ChatGPT as Your n-th Annotator: Experiments in Leveraging Large Language Models for Social Science Text Annotation in Slovak Language

Endre Hamerlik[1,2], Marek Šuppa[1,3], Miroslav Blšták[4], Jozef Kubík[1], Martin Takáč[1], Marián Šimko[4], Andrej Findor[5],

[1]Faculty of Mathematics Physics and Informatics, Comenius University in Bratislava, Slovakia
[2]ELKH Institute for Computer Science and Control (SZTAKI), Hungary
[3]Cisco Systems, Slovakia [4]Kempelen Institute of Intelligent Technologies, Slovakia
[5] Faculty of Social and Economic Sciences, Comenius University in Bratislava, Slovakia

Abstract

Large Language Models (LLMs) are increasingly influential in Computational Social Science, offering new methods for processing and analyzing data, particularly in lower-resource language contexts. This study explores the use of OpenAI's GPT-3.5 Turbo and GPT-4 for automating annotations for a unique news media dataset in a lower resourced language, focusing on stance classification tasks. Our results reveal that prompting in the native language, explanation generation, and advanced prompting strategies like Retrieval Augmented Generation and Chain of Thought prompting enhance LLM performance, particularly noting GPT-4's superiority in predicting stance. Further evaluation indicates that LLMs can serve as a useful tool for social science text annotation in lower resourced languages, notably in identifying inconsistencies in annotation guidelines and annotated datasets.

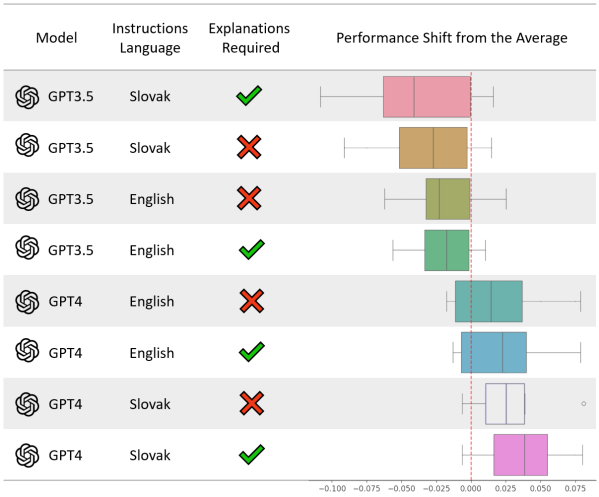

Figure 1: Prompting Strategy Grid Search: Slovak prompts exhibit the lowest effectiveness with `GPT-3.5` but secure the highest performance with `GPT-4`. Notably, incorporating explanations within prompts significantly enhances effectiveness across models. This boost is particularly pronounced for Slovak prompts used with `GPT-3.5` and English prompts with `GPT-4`. The red dashed line represents the 'zero shift' from the average performance, illustrating changes in F1 score relative to the average across all parameter combinations.

1 Introduction

The emergence of Large Language Models (LLMs) has not only revolutionized the field of natural language processing (NLP) (Min et al., 2023; Chang et al., 2023) but also significantly impacted social sciences (Teubner et al., 2023; Ziems et al., 2024). These models' ability to understand and generate human-like text has opened new avenues for analyzing complex social phenomena such as political discourse (Bornheim et al., 2023), public opinion (Lee et al., 2023), and media analysis (Jiang et al., 2023) with unprecedented precision.

This progress has set the stage for augmenting, or even substituting, human annotators in tasks demanding profound linguistic and semantic insights (Heseltine and Clemm von Hohenberg, 2024; Ollion et al., 2023; He et al., 2023). Our research explores the application of LLMs, coupled with sophisticated prompting strategies, to scrutinize Slovak news media content on migration, a topic with deep societal and political ramifications. The dataset, created for project MIMEDIS[1], seeks to unravel how media shapes public migration viewpoints, integrating computational and manual analysis. Our investigation pivots on employing LLMs for annotating social science materials in less commonly used languages, revealing that advanced prompting methods can position LLMs as viable alternatives or complements to traditional supervised fine-tuning. Our findings also suggest that non-English instructions could enhance LLM performance, as outlined in Figure 1. We hope that this study will help highlight the LLMs' potential as a helpful tool in annotating social science texts

[1]See https://cogsci.fmph.uniba.sk/MIMEDIS/index.html.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 81–89
July 5–10, 2020.

in lower resourced languages and instigate further development in this area.

2 Related Work

Large Language Models (LLMs) have garnered a significant amount of interest over the past few years, especially due to their unprecedented ability to generalize based just on zero-shot input, or from just a handful of examples, also known as few-shot learning. When combined with advanced prompting strategies such as Retrieval Augmented Generation (RAG) (Lewis et al., 2020) and Chain of Thought (CoT) prompting (Wei et al., 2022), this makes LLMs state-of-the-art methods for various NLP and text understanding tasks (Min et al., 2023).

Among other achievements, this has led to research that suggests that LLMs such as GPT-3.5-Turbo and GPT-4 can be adapted for annotation (He et al., 2023; Belal et al., 2023; Thapa et al., 2023) and in some cases even as a potential replacement for human annotation (Heseltine and Clemm von Hohenberg, 2024) as it was able to perform on-par or better than a human annotator (Gilardi et al., 2023). On the other hand, closer inspection by (Ollion et al., 2023) has found that "fewshot learners offer enticing, yet mixed results on text annotation tasks", suggesting that evidence for aforementioned claims is only partial at best.

Despite the partial evidence, LLMs still present an interesting option, particularly for languages which lack large-scale data resources and for which the cost of annotation is often significant due to the low number of native speakers and/or experts available, which is the case in our situation as well. Perhaps the most similar work to ours would be (Mets et al., 2023) in which the authors evaluate stance of sentences in Estonian news articles about immigration and compare the performance of supervised models with ChatGPT, finding that ChatGPT obtains similar performance. In contrast, in our work we explore a problem that can be viewed as multi-target and multi-class, we further consider the article-level as opposed to sentence-level stance, employ multiple LLMs (`GPT-3.5-Turbo` and `GPT-4`) and a number of advanced prompting strategies such as RAG and CoT prompting, which make our best performing LLMs capable of performing better than supervised models.

3 Dataset

To evaluate our models we utilize a specific Slovak dataset annotated for classification across various dimensions. The dataset aims to understand migration representation in Slovak media spanning from 2003 to 2022, targeting individual media outputs like articles and debate transcripts. We briefly outline the specific dimensions below.

Thematic Relevance Articles are classified based on relevance to human migration within the study period, marked as *strong*, *weak*, or *irrelevant*.

Geographical Relevance This categorization differentiates between articles *related to Slovakia* and those not.

Migration Direction It identifies if the migration is towards (*immigration*) or away from Slovakia (*emigration*).

Stance The media's stance toward migration is tagged as *positive*, *negative*, or *neutral* for the below listed targets: targeting migrants (*people*), facilitators of migration (*enablers*), and migration policies (*policies*). If a target is not mentioned in an article, annotators assign a label indicating the target is *not mentioned*.

As the scope of the Slovak media outputs between 2003 and 2020 is vast (we were able to obtain on the order of 800k items that contained migration-related keywords[2]), they were sampled in a stratified way on per-year basis. Each media output in the dataset was annotated by at least three different annotators via an Argilla[3] interface and only instances in which majority agreement was observed were included in the final dataset.

A visualization of the lengths of the media output contained in the respective subsets of the final dataset can be found in Figure 2.

As the majority of the media outputs in the final dataset are shorter, the truncation to 2,500 characters impacted 36%, 36% and 34% of the samples across the train, validation and test splits.

We conducted a similar analysis using the `gerulata/slovakbert` tokenizer[4] which is part of the SlovakBERT model.

The resulting distribution across the three splits can be seen in Figure 3. As the distribution in the figure suggests, the majority of media outputs

[2] See Table 2 for the list of the keywords.
[3] https://argilla.io
[4] https://huggingface.co/gerulata/slovakbert

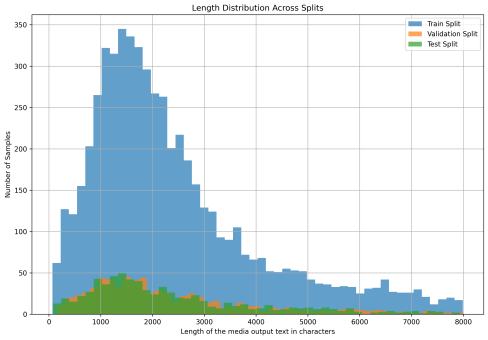

Figure 2: Character length distribution in the final dataset.

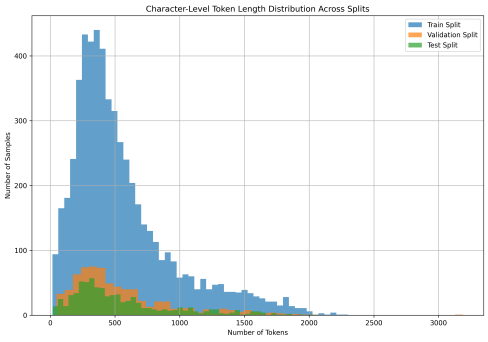

Figure 3: Token length distribution in the final dataset.

contained in the final dataset are shorter than 512 tokens. That being said, truncating the input to this token length impacted 41%, 40% and 40% of the samples in the train, validation and test set, respectively.

More information on dataset creation can be found in Section A.1.

The final dataset contains about 7.2k annotated articles, making it the largest Slovak classification dataset and the biggest in the realm of Political Social Science. More detailed statistics of the dataset can be found in Section A.2.

4 Methods

Our experimental methodology was informed by three key guidelines to ensure uniformity and comparability of results. To standardize input data, articles were cut to 2500 characters. This length not only aligns with the maximum input capacity of our baseline models but also helps in managing OpenAI credit usage efficiently.

The evaluation of annotations was carried out on the test split of the pre-annotated dataset. The experiments were structured with a single-task focus, i.e. there was a dedicated model for each task. Each model was tasked to predict a single label.

4.1 Grid search for prompting strategies

We initiated our experiments by evaluating different prompting strategies to identify critical hyperparameters, focusing on GPT-3.5 Turbo (specifically `gpt-3.5-turbo-0125`) and GPT-4 (`gpt-4-0125-preview`). Our analysis included examining the effect of prompting language on the process, particularly emphasizing that Slovak is considered troublesome in prompt engineering.[5] We compared the use of English and Slovak prompts, noting that prompts requiring detailed responses were more effective. This led to tests with and without such prompts.

4.2 Retrieval Augmented Generation

We incorporated Retrieval Augmented Generation (RAG) (Lewis and Oguz, 2020) into our experiments, leveraging its blend of retrieval-based and generative methods to augment prompts with relevant documents, thus improving response generation. SentenceBERT (Reimers and Gurevych, 2019) was used to embed data, aiding in the retrieval of the top-k (k=3 has been chosen based on preliminary tests and cost considerations) articles from the train set for model input. These articles, selected based on similarity in a vector database, were presented with the model prompts, details of which can be found in Appendix A.3. Our tests on `GPT-3.5` and `GPT-4` examined the effectiveness of various prompt languages and requiring explanations, in terms of the language used for prompts and the incorporation of explanation requests within the prompts.

4.3 Chain of Thoughts

Chain of Thought (CoT) prompting in Large Language Models (LLMs) is a strategic approach that prompts the model to reveal step-by-step reasoning before arriving at a conclusion, thereby improving the depth and logic of its outputs (Ma et al., 2023; Kojima et al., 2022). In our study, we embed CoT prompting within a dual-stage framework as discussed in Section 4.1. Initially, the prompt sets the stage with specific instructions, providing two annotated example articles and a system message highlighting the task's objective. Following this,

[5] See for instance https://community.openai.com/t/slovak-language-not-working-well/579305

the second stage of the prompt presents a system directive to choose an appropriate task label and includes a succinct request for the annotation of a given article. The detailed structure of this prompting strategy can be found in Appendix A.3.

4.4 Finetuned baselines

In order to provide a direct comparison with models within the standard supervised finetuning framework, we employ a selection of well established baselines relevant for the Slovak language: the mBERT model,[6] the multilingual version of BERT (Devlin et al., 2018), XLM-R,[7] (Conneau et al., 2019) a larger-scale pre-trained multilingual model based on the RoBERTa architecture and SlovakBERT,[8] (Pikuliak et al., 2021) a BERT-based model pretrained specifically on a large Slovak corpus and the current state-of-the-art model for many Slovak tasks. To provide uniformity across the evaluated models, we finetune all of them for five epochs using the AdamW optimizer and learning rate set at 2e-5. The models were provided the concatenation of the headline and the main text of the media output, and the inputs were further truncated at 512 tokens in order to conform to the requirements of BERT models.

5 Results

5.1 Baselines

As illustrated in Table 1, each baseline model surpassed the majority class baseline. Among them, the `slovakbert` models stood out, achieving the highest scores across most categories, with only a slight exception in `theme_relevance` where the difference was negligible and `target_people` where the difference was more pronounced.

5.2 Grid search for prompting strategies

The grid-search analysis showed performance differences across various setups. `GPT-3.5` performed best with English prompts, reaching a 70.22 F1 score, but slightly dropped to 69.44 when explanations were added. Conversely, `GPT-4` excelled with Slovak explanation prompts, achieving a 76.97 F1 score, a substantial rise from 75.21 with English prompts without explanations. This suggests that the explored task has a high language dependency and benefits from prompting in the language native to the input (Liu et al., 2024). Figure 1 elucidates the influence of various prompting parameters on the models' performance. Furthermore, Table 4 compiles the F1 scores for the different setups across all tasks within the dataset in a descending order, highlighting the relative advantages of specific configurations.

5.3 RAG and CoT experiment evaluations

The performance of `GPT-4 RAG` was notable in the collection of classification tasks, showing higher average accuracy. Its proficiency was especially prominent in geo-relevance prediction, where it outperformed other LLM experiments. Table 1 presents the top-performing configuration for each model. The success of `GPT-4 RAG` indicate the benefits of Retrieval Augmented Generation (RAG) in enhancing model capabilities, providing a significant enhancement compared to baseline models and underscoring the value of integrating external knowledge sources.

In a detailed comparison, `GPT-4 RAG` consistently surpassed `GPT-3.5 RAG`, with an impressive average F1 score difference of up to 6 F1 points. This underscores the advancements in `GPT-4`'s architecture and training compared to its predecessor. Interestingly, when employing the Chain of Thought (CoT) method, `GPT-3.5` achieved notable results in theme relevance and matched `GPT-4 RAG` in direction prediction accuracy, as indicated in the corresponding F1 scores.

However, the performance of `GPT-4 CoT` fell short of expectations, suggesting that the CoT method's performance might be task-dependent or influenced by specific model characteristics. This discrepancy invites further investigation into the CoT methodology's application in LLMs, potentially leading to innovative approaches like Retrieval Augmented Thoughts (Wang et al., 2024), which could merge the strengths of RAG and CoT for even more refined performance. This area represents a promising direction for future research, utilizing the synergy between different prompting strategies to enhance task-specific outcomes.

6 Discussion

The results in Table 1 show that RAG and CoT enhancements led `GPT-3.5` and `GPT-4` models to out-

[6] `bert-base-multilingual-cased`: https://huggingface.co/google-bert/bert-base-multilingual-cased

[7] `xlm-roberta-base`: https://huggingface.co/FacebookAI/xlm-roberta-base

[8] `slovakbert`: https://huggingface.co/gerulata/slovakbert

model	theme_relevance	geo_relevance	direction	target_people	target_enablers	target_policies
majority class	74.6898	62.3894	74.3386	36.6782	47.4637	30.6667
`bert-base-multilingual-cased`	**79.9007**	93.8053	78.8359	53.6332	47.8261	49.6667
`xlm-roberta-base`	78.1638	**94.6903**	80.9524	**59.1696**	50.0000	51.6667
`slovakbert`	79.1563	94.2477	**82.8042**	53.9792	**52.1739**	**53.6666**
`GPT-3.5 RAG`	85.4749	74.7967	90.9496	63.1090	47.1545	59.3968
`GPT-3.5 CoT`	**85.6346**	54.7170	**93.1686**	63.1090	45.9016	57.0755
`GPT-4 RAG`	83.5227	**96.7598**	**93.1686**	**69.7572**	**48.3871**	**65.3333**

Table 1: Micro F1 scores for various models and model types on the test set. As per the parameter search, GPT models were prompted in Slovak. The best performance is in bold.

perform finetuned baselines by up to 11 F1 points in most categories, except for `target_enablers`. However, some categories had low absolute F1 scores, the lowest being 48.3871. Analysis of RAG models in Figure 4 indicated a preference for "No Target" over "Positive" or "Negative" labels, suggesting these models aren't ready to replace human annotators in complex political topics like migration, yet. Although, a manual review of the models' explanations by one of the authors found them mostly logical, hinting at potential issues in the annotation guidelines or process rather than the models' capabilities. This is also reflected in the Inter-Annotator Agreement in Table 3, measured by Krippendorff's alpha (Castro, 2017), which indicated a relatively low agreement for many tasks.

In summary, while LLMs with advanced prompting have progressed, we do not yet find them to be viable replacements for human annotators in text annotation in the realm of Computational Social Science. They are, however, valuable for highlighting problems in annotation guidelines and datasets, effectively serving as an additional, or as the paper's title suggests, n-th, annotator. We leave further exploration of this concept as well as its potential implication to future work.

7 Conclusion

This study evaluates the performance of LLMs in automating stance classification tasks within a Slovak news media dataset, emphasizing the impact of advanced prompting strategies and native language instructions. The results indicate that while large language models (LLMs), particularly GPT-4, significantly outperform BERT-based baseline models, they still lack the ability to fully replace human annotators in complex tasks such as stance classification in political texts under the conditions of our experiments. However, their ability to uncover inconsistencies in annotation guidelines and datasets highlights their potential as valuable tools in social science research. The findings from this study have enabled the MIMEDIS project team to refine the annotation manual and to distinguish between inherently difficult tasks and those that are simply underdefined. Just like if chatGPT was our n-th expert annotator.

Limitations

- As our analysis has been done on a dataset in Slovak language, its conclusions might not be directly applicable to other languages.
- The analysis has been done using models which are accessed via paid APIs and might hence not be widely accessible.
- While article-level annotation and single-label classification were chosen to align with the goals of our project, we acknowledge that these choices may not suit all potential tasks, such as mention detection or cases involving multiple overlapping themes. Lower-level annotations would significantly increase the complexity and duration of the annotation process, making it impractical for our purposes. Additionally, we recognize that the lower IAA agreement observed for certain tasks may partially stem from these choices.
- We recognize the potential inconsistency in our methodology, where annotators had access to the full article text, while models like the LLM and transformer encoders processed only truncated versions (up to 2,500 chars or 512 tokens, respectively). This discrepancy could contribute to differences in performance and agreement.

Acknowledgments

We thank the 100 annotators who were instrumental in creating the dataset central to this work. This project was supported by grant APVV-21-0114.

References

Mohammad Belal, James She, and Simon Wong. 2023. Leveraging chatgpt as text annotation tool for sentiment analysis. *arXiv preprint arXiv:2306.17177*.

Tobias Bornheim, Niklas Grieger, Patrick Gustav Blaneck, and Stephan Bialonski. 2023. Speaker attribution in german parliamentary debates with qlora-adapted large language models. *ArXiv*, abs/2309.09902.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1):20531680241236239.

Yan Jiang, Ruihong Qiu, Yi Zhang, and P. Zhang. 2023. Balanced and explainable social media analysis for public health with large language models. In *Australasian Database Conference*.

