

Munich_Z@GermEval Shared Task 2025: When Prompting Is Not Enough: The Limits of Large Language Models in GermEval’s 2025 Harmful Content Detection Task

Florian Ludwig

ZITiS

Zamdorfer Straße 88, 81677 Munich

florian.ludwig@zitis.bund.de

Dr. Stefan Altmann

ZITiS

Zamdorfer Straße 88, 81677 Munich

stefan.altmann@zitis.bund.de

Abstract

In the context of automatic hate speech detection, effective methods for low-data scenarios are crucial due to the dynamic and evolving nature of harmful language, which demands frequent model retraining. Large Language Models (LLMs) are promising for this task because they can generalize tasks solely from task descriptions and thus reduce the need for extensive data annotation. We evaluate prompting strategies for subtask 2 of the *GermEval 2025 Harmful Content Detection* challenge, which involves classifying whether a tweet attacks the free democratic basic order. We show that techniques such as *Chain-of-Thought*, *In-Context Learning* or *Task Decomposition* outperform approaches like *Task Description*. However, a significant performance gap remains between our prompting-based strategies and the top-performing systems in the competition. These findings indicate that, under current conditions, prompting with open-source LLMs alone is insufficient to achieve state-of-the-art results in complex hate speech related classification tasks.

OFFENSIVE CONTENT WARNING: This work contains examples of hateful content. This is strictly for the purpose of enabling this research. Please be aware that this content could be offensive and cause you distress.

1 Introduction

Data annotation is considered a critical bottleneck in many machine learning projects, as it is resource intensive and time-consuming (Neves and Ševa, 2021; Anglin et al., 2022; Madukwe et al., 2020). This limitation is even more demanding in scenarios of low-resource data and of fast-paced dynamic changes, as is the case of hate speech, where the expression and targets of hate can constitute a small proportion of the material to be analyzed and change quickly. Thus, in automatic hate speech

detection, it is crucial to explore methods that perform well with limited data and reduce the need for model retraining (Lupu et al., 2023; Florio et al., 2020). Employing LLMs for this task is appealing, since these models can use their inherent knowledge and linguistic skills to generalize to various tasks based on task descriptions. Hence, these models are potentially able to adapt to different tasks without the need for extensive annotation of data (Ludwig et al., 2025).

In this work, we investigate different zero-shot and few-shot prompting methods to solve the second subtask of the *GermEval 2025* shared task, *Harmful Content Detection in Social Media* (Felser et al., 2025). The goal of this is to classify whether a tweet constitutes an attack on the free democratic basic order. Our contributions are as follows:

- We propose prompting strategies for analyzing different attacks on the basic democratic order
- We analyze the suitability and limitations of these strategies across various Large Language Models

Section 2 presents an overview of related works relevant to the study. Section 3 details the methodology, including data collection, the models applied, prompting strategies used, and the evaluation procedures. In Section 4, the results are provided along with their analysis and discussion. Finally, Section 5 concludes the work by summarizing the key findings and implications.

2 Related Works

The application of LLMs for hate speech detection has been investigated in several studies (Del Arco et al., 2023; Han and Tang, 2022; Chiu et al., 2021). The authors Kumarage et al. (2024) reported strong classification results using *GPT 3.5* and *Llama 2* (Touvron et al., 2023) for the task of recognizing hate speech. In contrast to our work, they applied

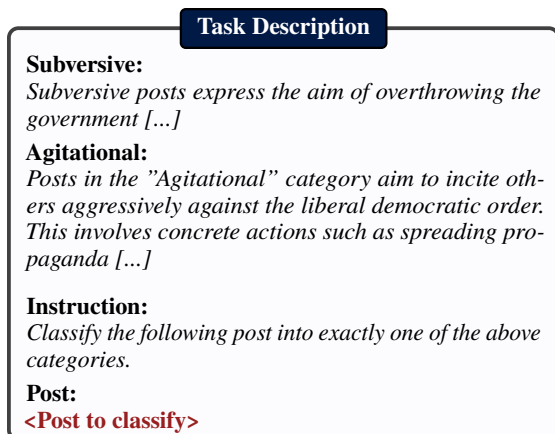


Figure 1: In the **Task Description** approach, we include the class definitions, provided by the *GermEval* challenge organizers, as context in the prompt.

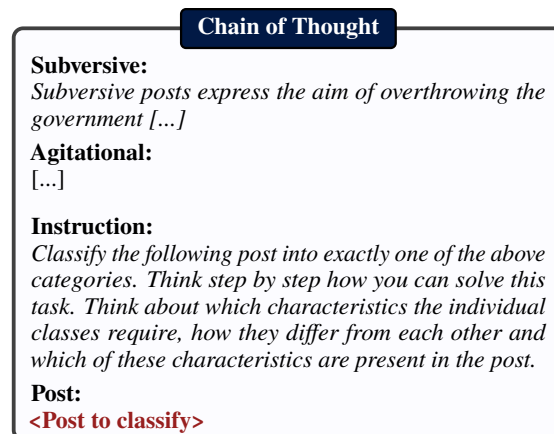


Figure 3: In the **Chain of Thought** approach, the model is guided to reason through how the task can be broken down into subtasks before solving it.

