Sample 10:

"Šlovek je čuda popi na fešte."

"Ta drugi dan ga je bolela glava."

"Ta drugi dan mu je kapa nuos."

Now to your task! You are to choose the more likely hypothesis given a premise. Take into account that we are either looking for a cause or an effect of the premise. Answer only with "1" or "2".

Premise: "Moje tielo je hitalo hlat na travo."

Question: "cause"

Hypothesis 1: "Sunce je hodilo van."

Hypothesis 2: "Trava je bila pokošena."

Answer:

Task prompt The ten-shot prompt, but without an answer provided. An example from the Croatian Chakavian dataset.

You will be given a task. The task definition is in English, but the task itself is in another language. You are to choose the more likely hypothesis given a premise. Take into account that we are either looking for a cause or an effect of the premise. Answer only with "1" or "2". Here are some examples of the task without a solution:

Example 1:

Premise: "Muški je otpra špino."

Question: "effect"

Hypothesis 1: "Školjka ot zahoda se je napunila z oduon."

Hypothesis 2: "Oda je počela teć z mlaznici."

Example 2:

Premise: "Mlada je našla neko blago va žitaricah."

Question: "effect"
Hypothesis 1: "Nalila je mlieko va škudelico."
Hypothesis 2: "Je zgubila tiek."
Example 3:
...
Example 10:
Premise: "Šlovek je čuda popi na fešte."
Question: "effect"
Hypothesis 1: "Ta drugi dan ga je bolela glava."
Hypothesis 2: "Ta drugi dan mu je kapa nuos."
Now to your task!
Premise: "Moje tielo je hitalo hlat na travo."
Question: "cause"
Hypothesis 1: "Sunce je hodilo van."
Hypothesis 2: "Trava je bila pokošena."
Answer: