Incorporating Dialect Understanding Into LLM Using RAG and Prompt Engineering Techniques for Causal Commonsense Reasoning

Benedikt Perak
University of Rijeka,
Faculty of Humanities
and Social Sciences,
AIRI Center
`bperak@uniri.hr`

Slobodan Beliga
University of Rijeka,
Faculty of Informatics and
Digital Technologies,
AIRI Center
`sbeliga@uniri.hr`

Ana Meštrović
University of Rijeka,
Faculty of Informatics and
Digital Technologies,
AIRI Center
`amestrovic@uniri.hr`

Abstract

The choice of plausible alternatives (COPA) task requires selecting the most plausible outcome from two choices based on understanding the causal relationships presented in a given text. This paper outlines several approaches and model adaptation strategies to the VarDial 2024 DIALECT-COPA shared task, focusing on causal commonsense reasoning in South-Slavic dialects. We utilize and evaluate the GPT-4 model in combination with various prompts engineering and the Retrieval-Augmented Generation (RAG) technique. Initially, we test and compare the performance of GPT-4 with simple and advanced prompts on the COPA task across three dialects: Cerkno, Chakavian and Torlak. Next, we enhance prompts using the RAG technique specifically for the Chakavian and Cerkno dialect. This involves creating an extended Chakavian-English and Cerkno-Slovene lexical dictionary and integrating it into the prompts. Our findings indicate that the most complex approach, which combines an advanced prompt with an injected dictionary, yields the highest performance on the DIALECT-COPA task.

1 Introduction

The choice of plausible alternatives task, introduced by Roemmele et al. in 2011 (Roemmele et al., 2011), presents a scenario where the model is required to comprehend cause-effect relationships from provided input and select the most plausible outcome from two choices. This task encapsulates various linguistic aspects, including lexical semantics, syntactic structure, and world knowledge, making it a challenging assessment for NLP systems (Ponti et al., 2020). It is an important benchmark in the domain of NLP, specifically designed to evaluate models' abilities in causal reasoning and inference making.

In recent years, the advancement of neural network architectures has led to the creation of a wide range of neural text representation models (Babić et al., 2020). Particularly transformer-based models like BERT (Devlin et al., 2018) and GPT (Radford et al., 2019), has significantly impacted the landscape of NLP tasks. Models pre-trained on extensive text corpora are commonly referred to as Large Language Models (LLMs). Through this process, they capture linguistic patterns and semantic connections, enabling them to perform better than traditional language models in a range of NLP tasks. Thus, LLMs have demonstrated remarkable performance across tasks such as text classification (Sun et al., 2019; Balkus and Yan, 2022), sentiment analysis (Babić et al., 2021; Beliga et al., 2021), paraphrase detection (Vrbanec and Meštrović, 2023), information spreading prediction (Meštrović et al., 2022), machine translation (Zhu et al., 2020; Yang et al., 2020), metaphor generation (Tong et al., 2024), question answering (Wang et al., 2019), etc.

Evaluating the performance of LLMs in tasks like COPA remains important to assess their true capabilities in causal reasoning inference. Furthermore, while LLMs exhibit strong performance in languages with abundant resources, such as English, their effectiveness decreases for small(er) dialects of non-major and low-resource languages (Kantharuban et al., 2023). Thus, it is important to explore novel approaches and model adaptation strategies that may enhance LLMs' abilities to comprehend dialects.

This paper aims to study language understanding in micro-dialects for moderately-resourced South-Slavic languages. Precisely, we experiment with the COPA task in Cerkno dialect (Slovenia), Chakavian dialect (Croatia) and Torlak dialect (Serbia). In addition, we evaluated some of our approaches to COPA task on datasets in several languages: English, Slovenian, Croatian, Serbian, and Macedonian language.

In this work, our focus is on examining and dif-

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 220–229
June 20, 2024 ©2024 Association for Computational Linguistics

ferentiating various methods specifically designed for the DIALECT-COPA task, with a particular emphasis on causal commonsense reasoning. The essence of this task lies in its ability to test models on understanding causal relationships within given scenarios, a crucial component of commonsense reasoning. For instance, consider the premise: "The girl found a bug in her cereal." In this case, the DIALECT-COPA task presents two possible effects: 1) "She poured milk in the bowl," and 2) "She lost her appetite." The challenge for the model is to deduce the most plausible effect of the initial event. While the first option is a neutral action that could occur in any context of preparing cereal, the second option directly relates to the discovery of a bug in the cereal, which would naturally lead to a loss of appetite. This task not only assesses the model's ability to infer logical consequences from specific events but also its capacity to navigate and understand nuanced human reactions, thereby evaluating its grasp of causal commonsense reasoning within varied contexts.