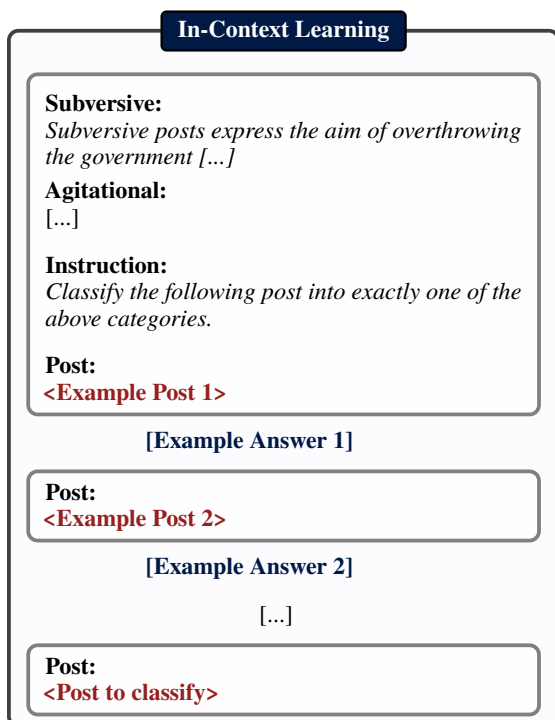


Figure 2: In **In-Context Learning** we provide few-shot examples in addition to the task descriptions in the prompt.

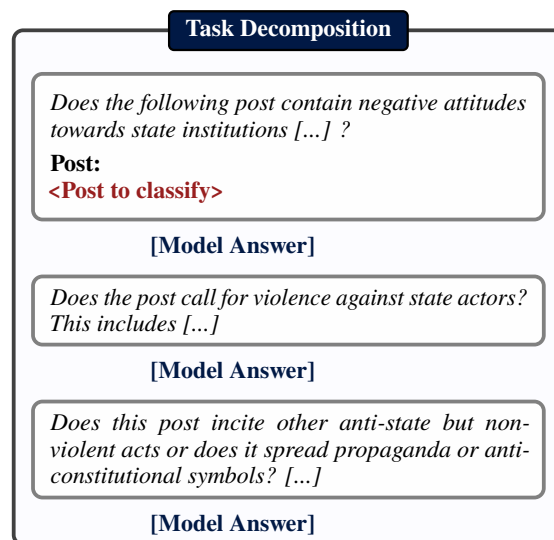


Figure 4: In the **Task Decomposition** approach, we divide the overall problem into subtasks, which are made available to the model as a chain of prompts.

these models to the HateCheck corpus (Röttger et al., 2021), which consists of rather simple, hand-crafted examples. Alkhamissi et al. (2022) investigated how hate speech recognition tasks can be prepared for large language models. The authors propose to split the task into simpler subtasks. They conducted task preparation for fine-tuning a BART model (Lewis et al., 2020). According to the authors the robustness of the classification depends on the prompt, its wording and the context provided. In contrast, our work employed inference-based

approaches, such as *Task Decomposition*, *Chain-of-Thought* prompting and *In-Context Learning*. Morbidoni and Sarra (2023) investigated how GPT 3.5 can be made to recognize misogyny. The authors formulated the task as a simple yes/no question and experimented with additional prompt context.

In our previous work (Ludwig et al., 2025) we investigated how LLMs can be used to detect criminally relevant hate speech. We investigated how these models can be conditioned at different levels of abstraction in legal systems to accomplish this task. A remarkable difference in performance between LLMs and legal experts was reported. While this work addressed how general legal knowledge can be brought into LLMs to recognize criminally-relevant posts, the current work focuses on the sub-

Approach	Model	Pre.	Rec.	F1
Uniform Rand.	<i>Dummy</i>	.29	.29	.29
Description	<i>Phi 4</i>	.43	.39	.41
	<i>Cmd-R+</i>	.34	.30	.31
	<i>Qwen</i>	.48	.43	.41
CoT	<i>Phi 4</i>	.67	.54	.53
	<i>Cmd-R+</i>	.58	.56	.56
	<i>Qwen</i>	.62	.54	.51
In-Context	<i>Phi 4</i>	.59	.52	.50
	<i>Cmd-R+</i>	.59	.57	.57
	<i>Qwen</i>	.67	.56	.55
Decomposition	<i>Phi 4</i>	.58	.50	.49
	<i>Cmd-R+</i>	.40	.44	.44
	<i>Qwen</i>	.56	.51	.49

Table 1: Classification metrics with respect to different approaches and models.

jective definitions of attacks on the free democratic basic order.

3 Methods and Experiments

In this section, we introduce the dataset, language models and prompting strategies, employed in our experiments.

3.1 Data

The training data of the harmful content detection task (Felser et al., 2025) consists of 7,454 German posts, containing different types of attacks on the free democratic basic order. The classes include (1) calling to overthrow the government (*Subversive*), (2) spreading propaganda (*Agitation*), (3) legitimate criticism of the government (*Criticism*) and (4) other posts, without any type of attack on the democratic order (*Nothing*). Data statistics and examples for each class can be seen in Appendix A, Table 5.

3.2 Models

For our experiments, we selected the following models due to their strong reasoning capabilities and state-of-the-art performance on a range of benchmarks, and demonstrated proficiency in multilingual tasks: *Phi-4 14B* (Abdin et al., 2024), known for its efficient scaling and reasoning depth; *Qwen-3 32B* (Yang et al., 2025), which excels in both language understanding and generation across diverse languages; and *Command-R+*¹, a model optimized for instruction following and retrieval-augmented generation.

¹<https://cohere.com/blog/command-r-plus-microsoft-azure>

Approach	Noth.	Crit.	Agit.	Subv.	Avg.
Description	.57	.67	.48	.26	.49
CoT	.60	.77	.56	.27	.55
In-Context	.58	.68	.67	.27	.55
Decomposition	.66	.59	.47	.24	.49
Average	.62	.68	.54	.26	.52

Table 2: Proportion of correctly classified posts by approach and class label, averaged over all models.

3.3 Prompting Strategies

We investigated and compared four different prompting strategies. For the **Task Description** approach (Figure 1 and Appendix B.1), we used the definitions provided by the *GermEval* challenge organizers and incorporated them into the prompt as contextual information. Our second approach (Figure 2, Appendix B.2) focuses on **In-Context Learning**. This is done by extending the first approach with examples selected from the trial set. Example answers to these prompts were formulated by the authors based on the provided class labels. Our third approach is **Chain-of-Thought (CoT)** prompting (Wei et al., 2022). In this approach (Figure 3, Appendix B.3), the model is explicitly encouraged to engage in intermediate reasoning by decomposing a complex task into a sequence of smaller and more manageable subtasks to be solved step by step. *CoT* prompting was shown to enhance the model’s accuracy on reasoning-intensive problems (Wang et al., 2023; Zhang et al., 2025; Yu et al., 2023). The fourth approach is based on **Task Decomposition** (Figure 4), in which the whole classification problem is manually divided into smaller sub-problems. In this case, the answers to the sub-problems are combined to provide the final answer to the complete problem (see Appendix B.4 for more details).

Similar to Ludwig et al. (2025), we used a two-step approach for retrieving the model answers. First, we let the language models freely generate their responses. Second, we prompted the models to format their responses in JSON format, which allowed us to analyze the answers automatically (see Appendix B.5 for further details).

3.4 Model Assessment

A balanced subset of the provided training data, consisting of 50 samples from each class, was extracted and used as a development set to test our prompting strategies. We optimized our prompts using the training set and then evaluated their final

Misclass. Models %	Noth.	Crit.	Agit.	Subv.
80%	.21	.11	.08	.67
100%	.10	.00	.02	.18

Table 3: Proportion of misclassified posts per class, averaged across models.

versions on this separate development set.

4 Results and Discussion

Table 1 presents classification results across various prompting strategies and language models on our development set. Almost all configurations significantly outperform the random baseline. Unlike Ludwig et al. (2025), we found that advanced prompting strategies (*Chain-of-Thought*, *In-Context Learning*, and *Task Decomposition*) outperformed the simpler *Task Description* approach. Notably, *Chain-of-Thought* surpassed manual *Task Decomposition*. No single model consistently outperformed others across all strategies. Our best approach achieved an F1-score of .57 on the development set, substantially lower than the competition winner’s F1-score of .71. This suggests that LLMs cannot yet be reliably or robustly prompted to solve complex classification tasks, such as those related to hate speech detection.

Table 2 reports the proportion of correctly classified samples on the development set across prompting strategies and class labels, averaged over all three models. The models performed worst on the Subversive class, with only 26% of samples correctly classified, which is substantially lower than for the other classes. We further present statistics on posts that were particularly challenging for LLMs to classify (Table 3). The table shows the proportion of posts misclassified by at least 80% of models (first row) or by all models (second row) across all strategies, highlighting that the *Nothing* class and, in line with previous results, especially the *Subversive* class, contained samples most challenging for the models. While 10% of examples of the *Nothing* class were misclassified by all models and approaches, 18% were misclassified for the *Subversive* class.

