Adaptation of Large Language Models for Spanish Text Generation in Responsible AI Problems

María Estrella Vallecillo Rodríguez

Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain

Abstract

The Internet and social networks are a fundamental part of people's lives. They use them for various aspects of their lives, such as interacting with people who are far away, keeping up to date with what is happening around the world or even expressing their own opinion on an issue. The problem is that appropriate content is not always posted on these networks, and they are sometimes used to promote offensive stereotypes or spread false information that can have very harmful effects on users. Therefore, the application of NLP (Natural Language Processing) techniques is very important, since it allows to control of the large volume of data that these networks contain. This doctoral thesis focuses on the use of Large Language Models (LLMs) for automatic generation of counter-arguments to messages posted on social networks to fight against misinformation and offensive messages.

Keywords

Large Language Models, Automatic Counter-argumentation, Natural Language Generation, Natural Language Processing

1. Justification of the proposed research

Nowadays, Internet and social networks are our most widely used means of communication. With them, we can connect with people who are far away and keep up to date with everything that is happening on the other side of the world. The problem is the use that internet users make of social networks is not always appropriate, using them on many occasions to spread false information, offend other people, promote stereotypes, or radicalise users, forcing them to share their ideology or opinion.

The strategy most commonly used so far with this type of content has been to delete the message containing erroneous or offensive information or to temporarily block users within the social network. However, this strategy can provoke a perception of censorship and limitation of freedom of expression in users who write these messages. This perception causes resistance to attitude change in the users who write these messages. Furthermore, in the case of offensive messages censorship does not address the underlying motives of hate crime and misses the opportunity to educate users [1]. For these reasons, a new strategy, automatic counter-argumentation generation, is explored. This strategy consists of the elaboration of a response containing truthful and solid arguments to refute the false information contained in the messages or the stereotype promoted with it. With this response we seek to combat misinformation and challenge the stereotypes that are generated in social networks with offensive messages, offering an alternative and constructive perspective to encourage empathy, understanding, and tolerance among users, thus promoting a more inclusive and respectful online environment.

The purpose of this thesis is to be able to elaborate systems that can refute with solid arguments these messages that are spread through the network with the simple aim of doing harm. To address this task, we will use large language models (LLMs), because their evolution has allowed us to have a breakthrough in solving complex tasks. For example, LLMs are able to understand a task through a small set of examples or to execute textual instructions. To solve these tasks, the language models have been trained with a massive amount of data, and have increased in size. However, recent work by Hoffmann et al. [2] shows that it is not always better to use larger models, but smaller models that have been trained with more data. In addition, training smaller models with more data means that such

Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.

mevallec@ujaen.es (M. E. V. Rodríguez)

0000-0001-7140-6268 (M. E. V. Rodríguez)

models use substantially less compute for fine-tuning and inference, greatly facilitating downstream usage. This is why we will try to train LLMs to adapt them to the problem we want to solve. In addition, we want to focus on Spanish, since it is a language that does not have as many resources as other languages in this task and it is a field that is not yet being explored in depth.

2. Related work

The automatic generation of counterarguments can be applied to multiple responsible AI problems, particularly in combating misinformation and offensive messages on social networks (currently these are the fields we are focusing on in this thesis). Imagine a scenario in which an AI system instantly responds to a harmful social media post that spreads misinformation about a public health issue or counters a hateful message directed at a minority group. By providing accurate information and promoting respectful dialogue, such systems can significantly improve the quality of online interactions and contribute to a more informed and inclusive society. To develop a system for these purposes it is crucial to explore some of the work that has been done in recent years on these tasks:

2.1. Misinformation

Among the texts that promote disinformation we find texts containing fallacies or persuasion techniques, conspiracy texts or even fake news.

Our interest in this area lies in studies such as [3] in which it is exposed that news headlines are more dangerous than the fake news itself, considering that a fake news already generates enough damage in our society.

In this area we find studies such as [4] in which they try to detect those users who manipulate information in social networks, the learning of causal models of disinformation and social manipulation, and the detection of disinformation generated by LLMs. Other studies such as [5] tries to address fake news detection by using LLMs and the prompt strategy known as Chain of Thought, as these models can generate logical reasoning that validates or criticizes news headlines. This strategy is beneficial as it combines the predictive ability of LLMs with the requirement for coherent explanations, facilitating not only the detection of fake news, but also providing transparent and reasoned justifications for each classification. Finally, there are studies such as [6] that propose advanced solutions for fact checking exposed in fake news. Therefore, they explore methods based on retrieval augmented generation (RAG). This work proposes two novel methodologies, Chain of RAG (CoRAG) and Tree of RAG (ToRAG). These approaches improve the accuracy of veracity predictions and explanation generation over the traditional fact-checking approach.

2.2. Offensive Messages

To address offensive messages with strategies based on counter-argumentation, there is a task known as automatic generation of counter-narratives. A counter-narrative can be defined as an elaborated response to negate a message in a respectful and constructive way. Such a response can be argued, providing accurate and truthful information (known as a rebuttal), or unargued, simply rejecting the idea of the message to which it responds (known as a rejection). Counter-narrative applied to eliminate hate speech aims to challenge the stereotypes contained in offensive messages and foster empathy, understanding and tolerance among social network users, thus promoting a more inclusive and respectful online environment.

In the field of automatic generation of counter-narratives, research has been conducted in three important areas: studies on the usefulness and benefits of counter-narratives, the creation of quality datasets for the generation of counter-narratives, and the exploration of various methods for automatic generation of counter-narratives.

Relating to studies of the benefits of counter-narratives, we find the work of Schieb and Preuss [7]. In this study, they show that the factors that define the success of counter-narratives are the proportion

of offensive messages that we encounter and the influence that the people in charge of carrying out the counter-narrative can exert on the undecided. Munger [8] and Mathew et al. [9] found that subjects who were educated through counter-narratives significantly reduced the use of racist insults on Twitter. Furthermore, Benesch [10] and Mathew et al. [11] suggest that the use of counter-narratives can be considered one of the most promising approaches against hate speech and the use of offensive messages.

Regarding the creation of datasets for the development of counter-narrative generating systems we find studies such as Guerini [12] where a corpus called "CONAN: Counter-narratives datasets to fight hate speech" [1] has been created, formed by four datasets useful to fight online hate speech through the generation of counter-narratives, including datasets for multiple offensive targets, including knowledge information or applied for different languages. In addition, other works presents a corpus designed to classify both offensive message and counter-narrative [13], or Mathew et al. [9] where a different approach is taken, as different social network user accounts are analyzed, (accounts that wrote many offensive messages as users who countered hate speech).

Looking at papers describing the methods used for automatic generation of counter-narratives we have works like Chung et al. [14] that provide an online platform to monitor and perform counter-narrative to Islamophobic messages. Qian et al. [15] approaches the task with three different methods (sequence-to-sequence models (SeqToSeq), variational autoencoders, and reinforcement learning), or the use of different linguistic models, including pretrained models and LLMs [16, 17, 18]. In addition, other studies [19, 20] includes external information to avoid model hallucinations or applying counternarrative generation to languages with fewer resources. A more recent and novel study is that of Bonaldi et al. [21], where researchers regulate the attention of Transformers models to improve the generalization capabilities of these models.

Finally, we should mention the work of Chung et al. [22], which presents an extensive study on the current state of automatic generation of counter-narratives, ranging from the systems that generate them and their evaluation to the datasets created to develop these systems (analyzing the languages of the resources and the sources from which the data are extracted). Although this review does not include any work in Spanish.

3. Hypothesis and objectives

We assume the following hypothesis: Given a large language model, to be able to solve different AI responsible problems by generating texts containing solid and truthful arguments in Spanish.

With this hypothesis, the following objectives are established:

- Analyze and characterize the problems of the "responsible artificial intelligence" related to language and where argumentation can be applied to solve them.
- Understand in depth the different types of texts that are used to misinform or promote stereotypes.
- Investigate existing resources related to counter-argumentation and develop new resources specific to Spanish.
- Conduct various experiments based on prompting or adapting language models (fine-tuning) using existing and new datasets.
- Participate in evaluation campaigns to assess and improve the systems developed.
- Share the results obtained with the scientific community by writing articles and propose the organization of shared tasks related to research.

At this stage of the thesis, we are at the first point.

4. Methodology and proposed experiments

Until now, we have focused on the automatic generation of counter-narratives. For this purpose, several experiments have been carried out related to the automatic generation of counter-narratives to combat

[1] https://github.com/marcoguerini/CONAN

offensive messages and the different stereotypes present in them. Therefore, two simple corpora and several systems based on prompting and adapted to Spanish have been developed to serve as a starting point and support for future work.

Regarding the problem of misinformation, we intend to access a knowledge base of current news and information in order to use this information to elaborate a message that tries to combat messages that are aimed at confusing people, either with fake news or with information that is not entirely true. To develop this kind of system, we are considering systems based on RAG (Retrieval Augmented Generation) [23] or LLM with different prompting strategies such as CoT or ToT.

4.1. Datasets

Two dataset for Spanish automatic counternarrative generation was generated. The first called CONAN-SP [24] is available in GitHub [2] and the second is entitled CONAN-MT-SP [25] that was used in the shared task RefutES hosted at IberLEF and is available in GitHub[3].

4.1.1. CONAN-SP

CONAN-SP is based on CONAN-KN [16] that consists of 195 HS-CN pairs covering multiple hate targets *(islamophobia, misogyny, antisemitism, racism, and homophobia)*, provided along with the relevant knowledge automatically retrieved. Since CONAN-KN is in English, we use DeepL, an automatic translator tool to translate English pairs to Spanish.

To construct CONAN-SP, we remove the pairs that contain duplicates of hate-speech texts and the examples used to calculate the agreement between annotators. The structure of CONAN-SP is the hate-speech provided by CONAN-KN and the counter-narrative texts generated by GPT-3.5 model. We do not apply any filter to the CN generated by GPT-3. Furthermore, we associated the target of the offensive comment with the hate speech and counter-narrative pair.

To obtain the CN generated by GPT-3.5, we follow 3 different prompt strategies:

- Exp1: General prompt task definition + 5 examples (1 for each target).
- Exp2: 5 Specific prompt (1 for target) task definition + 3 examples for the same target.
- Exp3: General prompt 5 examples (1 for each target)

Finally, we obtained 238 pairs of hate-speech and counter-narrative among the 3 experiments. All of these pairs are labeled by human annotators in different proposed metrics (Offensiveness, Stance, and Informativeness).

4.1.2. CONAN-MT-SP

CONAN-MT-SP is based in CONAN-MT and an automatic translation is carried out using the API of DeepL to obtain the CONAN-MT-SP (CONAN Multitarget in Spanish) corpus. CONAN-MT consists of 5003 HS-CN pairs covering multiple hate targets (DISABLED, JEWS, LGBT+, MIGRANTS, MUSLIMS, PEOPLE OF COLOR (POC), WOMEN).

Each instance of the CONAN-MT-SP consists of the HS and CN part translated directly into Spanish with DeepL from the CONAN Multitarget corpus, plus the CN generated by GPT-4 using a FSL (Few-Shot Learning) [26] prompt that consists of the task description, 8 examples of HS-CN pairs (one for each target) and the instruction. In addition, evaluations by human experts have also been included as part of the CONAN-MT-SP corpus.

The structure of CONAN-MT-SP is the hate-speech and counternarrative provided by CONAN-MT and the counter-narrative texts generated by GPT-4 model. Furthermore, we associated the values of the different metrics used in the manual evaluation carried by humans. The evaluation metrics are offensiveness, stance, informativeness, truthfulness, editing required, and a comparison between Human-Model.

[2] https://github.com/sinai-uja/CONAN-SP
[3] https://github.com/sinai-uja/CONAN-MT-SP

4.2. Developed systems

As we can see in the previous section, different prompting strategies based on FSL and ZSL (Zero-Shot Learning) [27] have been tested to generate the datasets. The difference between FSL and ZSL is based on whether we include some examples of the task to be solved (FSL) or not (ZSL) in the prompt we provide to the model.

In addition, to prove how the fine-tuning of the model works for automatic generation of counter-relata, the efficient LLM training strategy known as QLoRA [28] has been used to establish a reference model in RefutES, the task we proposed in Iberlef.

However, we want to continue exploring different prompt-based strategies such as those known as Chain-of-Thought [29], Tree-of-Thought or those based on a multi-step prompt in which we provide the different steps to elaborate a good counterfactual.

In addition, we want to test other LLMs fitting strategies such as LOw Memory Optimization (LOMO) [30] to know the differences between both methods or to be able to incorporate external information through a RAG-based system to the generated counter-narratives.

4.3. Participation in Shared Tasks

In order to study the performance of different LLMs and try to understand how the prompt we pass as input affects these models or their adaptation to the task, we have participated in two shared CLEF tasks.

- **Multilingual detoxification (PAN Lab 2024) [31].** The approach presentated is based in the use of LLMs with a prompt Chain of Thought Self-Consistency (CoT-SC) [32] strategy. This CoT-SC strategy consists of identifying the language of the toxic comment and then generating three different detoxified text proposals, the first proposal consists of removing the toxic words, the second of replacing the toxic words with neutral words, and the last of rewriting the toxic text in a neutral way. Subsequently, the selected LLM has to evaluate each generated neutral text according to the competition metrics. Finally, the model selects the best neutral text generated. Specifically with this proposal, we aim to evaluate the capacity of auto-evaluation and reasoning of LLM in different languages, including those with low resources.
- **Oppositional Author Analysis (PAN Lab 2024) [33].** This task is composed of 2 subtasks subtask 1 which consists of a binary classification between critical and conspiracy texts and subtask 2 which consists of a token-level classification of the element of the oppositional narrative. The proposed system for both subtasks consists of the use of LLMs (LLaMA3 or GPT-3.5) where we apply an instruction tuned for the specific subtask. We think that these types of models have more knowledge and can reason to distinguish each type of text or elements of the texts and the instruction tuned will potentiate this, helping the models to distinguish between the classes.

On the other hand, I was part of the organising committee of RefutES, a shared task organized at IberLEF as part of the International Conference of the Spanish Society for Natural Language Processing (SEPLN). The aim of RefutES is to promote the automatic counter-narratives generation in Spanish to reduce the amount of hate speech messages and their effects in social media platforms.

- **RefutES 2024.** We outline a task where participants must be able to generate a response to the offensive message in Spanish. The response should be reasoned, respectful, non-offensive, and contain information that is specific and truthful. For this first edition, the offended targets are: disabled, Jews, LGBT+, migrants, Muslims, people of colour, women and other groups. To establish a baseline benchmark, we performed experiments using ZSL prompt strategy and a fine-tuning of LLaMA2-13B-chat model using QLoRA. In this edition, 6 teams were registered from 4 different countries, but only 1 sent their submission and wrote their working notes.

5. Research Elements Proposed for Discussion

As I am still at the beginning of my research, so there are many issues to address and elements to propose and discuss. Some of them are as follows:

- **Adaptation and performance of LLMs in Spanish.** What adaptation techniques are most effective in improving the accuracy and coherence of texts generated in Spanish? What are the limitations of LLMs in generating argumentative text in Spanish? Is there a way to overcome these problems either by giving them more specific instruction or by training them to be more language literate?
- **Generation of counter-arguments.** What types of messages can be counter-argued? What types of counter-arguments exist? Which ones are valid to apply to responsible AI problems? Can these counter-arguments be oriented to the users who will receive the answer? How can we evaluate the quality and effectiveness of the generated counter-arguments?
- **Access to external information.** In which cases can we resort to external sources of information to elaborate a good counter-argument? How do we extract this information? How can we integrate external information with LLMs to elaborate counter-arguments with accurate information?
- **Responsibility and ethics in text generation.** How can biases present in language models be mitigated when generating text in Spanish, especially in the context of counterarguments? What measures can be implemented to ensure that the counterarguments generated do not perpetuate misinformation or bias? How can it be ensured that language models generate counterarguments that are responsible and ethical?

6. Acknowledgements

My sincere thanks to my thesis tutors Arturo Montejo Ráez and María Teresa Martín Valdivia for guiding me along this process, to the doctoral program of my beloved University of Jaen, and to the Centro de Estudios Avanzados en Tecnologías de la Información y Comunicación (CEATIC) for their support in this research experience. This work has been supported by project CONSENSO (PID2021-122263OB-C21) funded by Plan Nacional I+D+i from the Spanish Government.

References

[1] R. Miller, K. Liu, A. F. Ball, Critical counter-narrative as transformative methodology for educational equity, Review of Research in Education 44 (2020) 269–300. URL: https://doi.org/10.3102/0091732X20908501. doi:10.3102/0091732X20908501. arXiv:https://doi.org/10.3102/0091732X20908501.

[2] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J. W. Rae, L. Sifre, Training compute-optimal large language models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2024.

[3] J. Allen, D. J. Watts, D. G. Rand, Quantifying the impact of misinformation and vaccine-skeptical content on facebook, Science 384 (2024) eadk3451. URL: https://www.science.org/doi/abs/10.1126/science.adk3451. doi:10.1126/science.adk3451. arXiv:https://www.science.org/doi/pdf/10.1126/science.adk3451.

[4] Y. Zhang, K. Sharma, L. Du, Y. Liu, Toward mitigating misinformation and social media manipulation in llm era, in: Companion Proceedings of the ACM on Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1302–1305. URL: https://doi.org/10.1145/3589335.3641256. doi:10.1145/3589335.3641256.

[5] W. Kareem, N. Abbas, Fighting lies with intelligence: Using large language models and chain of thoughts technique to combat fake news, in: M. Bramer, F. Stahl (Eds.), Artificial Intelligence XL, Springer Nature Switzerland, Cham, 2023, pp. 253–258.
[6] M. A. Khaliq, P. Chang, M. Ma, B. Pflugfelder, F. Miletić, Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models, 2024. arXiv:2404.12065.
[7] C. Schieb, M. Preuss, Governing hate speech by means of counterspeech on facebook, 2016.
[8] K. Munger, Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment, Political Behavior 39 (2017) 629–649. URL: https://doi.org/10.1007/s11109-016-9373-5. doi:10.1007/s11109-016-9373-5.
[9] B. Mathew, N. Kumar, Ravina, P. Goyal, A. Mukherjee, Analyzing the hate and counter speech accounts on twitter, 2018. arXiv:1812.02712.
[10] S. Benesch, Countering Dangerous Speech: New Ideas for Genocide Prevention, 2014. URL: https://papers.ssrn.com/abstract=3686876. doi:10.2139/ssrn.3686876.
[11] B. Mathew, P. Saha, H. Tharad, S. Rajgaria, P. Singhania, S. K. Maity, P. Goyal, A. Mukherjee, Thou Shalt Not Hate: Countering Online Hate Speech, Proceedings of the International AAAI Conference on Web and Social Media 13 (2019) 369–380. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/3237. doi:10.1609/icwsm.v13i01.3237.
[12] M. Guerini, Counter-narratives datasets to fight hate speech, 2023. URL: https://github.com/marcoguerini/CONAN, original-date: 2019-05-30T09:48:42Z.
[13] J. Garland, K. Ghazi-Zahedi, J.-G. Young, L. Hébert-Dufresne, M. Galesic, Countering hate on social media: Large scale classification of hate and counter speech, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, 2020, pp. 102–112. URL: https://aclanthology.org/2020.alw-1.13. doi:10.18653/v1/2020.alw-1.13.
[14] Y.-L. Chung, S. S. Tekiroğlu, S. Tonelli, M. Guerini, Empowering NGOs in countering online hate messages, Online Social Networks and Media 24 (2021) 100150. URL: https://doi.org/10.1016%2Fj.osnem.2021.100150. doi:10.1016/j.osnem.2021.100150.
[15] J. Qian, A. Bethke, Y. Liu, E. Belding, W. Y. Wang, A benchmark dataset for learning to intervene in online hate speech, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4755–4764. URL: https://aclanthology.org/D19-1482. doi:10.18653/v1/D19-1482.
[16] Y.-L. Chung, S. S. Tekiroğlu, M. Guerini, Towards knowledge-grounded counter narrative generation for hate speech, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 899–914. URL: https://aclanthology.org/2021.findings-acl.79. doi:10.18653/v1/2021.findings-acl.79.
[17] S. Tekiroglu, H. Bonaldi, M. Fanton, M. Guerini, Using pre-trained language models for producing counter narratives against hate speech: a comparative study, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 3099–3114.
[18] H. Lee, Y. J. Na, H. Song, J. Shin, J. Park, ELF22: A context-based counter trolling dataset to combat Internet trolls, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3530–3541. URL: https://aclanthology.org/2022.lrec-1.378.
[19] M. Ashida, M. Komachi, Towards automatic generation of messages countering online hate speech and microaggressions, in: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 11–23. URL: https://aclanthology.org/2022.woah-1.2. doi:10.18653/v1/2022.woah-1.2.
[20] Y.-L. Chung, S. S. Tekiroğlu, M. Guerini, Italian counter narrative generation to fight online hate speech, in: Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it, 2020.
[21] H. Bonaldi, G. Attanasio, D. Nozza, M. Guerini, Weigh Your Own Words: Improving Hate Speech Counter Narrative Generation via Attention Regularization, in: Y.-L. Chung, H. Bonaldi,

G. Abercrombie, M. Guerini (Eds.), Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA), Association for Computational Linguistics, Prague, Czechia, 2023, pp. 13–28. URL: https://aclanthology.org/2023.cs4oa-1.2.

[22] Y.-L. Chung, G. Abercrombie, F. Enock, J. Bright, V. Rieser, Understanding Counterspeech for Online Harm Mitigation, 2023. URL: http://arxiv.org/abs/2307.04761. doi:10.48550/arXiv.2307.04761, arXiv:2307.04761 [cs].

[23] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL: https://arxiv.org/abs/2005.11401. arXiv:2005.11401.

[24] M. E. Vallecillo-Rodríguez, A. Montejo-Raéz, M. T. Martín-Valdivia, Automatic counter-narrative generation for hate speech in spanish, Procesamiento del Lenguaje Natural 71 (2023) 227–245.

[25] M.-E. Vallecillo-Rodríguez, M.-V. Cantero-Romero, I. Cabrera-De-Castro, A. Montejo-Ráez, M.-T. Martín-Valdivia, CONAN-MT-SP: A Spanish corpus for counternarrative using GPT models, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italy, 2024, pp. 3677–3688. URL: https://aclanthology.org/2024.lrec-main.326.

[26] A. Parnami, M. Lee, Learning from Few Examples: A Summary of Approaches to Few-Shot Learning, 2022. URL: http://arxiv.org/abs/2203.04291. doi:10.48550/arXiv.2203.04291, arXiv:2203.04291 [cs].

[27] W. Wang, V. W. Zheng, H. Yu, C. Miao, A survey of zero-shot learning: Settings, methods, and applications, ACM Trans. Intell. Syst. Technol. 10 (2019). URL: https://doi.org/10.1145/3293318. doi:10.1145/3293318.

[28] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient Finetuning of Quantized LLMs, 2023. URL: http://arxiv.org/abs/2305.14314. doi:10.48550/arXiv.2305.14314, arXiv:2305.14314 [cs].

[29] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Advances in Neural Information Processing Systems 35 (2022) 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[30] K. Lv, Y. Yang, T. Liu, Q. Gao, Q. Guo, X. Qiu, Full parameter fine-tuning for large language models with limited resources, 2024. arXiv:2306.09782.

[31] D. Dementieva, D. Moskovskiy, N. Babakov, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. Ustalov, E. Stakovskii, A. Smirnova, A. Elnagar, A. Mukherjee, A. Panchenko, Overview of the multilingual text detoxification task at pan 2024, in: CEUR Workshop Proceedings, CEUR-WS.org, 2024. URL: https://pan.webis.de/clef24/pan24-web/text-detoxification.html.

[32] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, 2023. URL: https://arxiv.org/abs/2203.11171. arXiv:2203.11171.

[33] D. Korenčić, B. Chulvi, X. B. Casals, M. Taulé, P. Rosso, F. Rangel, Overview of the oppositional thinking analysis pan task at clef 2024, in: G. Faggioli, N. Ferro, P. Galuvakova, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024. URL: https://pan.webis.de/clef24/pan24-web/oppositional-thinking-analysis.html.