We consider and examine four distinct model adaptation strategies of utilising GPT-4 in the DIALECT-COPA task as follows.

1. Simple Prompt Engineering for GPT-4: This initial method involves straightforward simple prompt engineering with GPT-4, employing an iterative, 0-shot framework. It serves as our baseline, testing the model's innate ability to understand and reason about the given dialectal inputs without prior examples.

2. Advanced Prompt Engineering for GPT-4: Building on the first approach, we introduce advanced prompt engineering techniques that incorporate explicit reasoning instructions. This method enhances the model's capacity for logical deduction and causal inference, aimed at improving its performance on the task.

3. Simple Prompt Engineering with Retrieval-Augmented Generation (RAG): The third strategy extends the simple prompt engineering approach by integrating the RAG technique. This implementation includes an expanded lexical database, featuring 11,000 Chakavian and 4,000 Cerkno lexical entries, to facilitate deeper understanding and generation capabilities in these dialects.

4. Advanced Prompt Engineering with RAG for GPT-4: Our most sophisticated approach combines advanced prompt engineering, reasoning instructions, and RAG (see Fig. 1). This comprehensive method leverages the expanded dictionaries — comprising 11,000 Chakavian and 4,000 Cerkno lexical items — and integrates reasoning instructions to optimize the model's performance on the DIALECT-COPA task by enhancing its reasoning capabilities and dialect understanding.

Augmenting GPT models with dialect dictionaries represents a strategic enhancement aimed at enhancing LLMs' efficacy in dialect-rich linguistic environments. This augmentation approach, by integrating dialect-specific lexical resources into the model's pre-training regimen, is designed to elevate the model's comprehension and operational performance across varied linguistic landscapes. Such a strategy not only promises improvements in understanding diverse dialects but also champions the cause of inclusivity and accuracy in natural language processing applications.

To complement this augmentation strategy, we incorporate a novel self-referral grading task. This mechanism serves as an internal feedback loop, enabling the model to assess its own performance and adapt more effectively to the nuances of different dialects. This framework showcases a pioneering blend of dialect augmentation and self-evaluation capabilities, setting new approach for adaptability and accuracy in processing dialectical variations within LLMs.

2 Experiment Setup

2.1 DIALECT-COPA Shared Task

The DIALECT-COPA task, a component of the VarDial 2024 shared tasks, presents a unique challenge in understanding causal relationships within scenarios expressed in various South-Slavic dialects (Ljubešić et al., 2024; Chifu et al., 2024). The task involves selecting the most plausible outcome from two options provided, based on the comprehension of causal connections embedded within the context of these dialectal variations. This task aims to explore and evaluate models' capabilities in reasoning across different South-Slavic languages and dialects, emphasizing the nuanced linguistic differences and their impact on reasoning abilities.

Within the task, participants are presented with datasets containing sentences written in South-Slavic dialects, along with adequate annotations of sentences designed to probe the understanding of causal relationships within these linguistic contexts. The datasets encompass a diverse range of dialects, namely Cerkno, Chakavian, and Torlak. This broad representation of dialects enables a comprehensive

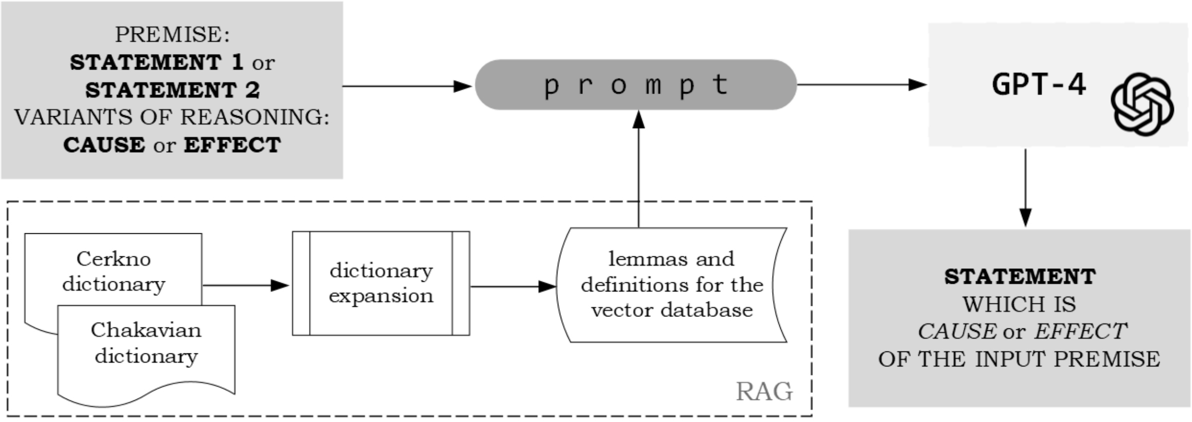

Figure 1: Framework whick inlude advanced prompt engineering and Retrieval Augmented Generation using external data incorporated in dialect dictionaries.

examination of how language variations influence the interpretation and identification of causal connections within the given scenarios. In addition, the organizers have provided several datasets in various languages, such as English, Slovenian, Croatian, Serbian, and Macedonian, annotated specifically for the COPA task. These datasets can be utilized for training and validating the proposed models, enhancing the robustness and versatility of the solutions developed.

To address the complexities of reasoning within dialectal contexts, participants are required to analyze the provided sentences and discern the underlying causal relationships. This task challenges participants to navigate the linguistic intricacies inherent in South-Slavic dialects, such as variations in vocabulary, syntax, and grammatical structures, which may impact the interpretation of causal connections. By engaging with dialectal variations, participants have the opportunity to explore the intersection of language diversity and causal reasoning, thereby contributing to a deeper understanding of how linguistic differences shape cognitive processes.

Taking the example: *The girl found a bug in her cereal.* is translated in Croatian (Hr): *Djevojka je pronašla kukca u žitaricama.*, in Croatian dialect Chakavian (Hr-Ckm) as: *Mlada je našla neko blago va žitaricah.* with options: 1) *Nalila je mlijeko u zdjelicu.* (Hr) and *Nalila je mlieko va škudelico.* (Hr-Ckm) (She poured milk in the bowl.), or 2) *Izgubila je apetit.* (Hr) *Je zgubila tiek.* (Hr-Ckm) (She lost her appetite.) In this analysis, the primary challenge revolves around the semantic ambiguity inherent in language translation, exemplified by the Croatian noun *kukac* and its Chakavian dialect translation *blago*. The term "blago" in Standard Croatian predominantly connotes "treasure" and, to a lesser extent, "cattle". It is notably perplexing to encounter its use in the Chakavian dialect where it assumes the meaning "bug".

This example highlights the broader linguistic issue of polysemy—where a single word has multiple meanings depending on context—which poses significant challenges for accurate translation and understanding. Such ambiguities are especially pronounced in dialects and regional languages, which may not be sufficiently documented, thereby complicating the task of linguistic models like Large Language Models (LLMs) in processing and interpreting such data accurately. For instance, with no data implying the meaning "bug" the reasoning is guided in the direction of misleading translation *The girl found a treasure in her cereal.* with no meaningful connotation to the either of two causal options, which often leads to phenomenon referred as AI semantic hallucination. This situation underscores the necessity for enhanced data inclusivity and sophisticated contextual analysis capabilities in language processing technologies.

2.2 Datsets

We use several available datasets annotated for the COPA task: (i) COPA-HR for Croatian language (Ljubešić, 2021), (ii) COPA-SR for Serbian language (Ljubešić et al., 2022b) and (iii) COPA-MK for Macedonian language (Ljubešić et al., 2022a). In addition, we also used data sets for the corresponding dialects, Chakavian from Croatian, Cerkno from Slovenian and Torlak from Serbian, as well as a data set for the English language.

All datasets use the Latin alphabet, but for the Macedonian and Serbian languages, as well as for the Torlak dialect, we used a dataset written in Cyrillic. The statistics of the distribution of examples into sets for training, validation and testing are shown in the Table 1.