Daiki Kojima, Sho Oura, Yusuke Iwasawa, and Yutaka Matsuo. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

S. Lee, T. Q. Peng, M. H. Goldberg, S. A. Rosenthal, John E. Kotcher, Edward W Maibach, and Anthony Leiserowitz. 2023. Can large language models capture public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *ArXiv*, abs/2311.00217.

Patrick Lewis and Sergey Edunov Danqi Chen Mandar Joshi Mike Lewis Luke Zettlemoyer Veselin Stoyanov Oguz, Bhuwan Dhingra. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

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. *arXiv preprint arXiv:2403.10258*.

Xilai Ma, Jing Li, and Min Zhang. 2023. Chain of thought with explicit evidence reasoning for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352, Singapore. Association for Computational Linguistics.

Marko Mets, Andres Karjus, Indrek Ibrus, and Maximilian Schich. 2023. Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media. *ArXiv*, abs/2305.13047.

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 Computing Surveys*, 56(2):1–40.

Etienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2023. Chatgpt for text annotation? mind the hype! *SocArXiv. October*, 4.

Matú Pikuliak, Stefan Grivalsky, Martin Konopka, Miroslav Blták, Martin Tamajka, Viktor Bachrat'y, Marián Simko, Pavol Balázik, Michal Trnka, and Filip Uhl'arik. 2021. Slovakbert: Slovak masked language model. *ArXiv*, abs/2109.15254.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Timm Teubner, Christoph M Flath, Christof Weinhardt, Wil van der Aalst, and Oliver Hinz. 2023. Welcome to the era of chatgpt et al. the prospects of large language models. *Business & Information Systems Engineering*, 65(2):95–101.

Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. 2024. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, pages 1–55.

A Appendix

A.1 Dataset creation

In order to arrive at a dataset representative of the migration-related discourse in Slovak media, a multi-step approach was applied.

First, an export of all Slovak media outputs that contained at least one of the migration-related keywords in one of their possible lexical forms. The full list of lemmas can be found in Table 2. This process has yielded 802,503 media outputs in total.

Second, the media outputs were filtered for length, where only those with the length of less than 8,000 characters (on the order of 1,000 words) as these were found to be long listings of for instance the TV programme for a specific day or listings of news agency output for a specific day, which would not materially contribute to the aim of our analysis. This process has filtered out 78,763 media outputs, representing 9.82% in total.

Finally, the export was sampled on per-year basis in a stratified in order for smaller batches of media output to be supplied to the annotators. This was done primarily to ensure the distribution of migration-related media outputs in the final dataset across the years is as close as possible to that of the aforementioned export, which is thought to be derived from all of the media output produced in Slovak between 2003 and 2020.

A.2 Dataset statistics

The Table 5 describes the distribution of samples across the various configurations which are based on the categories discussed in Section 3.

Slovak Lemma	English Translation
migrant	migrant
migrantka	female migrant
imigrant	immigrant
imigrantka	female immigrant
emigrant	emigrant
emigrantka	female emigrant
utečenec	refugee
utečenka	female refugee
utečenkyňa	female refugee (alternative form)
odídenec	displaced person
odídenka	displaced female
odídenkyňa	displaced female (alternative form)
azylant	asylum seeker
azylantka	female asylum seeker
cudzinec	foreigner
cudzinka	female foreigner
expat	expat
expatka	female expat
expatriant	expatriate
expatriantka	female expatriate
vysťahovalec	emigrant
vysťahovalkyňa	female emigrant
vyhnanec	exile
vyhnankyňa	female exile
exulant	exile
exulantka	female exile
vysťahovalectvo	emigration
azyl	asylum
migrácia	migration
imigrácia	immigration
emigrácia	emigration
migračný	migration-related
migrantský	migrant-related
imigrantský	immigrant-related
emigrantský	emigrant-related
utečenecký	refugee-related
odídenecký	displaced-related
cudzinecký	foreigner-related
vysťahovalecký	emigrant-related
migrantov	migrants (plural)
migrantkin	female migrants (plural)
utečencov	refugees (plural)
utečenkin	female refugees (plural)
imigrantov	immigrants (plural)
imigrantkin	female immigrants (plural)
odídencov	displaced persons (plural)
odídenkin	displaced females (plural)
emigrovať	to emigrate
imigrovať	to immigrate
migrovať	to migrate