We manually analyzed these posts and the corresponding language model responses. We observed that 50% of the normal posts, which were misclassified by all models and approaches, contained a call for violent action. These posts were, however, annotated as normal, because the call for violence is not clearly directed towards the government or its offi-

Post	Label
Cut off his stumps and hang him by the feet 🤬🤬🤬 ... far too little, great justice!	Nothing
Carry them? Better to have them flogged under Islamic law. 🕌	Nothing
Someone should shoot him! 🧠	Nothing
[...] Enough of that 🤬 #refugeesnotwelcome [...] #MerkelMustGo	Subversive
Merkel must go, and every asylum seeker who doesn’t behave follow ...	Subversive
Encourage family and friends to vote nationalist in the elections! #MerkelMustGo #PartyBlocMustGo #Islamization	Subversive

Table 4: Example posts, which were misclassified by all models across all prompting schemes.

cials. This indicates that models struggle to grasp the task’s fine-grained nuances, in line with previous findings (Ludwig et al., 2025). As for the posts misclassified as belonging to the *Subversive* class we observed that about 64% of these posts contain a call to dismiss the chancellor or the governing party. However, it is not clear from the posts that this action is to be carried out by force, as required to classify the posts as being subversive. These posts might be considered as borderline cases, which are evaluated differently by language models due to their inherent biases and conceptual understanding. The subjectivity and difficulty of hate speech annotation (Dehghan et al., 2025; Fleisig et al., 2023; Poletto et al., 2019) and inherent biases in language models (Navigli et al., 2023; Liang et al., 2021; Ferrara, 2023) are well-known problems in the research community. Examples of the discussed cases are shown in Table 4.²

5 Conclusion

In this work, we examined various prompting strategies for language models aimed at classifying tweets according to different forms of attacks on the free democratic basic order. Our findings suggest that while LLMs are generally effective in classifying these attacks, their inherent conceptual biases can limit prompt-based control, resulting in suboptimal performances on the second subtask of the *GermEval’s 2025 Harmful Content Detection* (Felser et al., 2025). In contrast, smaller models like BERT (Devlin et al., 2019) may offer greater adaptability through fine-tuning for highly specific tasks. Due to their strong linguistic abilities, LLMs

²These are English translations of posts; original posts were written in German.

are well suited for challenging tasks such as analyzing hate speech - provided that their internalized concepts required for their analyses are properly aligned with the intended goals of the task. Consequently, future work should focus on methods, which are able to achieve language model alignment for hate speech related tasks in a data- and annotation-efficient manner to combine the best of both worlds: strong linguistic capabilities and rapid model adaptation.

Limitations

Our study has some limitations. First, we focus exclusively on subtask 2 of the *GermEval 2025* challenge. While this allows for a more targeted evaluation, it limits the generalizability of our findings to other subtasks or related tasks. Second, we restrict our experiments to three open-source language models. This choice ensures transparency and reproducibility but excludes stronger proprietary models such as *GPT-4*, which could potentially yield better results. Third, we evaluate only a subset of existing prompting strategies. Although we include widely used techniques such as *Chain-of-Thought* and *In-Context Learning*, other recent approaches remain unexplored due to computational and scope constraints. Future work should address these limitations by expanding the range of tasks, models, and prompting methods.

Ethics Statement

We use publicly available, anonymized data from the *GermEval 2025* shared task. While this ensures reproducibility, we acknowledge that the dataset may contain biases and harmful content. Hate speech detection remains a subjective and context-dependent task, and automated systems may produce false positives or negatives. Our models are intended solely for research and should not be deployed in real-world applications without careful consideration of fairness, societal impact, and the need for human oversight.

References

Marah Abidin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. [Phi-4 technical report](#). *arXiv preprint arXiv:2412.08905*.

Badr Alkhamissi, Faisal Ladhak, Srinivasan Iyer, Veselin Stoyanov, Zornitsa Kozareva, Xian Li, Pascale Fung, Lambert Mathias, Asli Celikyilmaz, and Mona Diab. 2022. [Token: Task decomposition and knowledge infusion for few-shot hate speech detection](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2120.

Kylie Anglin, Arielle Boguslav, and Todd Hall. 2022. [Improving the science of annotation for natural language processing: The use of the single-case study for piloting annotation projects](#). *Journal of Data Science*, 20(3).

Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2021. [Detecting hate speech with gpt-3](#). *arXiv preprint arXiv:2103.12407*.

Somaiyeh Dehghan, Mehmet Umut Sen, and Berrin Yanikoglu. 2025. [Dealing with annotator disagreement in hate speech classification](#). *arXiv preprint arXiv:2502.08266*.

Flor Miriam Plaza Del Arco, Debora Nozza, and Dirk Hovy. 2023. [Respectful or toxic? using zero-shot learning with language models to detect hate speech](#). In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Jenny Felser, Michael Spranger, and Melanie Siegel. 2025. Overview of the *germeval 2025* shared task on harmful content detection. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*.

Emilio Ferrara. 2023. [Should chatgpt be biased? challenges and risks of bias in large language models](#). *arXiv preprint arXiv:2304.03738*.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. [When the majority is wrong: Modeling annotator disagreement for subjective tasks](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. [Time of your hate: The challenge of time in hate speech detection on social media](#). *Applied Sciences*, 10(12):4180.