Table 1: Datasets distribution for languages and South-Slavic dialects.

Lang./Dialect	**Train**	**Val.**	**Test**
English	400	100	-
Slovenian	400	100	-
Croatian	400	100	-
Serbian	400	100	-
Macedonian	400	100	-
Cerkno Dialect	400	100	500
Chakavian Dialect	400	100	500
Torlak Dialect	400	100	500

2.3 Prompt Engineering

Prompt engineering has emerged as a valuable technique for improving the effectiveness of LLMs (Reynolds and McDonell, 2021). This method leverages task-specific instructions to expand model capabilities without altering the core model parameters.

In this research, we explore various strategies to formulate prompts to Generative Pre-trained Transformer 4 (GPT-4) (OpenAI et al., 2023) for the COPA task. In the first step, we experiment with the 0-shot learning approach defining a simple prompt and iterating it across all examples in the datasets, for all languages. Here is the prompt used for generating responses in the simple version.

```
This is a reasoning task
with statement in {language} language.
Given the PREMISE = {premise}
please determine {question} of the PREMISE,
choosing between two statements
offered in choices:
CHOICE1 = {choice1}
or
CHOICE2 = {choice2}
Format response in JSON format
with keys:
index:{index}, label: integer(range: 0-1).
```

Next, we pursue the 0-shot learning approach further by providing more comprehensive instructions that incorporates reasoning on the causal relationships within the given sentences. We extend the previous prompt (simple prompt) to advanced prompt with a following instruction:

```
This is statement in {language}.
Based on the PREMISE = {premise}
please determine what could be
the plausible {"cause"|"effect"}
of the PREMISE,
choosing between two choices:
CHOICE1 = {choice1}
or
CHOICE2 = {choice2}
Reason about this problem step by step.
```

Again, we iterate the advanced prompt across all examples in the datasets, for all languages.

In this approach we were able to analyze the reasoning process in the LLM, as well as the elements guiding the erroneous decision in the system. For instance, in the example with the girl and the bug in her cereal, the reasoning in English is correct:

"Finding a bug in one's cereal is generally an unpleasant and unexpected experience that could lead to a loss of appetite. The presence of the bug would likely overshadow the normal routine of adding milk to the cereal, as the discovery would be off-putting and could cause disgust. Therefore, it is more plausible that the effect of finding a bug in her cereal would be that she lost her appetite rather than her continuing with the meal preparation by pouring milk into the bowl."

However, the Chakavian reasoning explanation is plagued with wrong translation of nouns *bug* (Chak. blago) and *appetite* (Chak. tiek). The GPT4 reasoning is as follows: "The premise 'Mlada je našla neko blago va žitaricah' translates to 'The young woman found some treasure in the cereals'. Neither of the choices directly relate to finding treasure in cereals. However, CHOICE_0 'Nalila je mlieko va škudelico' translates to 'She poured milk into a bowl', which could be indirectly related to cereals as one might pour milk into a bowl of cereal. CHOICE_1 'Je zgubila tiek' does not seem to be related to the premise as it translates to 'She lost a thread', which has no apparent connection to finding treasure or cereals. Therefore, CHOICE_0 is selected as it has a weak indirect connection to the premise.".

Consequently, without a comprehensive resolution of the lexical concepts embedded within the sentence, the effectiveness of the reasoning prompt in facilitating accurate comprehension and analysis is significantly diminished. This highlights the critical need for precise semantic interpretation to ensure that cognitive processing mechanisms can effectively engage with and extrapolate meaningful insights from the textual content presented.

2.4 Retrieval-Augmented Generation

In the next step, we propose an approach that combines prompts using GPT-4 and RAG techniques for dialect processing. RAG is a general-purpose fine-tuning technique which combines pre-trained parametric and non-parametric memory for language generation (Lewis et al., 2020). This approach has shown promise through the integration of knowledge from external databases, resulting in improved accuracy and credibility of the models, especially for tasks requiring substantial domain knowledge (Gao et al., 2023). It also facilitates ongoing knowledge updates and the incorporation of domain-specific information.

Within the framework of the DIALECT-COPA task, we utilized an external knowledge base, specifically, a dictionary of the Chakavian dialect and a dictionary of the Cerkno dialect (we could not find a freely available version of the Torlak dialect dictionary). By including dialect dictionaries as external knowledge into the model we enhance our model's understanding of this linguistic variety.

The cornerstone of the Chakavian dictionary dictionary is the work of Cvjetana Miletić (Miletić, 2019), which catalogues approximately 10,800 words, predominantly from the Kastav region in Croatia. In addition, an online version of the Cerkno dictionary called Ana Mičken's zbíerka crklajnskih besít[1], was used as a source of external knowledge for the Cerkno dialect. This dictionary contains about 4000 lemmas with corresponding descriptions and some linguistic examples of usage.

In dialectal dictionaries, in addition to lemmas and corresponding definitions, we also found determinants about the meaning and grammatical category of words. However, we expanded the dictionaries in order to offer as much external knowledge as possible to the model. Thus, we expanded the dictionaries to have the following determinants: lemmas in original dialectal form, lemmas translated into English, examples of use, examples of use translated into English, definitions, expanded definitions and suggestions of use offered by GPT.

We thus proceeded with broadening definitions and examples of usage within the Chakavian and Cerkno contexts, as well as facilitating automatic translations into English. The prompt used for expanding the dictionary examples was structured as follows:

```
Given a dictionary entry "{line}",
expand the definition and provide usage
examples in {language_dict},
with the source lexeme in {language_source}.
Format the response in JSON, including:
'definition': A string containing
the expanded definition,
'GPT_suggestion_of_use':
Two examples of usage in {language_source},
'source_nd_translated':
The source lexeme translated into English,
'example_of_use_translated':
Usage examples translated into English.
```

This methodology was designed to enrich the dictionary and augment GPT-4's ability to comprehend and utilize terms and expressions unique to the Chakavian dialect.

Dictionaries containing linguistic data from the Croatian Chakavian and Slovenian Cerkno dialect were processed using a Retrieval-Augmented Generation (RAG) approach. The process began by transforming data into a CSV format for ease of manipulation and analysis. Following data importation, specialized Recursive Character Text Splitter tool in Langchain library was employed to segment the text into manageable parts, enhancing the handling and vectorization of the data, using the chunk size 200 characters with 50 characters overlap.

The segmented data was then vectorized using an embedding function that facilitated the creation of a persistent vector store in a Chroma database. This vector store serves as a retriever, enabling the efficient retrieval of vectorized text segments based on their semantic content. This setup was designed for integrating the enriched dictionaries with advanced language models, thereby allowing for more contextually aware processing and generation of text based on the Chakavian and Cerkno dialect, respectively.

Enriched dictionaries were used as context in conjunction with the first two previously described prompt strategies (simple prompt engineering strategy and advanced prompt engineering strategy) by inserting the whole enriched dictionary into the prompt.

```
This is a reasoning task
```

---

[1] https://www2.arnes.si/ supmrazp/zbirkacb.htm

```
with statement in {language} language.
Using aditional knowledge
about the {language} language
from the provided dictionary in the
{context}, answer the following
Question:{question}
```

By introducing retrieval augmentation approach, we further evaluated the model's performance on the DIALECT-COPA task.

In addition to the augmentation approach, we introduced a self-referential grading aspect into the prompt. This mechanism establishes an internal feedback loop, empowering the model to autonomously evaluate its own performance. By integrating this self-referral aspect, the proposed approach provides a deeper insight into the reasoning procedures within the DIALECT-COPA task. For this purpose, we define a specific prompt in which we instruct the model to translate the premise and choices into the English language using the dialect dictionary. And then we instruct the model to provide a reasoning for the choice it made. The prompt is formulated as follows:

```
Translate the premise and choices using
the dictionary knowledge in the {context},
especially if the connection of the
premises and choices is apparently weak,
or you do not understand the phrase or
word and the choices.
Provide a reasoning for the choice
you made.
Assess your certainty in a range 0-1.
Format response in JSON format with
keys: index:{index},
label: integer(range: 0-1),
reasoning: str,
certainty: float (range:0-1).
```

The inclusion of a self-referential grading mechanism called *certainty* within the prompt is using the model's autonomous evaluative capabilities with a potential to elevate its ability to introspect and rationalize decision-making processes. Although we did not proceed with the feedback mechanisms based on these values, this approach can be further developed as a framework for the development of more self-aware and adaptive language processing systems in future linguistic research.

3 Results

The initial measurement for English, Macedonian, Serbian, Slovenian, and Croatian was performed on the training data[2] to serve as an orientation and provide an initial insight into the success of the GPT-4 0-shot approach. The performance of the dialects (Chakavian, Cerkno and Torlak) was also measured on available training datasets. Thus, the baseline performance of the GPT-4 model was obtained for all languages in their standard variant, as well as the baseline values for 3 dialects.

The left part of Table 2 shows the model accuracy results for the GPT-4 0-shot with simple prompt (left) and GPT-4 0-shot with advanced prompt framework (right). The results are shown in 3 columns in both cases. The first column contains the measured values for the classification task where the question is *cause*, the second column shows the accuracy when the question is *effect*, and the third column presents a question-independent accuracy (i.e. accuracy on average).

As expected, in the 0-shot - simple prompt variant, but also in the 0-shot - advanced prompt variant demanding iterative reasoning, the model for the English language achieves the highest accuracy (96% and 98.3% on average, respectively). This is expected given the large amount of English language data available to train the model in contrast to the availability of data in other languages. Unexpected, the success of the model in Slovenian and English is identical in the 0-shot - simple prompt variant.

If we consider the 0-shot - simple prompt cases in more detail (left part of the Table 2), two observed facts are interesting. First, the other languages are not far behind the results for English and Slovenian. English, and Slovenian reaches an accuracy on average 96%, while all other languages are below that, but none are below 91%. They are behind by a small number of percentage points. Second, the differences in results with respect to the examination of *cause* or *effect* do not oscillate drastically. The differences are up to 4 percentage points in all cases, except for the Croatian and Serbian, where the deviation is approximately 6 or 8%. Third, also expected, the results for dialects are significantly worse than the results of the same languages in standard varieties. 4% worse for Serbian, 18.3% for Croatian, and even 31.7% for the case of Slovenian.

Unexpectedly, slightly worse results are achieved in the case where the baseline for GPT-4

---

[2] training and validation data are available at https://github.com/clarinsi/dialect-copa/?tab=readme-ov-file

Language	**Dialect**	**Alphabet**	**GPT-4 0-shot - simple**			**GPT-4 0-shot - advanced**			$\overline{\Delta}$
			Cause	**Effect**	**Avg**	**Cause**	**Effect**	**Avg**	
ENG	-	Latin	**.98**	**.941**	**.96**	**.985**	**.98**	**.983**	**+.0225**
MCD	-	Cyrillic	.934	.896	.915	.894	.931	.913	-.0025
SRB	-	Cyrillic	.965	.881	.923	.939	.936	.938	**+.015**
SRB	Torlak	Cyrillic	.894	.871	.883	.828	.881	.855	-.0275
SLO	-	Latin	**.98**	**.941**	**.96**	.904	.921	.913	-.0475
SLO	Cerkno	Latin	.621	.663	.643	.581	.604	.593	-.05
CRO	-	Latin	.97	.916	.943	.965	.901	.933	-.01
CRO	Chakavian	Latin	.748	.772	.76	.677	.629	.653	-.1075

Table 2: Achieved results in terms of accuracy for GPT-4 0-shot - simple promt (left) and GPT-4 0-shot - advanced prompt (right) in terms of accuracy representing the baselines for English, Macedonian, Serbian, Slovenian and Croatian as well as Torlak, Cerkno and Chakavian dialects. The delta column ($\overline{\Delta}$) indicates the average change in accuracy, and refers to the average difference in accuracy achieved by the andvanced and simple prompts.

0-shot are prompted with more sophisticated prompt (right part of the Table 2) against plain 0-shot. This is the case for Macedonian (-0.2%), Croatian (-1%), and Slovenian (-4.7%). However, the results are favorable for Serbian (+1.5% improvement) and English (+2.3% improvement) when sophisticated reasoning is prompted. Although small, a positive change in the performance of the advanced prompt relative to the simple prompt is evident for English (+.0225) and Serbian (+.015) in the last column of the Table 2, where $\overline{\Delta}$ indicates the average accuracy difference between the advanced and simple prompts. A positive value indicates the advantage of advanced prompts over simple prompts.