Table 2: The terms used to search for Slovak migration-related news outputs and their English translations

A.3 Prompting strategy

CoT Prompt structure

`sys message1:`
Try to think about why the given annotations might be correct

`human message1:`
Extract from the Annotation manual, including 2 annotated examples

`response1:`
The Annotations are correct...

`sys message2:`
You are an expert Slovak annotator. Your answers should ONLY contain ONE of the following labels:
labels

`human message2:`
'These are just a few examples. Please annotate the text below following the scheme of the examples provided above:
Article to be annotated

`response2:`
Annotations

RAG Prompt structure

`sys message:`
You are an expert Slovak annotator. Your answers should ONLY contain ONE of the following labels:
labels

`human message:`
First, I will give you some annotated examples:
##Annotated examples from the vector db of train and valid sets##

'These are just a few examples. Please annotate the text below following the scheme of the examples provided above:
Article to be annotated

`response:`
Annotations

A.4 Inter Annotator Agreement

Task	Krippendorff's alpha
theme_relevance	0.3258
geo_relevance	0.7375
direction	0.3627
target_people	0.1754
target_enablers	0.1167
target_policies	0.1958

Table 3: Inter Annotator Agreement between the human annotators represented as Krippendorff's alpha.

A.5 Grid search results

Parameters			Average F1 Score
model_name	sk_prompts	explanations	
GPT-4	True	True	76.9660
GPT-4	True	False	76.2066
GPT-4	False	True	75.6418
GPT-4	False	False	75.2091
GPT-3.5	False	False	71.4464
GPT-3.5	False	True	71.3757
GPT-3.5	True	False	70.2208
GPT-3.5	True	True	69.4447

Table 4: Average F1 Scores for different prompting strategies (grid search results).

Configuration	Train	Validation	Test
default	5828	728	729
theme_relevance	3316	413	403
geo_relevance	3727	455	452
direction	3097	395	378
target_people	2394	296	289
target_enablers	2317	275	276
target_policies	2423	290	300

Table 5: Dataset statistics across various subsets

A.6 Confusion Matrices

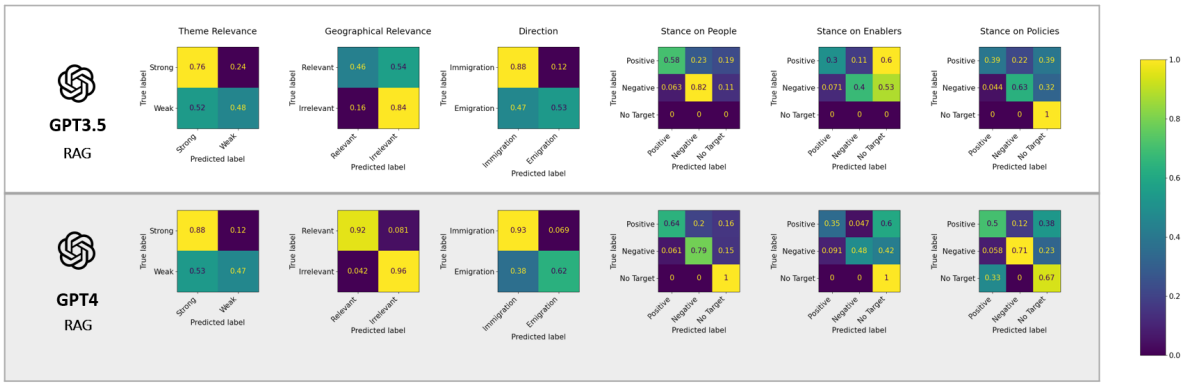

Figure 4: Confusion Matrices for `GPT-3.5 RAG` and `GPT-4 RAG`