Lawrence Han and Hao Tang. 2022. [Designing of prompts for hate speech recognition with in-context learning](#). In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 319–320. IEEE.

- Tharindu Kumarage, Amrita Bhattacharjee, and Joshua Garland. 2024. [Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection](#). *arXiv preprint arXiv:2403.08035*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. [Towards understanding and mitigating social biases in language models](#). In *International conference on machine learning*, pages 6565–6576. PMLR.
- Florian Ludwig, Torsten Zesch, and Frederike Zufall. 2025. [Conditioning large language models on legal systems? detecting punishable hate speech](#). *arXiv preprint arXiv:2506.03009*.
- Yonatan Lupu, Richard Sear, Nicolas Velásquez, Rhys Leahy, Nicholas Johnson Restrepo, Beth Goldberg, and Neil F Johnson. 2023. [Offline events and online hate](#). *PLoS one*, 18(1):e0278511.
- Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. [In data we trust: A critical analysis of hate speech detection datasets](#). In *Proceedings of the fourth workshop on online abuse and harms*, pages 150–161.
- Christian Morbidoni and Annalina Sarra. 2023. [Can llms assist humans in assessing online misogyny? experiments with gpt-3.5](#). In *GENERALE Workshop @ CHIItaly 2023*.
- Roberto Navigli, Simone Conia, and Björn Ross. 2023. [Biases in large language models: origins, inventory, and discussion](#). *ACM Journal of Data and Information Quality*, 15(2):1–21.
- Mariana Neves and Jurica Ševa. 2021. [An extensive review of tools for manual annotation of documents](#). *Briefings in bioinformatics*, 22(1):146–163.
- Fabio Poletto, Valerio Basile, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2019. [Annotating hate speech: Three schemes at comparison](#). In *CEUR WORKSHOP PROCEEDINGS*, volume 2481, pages 1–8. CEUR-WS.
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. [Hatecheck: Functional tests for hate speech detection models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.
- Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. [Towards understanding chain-of-thought prompting: An empirical study of what matters](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). *Advances in neural information processing systems*, 35:24824–24837.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. [Towards better chain-of-thought prompting strategies: A survey](#). *arXiv preprint arXiv:2310.04959*.
- Yufeng Zhang, Xuepeng Wang, Lingxiang Wu, and Jinqiao Wang. 2025. [Enhancing chain of thought prompting in large language models via reasoning patterns](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25985–25993.

Appendix

A Data

Class	Counts	Original Examples	Translated Examples
Nothing	6,277	Erwache Deutschland 	Awaken, Germany 
		Heil Deutschland !	Hail Germany!
		Es wäre traurig wenn man PEGIDA nur auf die unkontrollierte Zuwanderung reduzieren würde.	It would be sad if PEGIDA were reduced only to uncontrolled immigration.
Criticism	804	Dämonisierung des Gegners ist immer ein notwendiger Teil der Strategie. Jetzt sind wir an der Reihe.	Demonizing the opponent is always a necessary part of the strategy. Now it's our turn.
		wobei die oxen und Esel in Berlin im Bundestag sitzen und zwar alle	where the oxen and donkeys sit in Berlin in the Bundestag, all of them.
		Gar kein Geld sollten die bekommen, wozu auch, die bekommen doch eh schon alles in den Arsch geschoben.	They shouldn't get any money at all, why should they, they already have everything shoved up their ass anyway.
Agitation	313	1488  	1488  
		Heil Hitler! Wir werden uns nicht erweichen lassen! #Vorwärts #SiegHeil #1488	Heil Hitler! We will not be softened! #Forward #SiegHeil #1488
		Blut und Ehre, Hakenkreuz über Europa!	Blood and Honor, swastika over Europe!
Subversive	60	Ihr werdet Frau Merkel stürzen, Die Welt schaut auf euchsuper	You will overthrow Mrs. Merkel, the world is watching you... great
		Kann den nicht einer erschießen - so mit Auftragskiller - am besten vielleicht die komplette Regierung???	Can't someone shoot him – like with a hitman – maybe even the entire government???
		RT @name: @name @news Die EU spaltet Europa. Merkel spaltet die EU. Merkel muss weg. Die EU muss weg. Damit Europa wieder atmen kann.	RT @name: @name @news The EU is dividing Europe. Merkel is dividing the EU. Merkel must go. The EU must go. So that Europe can breathe again.

Table 5: Examples of the provided training data of the *GermEval's 2025 Harmful Content Detection Task*.

B Language Model prompts

B.1 Task Description

Original prompt, written in German.

Task Description

A: „Subversive“

Posts der Kategorie „Subversive“ drücken das Ziel aus, die Regierung durch gewaltsame Handlungen zu stürzen und somit die staatliche Ordnung zu beseitigen. Dies kann in Form von Ankündigungen, Drohungen oder Aufrufen zum Handeln geschehen.