The preliminary results and findings from this study provide initial evidence that further investigation into sophisticated causal reasoning is warranted, given its modest improvements in accuracy. This suggests potential benefits that could enhance understanding and application in related fields. Moreover, it gave us a strong indication that the model needs an external source of dialect knowledge and the need to apply the RAG paradigm to augment the model's ability to more efficiently handle a dialect it was not initially trained for.

Therefore, in Table 3 experimental results based on advanced prompt engineering which includes iterative reasoning and the RAG approach were presented only for dialects. Test data prepared for the DIALECT-COPA shared task were used (Chifu et al., 2024; Ljubešić et al., 2024) .

Approach	**Cer.**	**Cha.**	**Tor.**
`GPT-4 0-shot-simple`	.664	**.774**	**.894**
`GPT-4 0-shot-advanc.`	.608	.664	.806
`GPT-4-augm-simple`	.688	.76	-
`GPT-4-augm-advanc.`	**.708**	.764	-

Table 3: Results in terms of accuracy for 4 different approaches: (1) GPT-4 0-shot - simple promt, (2) GPT-4 0-shot - advanced prompt, (3) GPT-4 augmented with external knowledge for Cerkno and Chakavian dialect, and (4) GPT-4 augmented with external knowledge for Cerkno and Chakavian dialect and advanced prompt engineering.

Results in terms of accuracy for upgrading the basic 0-shot technique with an iterative reasoning prompt (GPT-4 0-shot-advanced) did not yield fruitful results. The success of such an approach for all dialects recorded a negative effect, namely -5.6% for Cerkno, -8.8% for Torlak and -11% for the Chakavian dialect.

Augmentation of the model with external knowledge (GPT-4-augm-simple approach) from the dialect dictionaries for Chakavian and Cerkno managed to improve the accuracy of Cerkno to 68.8%, which is 2.4% better than the initial 0-shot. However, this was not the case for the Chakavian dialect. One possible reason for this outcome is the incompleteness of the dictionary and lack of polysemous definitions. For example, there is no relation of the Chakavian word *blago* to suggested English equivalent *bug*. Therefore, one way to improve this process would require a better dictionary resources, or even a different way of representing the polysemous nature of lexical networks, such as using lexical graphs (Ban Kirigin et al., 2022; Perak and

Kirigin, 2023).

Finally, with RAG and prompt engineering that requires sophisticated reasoning (GPT-4-augm-advanced), progress is made in both dialects. With such an approach, Chakavian still lags behind the basic 0-shot approach, but Cerkno reaches 70.8%, which is 4.4% more successful than the initial 0-shot.

Experimental findings suggest that the advanced prompt in our study tends to underperform when dealing with languages and dialects that are less familiar to the GPT-4 model. This observation is further substantiated by the statistical measure of Pearson's correlation, which explores the relationship between the simple prompt and the delta (Δ)[3]. The correlations for all languages are negative. The most pronounced values are observed for English with statistically significant correlations ($r = -.7655, p = 3.1040 \times 10^{-78}, \alpha = .05$) and Serbian ($r = -.6202, p = 6.9432 \times 10^{-44}, \alpha = .05$), which is expected given their positive average delta values (refer to the rightmost column in Table 2). Correlation values are also negative for all other languages, but significantly lower, and extremely low for dialects (although for Chakavian and Cerkno are not statistically significant).

4 Conclusion

In this paper, we propose and compare several model adaptation strategies for DIALECT-COPA task. We combine prompt engineering and RAG techniques to enhance the dialect understanding of GPT-4 model.

Research indicated that sophisticated causal reasoning has slight advantages in accuracy, suggesting the need for further investigation. The model requires external dialect knowledge and the application of the RAG paradigm for more efficient handling of dialects for which it was not initially trained. The enhancement of the basic 0-shot method with iterative reasoning did not produce significant results, negatively impacting Cerkno and Chakavian dialects. The model's augmentation with external knowledge from Chakavian and Cerkno dialect dictionaries increased Cerkno's accuracy to 68.8%, a 2.4% improvement over the initial 0-shot. Lastly, employing RAG and prompt engineering that demands complex reasoning led to improvements in both dialects. Although Chakavian still trails the basic 0-shot method, Cerkno achieves an accuracy of 70.8%, marking a 4.4% improvement over the initial 0-shot.