B: „Agitational“

Posts der Kategorie „Agitational“ haben das Ziel, andere Personen in aggressiver Art und Weise gegen die freiheitlich-demokratische Grundordnung aufzuhetzen. Dabei werden konkrete Maßnahmen wie beispielsweise die Verbreitung von Propaganda oder Beschädigung von Staatssymbolen ergriffen.

C: „Critical“

Der Tweet äußert Kritik an der Regierung, an Amtsträgern, Beamten, Behörden oder politischen Parteien, ohne konkrete Absichten gegen die freiheitlich-demokratische Grundordnung anzudeuten.

D: „Nothing“

Posts der Kategorie „Nothing“ richten sich weder gegen die Regierung noch gegen die freiheitlich demokratische Grundordnung.

Instruction:

Ordne den nachfolgenden Post in eine der oben genannte Kategorien (A-D) ein. Entscheide dich immer für genau eine Kategorie. Stelle deine Argumente in kurzer Form dar.

Post:

<Post to classify>

English translation of the original prompt.

Aufgabenbeschreibung

A: „Subversive“

Posts in the "Subversive" category express the goal of overthrowing the government through violent actions and thereby eliminating the state order. This can take the form of announcements, threats, or calls to action.

B: „Agitational“

Posts in the "Agitational" category aim to incite others aggressively against the liberal democratic order. This involves concrete actions such as spreading propaganda or damaging state symbols.

C: „Critical“

The tweet expresses criticism of the government, officials, civil servants, authorities, or political parties without indicating any concrete intentions against the free democratic basic order.

D: „Nothing“

Posts in the "Nothing" category are directed neither against the government nor against the liberal democratic order.

Instruction:

Classify the following post into one of the categories (A–D) mentioned above. Always choose exactly one category. Briefly explain your reasoning.

Post:

<Post to classify>

B.2 In-Context Learning

B.2.1 Example Answers

Original *In-Context Learning* examples and corresponding answers, written in German.

In-Context Learning Example	Manually Crafted Answer
Unsere ehre heisst treue, Kameraden!	Der Post verwendet eine verbotene Phrase (unsere ehre heisst treue) und verbreitet somit verfassungsfeindliche Symbole. Dementsprechend ist der Post als unter der Kategorie Agitational einzuordnen.
Es gibt immer noch eine Alternative, die man wählen kann ...	Der Post drückt Unzufriedenheit mit der Regierung aus und weist darauf hin, eine andere Partei zu wählen. Somit handelt es sich um legale Regierungskritik. Der Post ist dementsprechend unter der Kategorie Critical einzuordnen.
Mal sehen wann wir an der Erderwärmung schuld sind ;)	Der post enthält keine Kommentare, die sich gegen die Regierung wenden. Somit ist dieser Post unter der Kategorie Nothing einzuordnen.
Wir sollten unsere Fahne vor der Regierung schützen und gegen sie kämpfen!	Dieser Post spricht von eine Kampf gegen die Regierung, was einem Aufruf zur gewaltsamen Handlung entspricht. Somit ist der Post als Subversive einzustufen.
Alles für Deutschland! Wir werden die Demokraten vernichten! #Hammerskins88 #HH18	Die Phrase „Alles für Deutschland“ zählt in Deutschland zu verbotener Symbolik. Somit ist dieser Post als Agitational einzustufen.
Danke allen Patrioten!	In diesem Post wird keine Regierungskritik oder negative Einstellung gegen über der demokratischen Grundordnung deutlich. Somit ist der Post der Kategorie Nothing zuzuordnen.
Wenn wir nicht bald handeln, werden wir noch mehr Schaden erleiden. Zeit für einen Aufstand!	Dieser Post ruft zu einem Aufstand gegenüber der Regierung auf. Somit ist der Post der Kategorie Subversive zuzuordnen.
Die schädliche Arbeitsagentur soll mal in den Spiegel schauen.	Der Post drückt Unzufriedenheit gegenüber der Agentur für Arbeit aus. Der Ausdruck wird somit als legitime Kritik verstanden, wodurch der Post der Klasse Critical zugeordnet werden kann.

In-Context Learning examples and corresponding answers, translated into English.

In-Context Learning Example	Manually Crafted Answer
Our honor is loyalty, comrades!	The post uses a banned phrase (“Our honor is loyalty”) and thus disseminates unconstitutional symbols. Accordingly, the post should be classified under the category of agitational content.
There is still an alternative one can vote for...	The post expresses dissatisfaction with the government and suggests voting for another party. Therefore, it constitutes legal government criticism. Accordingly, the post should be classified under the category Critical.
Let’s see when we’ll be blamed for global warming ;)	The post does not contain any comments directed against the government. Therefore, this post should be classified under the category Nothing.
We should protect our flag from the government and fight against them!	This post talks about a fight against the government, which corresponds to a call for violent action. Therefore, the post should be classified as Subversive.
Everything for Germany! We will destroy the Democrats! #Hammerskins88 #HH18	The phrase “Alles für Deutschland” is considered banned symbolism in Germany. Therefore, this post should be classified as Agitational.
Thanks to all patriots!	This post does not express any government criticism or negative attitude towards the democratic constitutional order. Therefore, the post should be assigned to the category Nothing.
If we don’t act soon, we will suffer even more damage. Time for an uprising!	This post calls for an uprising against the government. Therefore, the post should be assigned to the category Subversive.
The harmful employment agency should look in the mirror.	The post expresses dissatisfaction with the Employment Agency. The statement is therefore understood as legitimate criticism, allowing the post to be assigned to the category Critical.

B.2.2 Prompt Design

Original prompt, written in German.

In-Context Learning

A: „Subversive“
Posts der Kategorie „Subversive“ drücken das Ziel aus, die Regierung durch gewaltsame Handlungen zu stürzen und somit die staatliche Ordnung zu beseitigen. Dies kann in Form von Ankündigungen, Drohungen oder Aufrufen zum Handeln geschehen.

B: „Agitational“
Posts der Kategorie „Agitational“ haben das Ziel, andere Personen in aggressiver Art und Weise gegen die freiheitlich-demokratische Grundordnung aufzuhetzen. Dabei werden konkrete Maßnahmen wie beispielsweise die Verbreitung von Propaganda oder Beschädigung von Staatssymbolen ergriffen.

C: „Critical“
Der Tweet äußert Kritik an der Regierung, an Amtsträgern, Beamten, Behörden oder politischen Parteien, ohne konkrete Absichten gegen die freiheitlich-demokratische Grundordnung anzudeuten.

D: „Nothing“
Posts der Kategorie „Nothing“ richten sich weder gegen die Regierung noch gegen die freiheitlich demokratische Grundordnung.

Instruction:
Ordne den nachfolgenden Post in eine der oben genannte Kategorien (A-D) ein. Entscheide dich immer für genau eine Kategorie. Stelle deine Argumente in kurzer Form dar.

Post:
<Example 1>

[Example Answer 1]

Post:
<Example Post 2>

[Example Answer 2]

Post:
<Example Post 3>

[Example Answer 3]

[...]

Post:
<Example Post 7>

[Example Answer 7]

Post:
<Example Post 8>

[Example Answer 8]

Post:
<Post to classify>

English translation of the original prompt.

In-Context Learning

A: „Subversive“
Posts in the "Subversive" category express the goal of overthrowing the government through violent actions and thereby eliminating the state order. This can take the form of announcements, threats, or calls to action.

B: „Agitational“
Posts in the "Agitational" category aim to incite others aggressively against the liberal democratic order. This involves concrete actions such as spreading propaganda or damaging state symbols.

C: „Critical“
The tweet expresses criticism of the government, officials, civil servants, authorities, or political parties without indicating any concrete intentions against the free democratic basic order.

D: „Nothing“
Posts in the "Nothing" category are directed neither against the government nor against the liberal democratic order.

Instruction:
Classify the following post into one of the categories (A–D) mentioned above. Always choose exactly one category. Briefly explain your reasoning.

Post:
<Example 1>

[Example Answer 1]

Post:
<Example Post 2>

[Example Answer 2]

Post:
<Example Post 3>

[Example Answer 3]

[...]

Post:
<Example Post 7>

[Example Answer 7]

Post:
<Example Post 8>

[Example Answer 8]

Post:
<Post to classify>

B.3 Chain of Thought Prompting

Original prompt, written in German.

Chain of Thought

A: „Subversive“

Posts der Kategorie „Subversive“ drücken das Ziel aus, die Regierung durch gewaltsame Handlungen zu stürzen und somit die staatliche Ordnung zu beseitigen. Dies kann in Form von Ankündigungen, Drohungen oder Aufrufen zum Handeln geschehen.

B: „Agitational“

Posts der Kategorie „Agitational“ haben das Ziel, andere Personen in aggressiver Art und Weise gegen die freiheitlich-demokratische Grundordnung aufzuhetzen. Dabei werden konkrete Maßnahmen wie beispielsweise die Verbreitung von Propaganda oder Beschädigung von Staatssymbolen ergriffen.

C: „Critical“

Der Tweet äußert Kritik an der Regierung, an Amtsträgern, Beamten, Behörden oder politischen Parteien, ohne konkrete Absichten gegen die freiheitlich-demokratische Grundordnung anzudeuten.

D: „Nothing“

Posts der Kategorie „Nothing“ richten sich weder gegen die Regierung noch gegen die freiheitlich-demokratische Grundordnung.

Instruction:

Ordne den nachfolgenden Post in eine der oben genannten Kategorien (A-D) ein. Entscheide dich immer für genau eine Kategorie. Denke dabei Schritt für Schritt, wie man diese Aufgabe lösen kann. Denke darüber nach, welche Merkmale die einzelnen Klassen erfordern, wie sie sich voneinander unterscheiden und welche dieser Merkmale im Post vorhanden sind.