The top achievements of our UNIRI team, as detailed in this paper, hold the second position overall when oposed with the outcomes reported by other teams in the DIALECT-COPA Task on Causal Commonsense Reasoning, a part of the VarDial Evaluation Campaign 2024. Specifically, for the Cerkno dialect, we achieved the second-best result using a simple prompt, with an accuracy of .774. In the case of the Chacavian dialect, we obtained the third-best result by combining an advanced prompt with reasoning and the RAG approach, achieving an accuracy of .708.

The main contribution of this paper is the proposed approach, which utilizes prompt engineering alongside the RAG technique. In this method, RAG facilitates enhancements by integrating Chakavian and Cerkno dictionaries into the prompt. During testing, we showcased that prompts augmented with RAG outperform those without RAG on the Cerkno test dataset. To the best of our knowledge, this represents the first attempt to integrate dialect dictionaries into RAG with the objective of addressing the COPA task focused on causal commonsense reasoning in South-Slavic dialects.

Future work will be concentrated on extending these methods to encompass other dialects and tasks akin to COPA. This includes exploring the adaptation of similar approaches to additional South-Slavic dialects and extending the application to a broader spectrum of tasks requiring nuanced linguistic understanding and reasoning abilities.

Acknowledgments

This work has been fully supported by the University of Rijeka projects uniri-mladi-drustv-23-33, uniri-drustv-23-95, uniri-class-1.

References

Karlo Babić, Sanda Martinčić-Ipšić, and Ana Meštrović. 2020. Survey of neural text representation models. *Information*, 11(11):511.

Karlo Babić, Milan Petrović, Slobodan Beliga, Sanda Martinčić-Ipšić, Mihaela Matešić, and Ana Meštrović. 2021. Characterisation of covid-19-related tweets in the croatian language: framework based on the cro-cov-csebert model. *Applied Sciences*, 11(21):10442.

---

[3] Delta (Δ) represents an absolute value that is calculated as the difference in accuracy achieved by a advanced and simple prompt.

Salvador V Balkus and Donghui Yan. 2022. Improving short text classification with augmented data using gpt-3. *Natural Language Engineering*, pages 1–30.

Tajana Ban Kirigin, Sanda Bujačić Babić, and Benedikt Perak. 2022. Graph-based taxonomic semantic class labeling. *Future Internet*, 14(12):383.

Slobodan Beliga, Sanda Martinčić-Ipšić, Mihaela Matešić, Irena Petrijevčanin Vuksanović, and Ana Meštrović. 2021. Infoveillance of the croatian online media during the covid-19 pandemic: one-year longitudinal study using natural language processing. *JMIR public health and surveillance*, 7(12):e31540.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. *arXiv preprint arXiv:2310.15135*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Nikola Ljubešić. 2021. Choice of plausible alternatives dataset in croatian copa-hr.

Nikola Ljubešić, Mirjana Starović, Taja Kuzman, and Tanja Samardžić. 2022a. Choice of plausible alternatives dataset in serbian copa-mk.

Nikola Ljubešić, Mirjana Starović, Taja Kuzman, and Tanja Samardžić. 2022b. Choice of plausible alternatives dataset in serbian copa-sr.

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.

Ana Meštrović, Milan Petrović, and Slobodan Beliga. 2022. Retweet prediction based on heterogeneous data sources: the combination of text and multilayer network features. *Applied Sciences*, 12(21):11216.

Cvjetana Miletić. 2019. *Slovar kastavskog govora [Dictionary of the Kastav Speech]*. Udruga Čakavski senjali, Kastav.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. Computer Science 2303.08774, arXiv.

Benedikt Perak and Tajana Ban Kirigin. 2023. Construction grammar conceptual network: Coordination-based graph method for semantic association analysis. *Natural Language Engineering*, 29(3):584–614.

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*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

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*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, pages 194–206. Springer.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for llms. *arXiv preprint arXiv:2403.11810*.

Tedo Vrbanec and Ana Meštrović. 2023. Comparison study of unsupervised paraphrase detection: Deep learning—the key for semantic similarity detection. *Expert systems*, 40(9):e13386.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167*.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. Towards making the most of bert in neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9378–9385.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.