Post:

<Post to classify>

English translation of the original prompt.

Chain of Thought

A: „Subversive“

Posts in the "Subversive" category express the goal of overthrowing the government through violent actions and thereby eliminating the state order. This can take the form of announcements, threats, or calls to action.

B: „Agitational“

Posts in the "Agitational" category aim to incite others aggressively against the liberal democratic order. This involves concrete actions such as spreading propaganda or damaging state symbols.

C: „Critical“

The tweet expresses criticism of the government, officials, civil servants, authorities, or political parties without indicating any concrete intentions against the free democratic basic order.

D: „Nothing“

Posts in the "Nothing" category are directed neither against the government nor against the liberal democratic order.

Instruction:

Classify the following post into one of the categories (A–D) mentioned above. Always choose exactly one category. Think through the task step by step. Consider what characteristics each category requires, how they differ from one another, and which of these characteristics are present in the post.

Post:

<Post to classify>

B.4 Task Decomposition

B.4.1 Prompt Design

Original prompt, written in German.

Task Decomposition

Enthält der Beitrag eine negative Einstellung, Handlungen oder Kritik gegenüber staatlichen Institutionen, staatlichen Akteuren oder gegen die demokratische Grundordnung? Antworte kurz mit Ja oder Nein.

Post:
<Post to classify>

[Model Answer 1]

Wird durch den Post zu Gewalt gegen staatliche Akteure oder deren Vertreter aufgerufen? Dies kann beispielsweise die gewaltsame Absetzung von Politikern beinhalten. Antworte kurz mit Ja oder Nein.

[Model Answer 2]

Wird durch diesen Post zu anderen staatsfeindlichen, aber gewaltfreien, Handlungen aufgerufen oder wird durch den Post Propaganda beziehungsweise verfassungsfeindliche Symbole verbreitet? Dazu zählen beispielsweise Hakenkreuze oder bestimmte Aussagen aus der NS-Zeit. Antworte kurz mit Ja oder Nein.

[Model Answer 3]

English translation of the original prompt.

Task Decomposition

Does the post contain a negative attitude, actions, or criticism toward state institutions, state actors, or the democratic constitutional order? Answer briefly with Yes or No.

Post:
<Post to classify>

[Model Answer 1]

Does the post call for violence against state actors or their representatives? This may include, for example, the violent removal of politicians. Answer briefly with Yes or No.

[Model Answer 2]

Does this post call for other anti-government but non-violent actions, or does it spread propaganda or unconstitutional symbols? This includes, for example, swastikas or specific statements from the Nazi era. Answer briefly with Yes or No.

[Model Answer 3]

B.4.2 Label Assignment

Answer 1	Answer 2	Answer 3	Class Label
No	No	No	Nothing
No	No	Yes	Criticism
No	yes	No	Subversive
No	yes	yes	Subversive
yes	No	No	Agitation
yes	No	yes	Agitation
yes	yes	No	Subversive
yes	yes	yes	Subversive

Table 6: Label assignment for the *Task Decomposition* approach for all combinations of model answers.

Post
[...] Genug davon 🙄 #refugeesnotwelcome [...] #MerkelMussWeg
Merkel muss gehen, und jeder Asylbewerber, der sich nicht benimmt, muss folgen ...
Ermutigen Sie Familie und Freunde, bei den Wahlen nationalistisch zu wählen! #MerkelMussWeg #Islamisierung

B.5 Answer Formatting

Original prompt, written in German.

Multi-Step JSON Formatierung

Fasse deine Antwort in JSON-Format zusammen. Nutze dabei folgendes Format:

```
{
  "Antwort": <Klasse>
}
```

Wobei für <Klasse> entweder „Nothing“, „Critical“, „Agitational“ oder „Subversive“ eingesetzt werden können.

English translation of the original prompt.

Multi-Step JSON Formatting

Summarize your answer in JSON format. Use the following structure:

```
{
  "Answer": <Class>
}
```

Whereby either "Nothing", "Critical", "Agitational" or "Subversive" can be used for <Class>.

Multi-Step JSON Formatierung

Fasse deine Antworten in JSON-Format zusammen. Nutze dabei folgendes Format:

```
{
  'negative Einstellung': <Antwort 1>
  'Gewalt': <Antwort 2>
  'andere Handlungen': <Antwort 3>
}
```

Wobei für <Antwort 1>, <Antwort 2> und <Antwort 3> entweder Ja oder Nein eingesetzt werden können.

Multi-Step JSON Formatting

Summarize your answers in JSON format. Use the following structure:

```
{
  'negative attitude': <Answer 1>
  'Violence': <Answer 2>
  'other Actions': <Answer 3>
}
```

Either Yes or No can be used for <Answer 1>, <Answer 2> and <Answer 3>.