mdok-style at SemEval-2026 Task 10: Finetuning LLMs for Conspiracy Detection

Dominik Macko
Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia
dominik.macko@kinit.sk

Abstract

SemEval-2026 Task 10 is focused on conspiracy detection. Specifically, the goal is to detect whether a Reddit comment expresses a conspiracy belief. Our submitted mdok-style system utilizes data augmentation and self-training (to cope with a rather small amount of training data) to finetune the Qwen3-32B model for a binary text-classification task. The submitted system is very competitive, ranking in the **85th percentile** (8th out of 52 submissions). The results shown that our approach, which originated in machine-generated text detection, can be used for conspiracy detection as well.

1 Introduction

The Psycholinguistic Conspiracy Marker Extraction and Detection (PsyCoMark) shared task (SemEval-2026 Task 10) combines psychology and natural language processing disciplines to shed light on how conspiracy theories are expressed in social media (Reddit) conversations (Samory et al., 2026). It includes two subtasks, while we have focused only on **subtask 2** that is dedicated to binary detection of conspiracy belief in a textual comment. Specifically, the goal is to classify English Reddit comments as either conspiracy-related or not (in a Yes/No manner). The training data contains 1,715 positive samples and 2,263 negative samples. The texts have been filtered to a length of 160 to 1,000 characters.

The proposed system is heavily based on mdok (**m**achine **d**etector **o**f **K**InIT) (Macko, 2025), a robust detector of machine-generated text winning a shared task of PAN@CLEF2025 (Bevendorff et al., 2025). Our final submitted system (**mdok-style conspiracy detector**)[1] utilizes data augmentation and self-training to cope with a rather small amount of training data. It is based on the **Qwen3-32B** model (Yang et al., 2025), finetuned using QLoRA parameter-efficient finetuning technique (PEFT) for a binary sequence classification task.

2 Related Works and Background

Conspiracy detection is part of the broader field of misinformation detection. Early approaches framed the task as a text classification problem using traditional machine learning and NLP techniques, relying on lexical and stylistic features with classification models, such as SVM (support vector machines) (Cortes and Vapnik, 1995). Subsequent research introduced psycho-linguistic and emotional features, showing that conspiracy-related narratives exhibit distinct linguistic patterns such as heightened certainty, emotional intensity, and distrust framing. Incorporating such signals improves detection performance beyond surface-level textual representations (Giachanou et al., 2023). More recently, large language models (LLMs) have been applied to conspiracy detection tasks due to their strong contextual understanding. While LLMs improve generalization, they may suffer from hallucination and precision issues. Emotion-aware finetuned LLM frameworks have been proposed to integrate affective signals into transformer-based architectures, achieving improved performance over vanilla LLM baselines (Liu et al., 2024).

For the task of binary conspiracy detection within a given text, we have utilized our experience in a similar task focused on machine-generated text detection. We have transferred our best approach to this kind of a different problem.

We have already started to experiment with finetuning LLMs for a binary classification task at SemEval-2024 Task 8 (Spiegel and Macko, 2024b), where we have observed a better text-classification performance of a finetuned LLM (size of 7B parameters) in comparison to the traditionally used small pre-trained (BERT-like) models. We have further enhanced the robustness of the finetuning

[1] https://github.com/kinit-sk/mdok-style-psycomark2026

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 446–450

process (Macko et al., 2025), which has finally resulted in the mdok (Macko, 2025) finetuning approach. mdok has ranked 1st in both subtasks of PAN@CLEF2025 (Bevendorff et al., 2025), one of which was focused on robust binary text classification (specifically, the detection of machine-generated texts). In this shared task of SemEval-2026 Task 10, we utilize the mdok's obfuscation and anonymization of training data as data augmentation techniques and combine them with self-training to cope with a small amount of labeled data.

Self-training (Amini et al., 2025) is an iterative semi-supervised learning technique, where a model acts as its own teacher to leverage unlabeled data. Firstly, a so-called teacher model is trained on a small amount of data with available "golden" labels (often labeled by humans). The weakly trained model is then used to predict "silver" labels (a.k.a., weak labels or pseudo labels, i.e., they may contain erroneously labeled samples) using any unlabeled data. Since such labels might be inaccurate, usually only the most confident predictions are used to retrain the model. This process can repeat, allowing to continuously expand the training set. However, the errors from the first iteration of silver labels can propagate throughout the learning.

3 System Overview

As mentioned above, the PsyCoMark shared task contains rather small amount of train data (at least for bigger LLMs). To cope with the problem we are proposing a combination of data-augmentation techniques as well as using self-training to incrementally label the provided unlabeled dev and test sets.

3.1 Data Augmentation

In our mdok-style approach for binary conspiracy detection, we use four data-augmentation techniques described below. To use each technique, we have copied the training texts that have been afterwards modified by the corresponding technique. In order to reduce prevalence of augmented texts (in comparison to original texts), we have used only 10% out of each technique to combine into the training. De-duplication of the final train set removes the redundant texts (i.e., unmodified copies).

Anonymization. The orignal PsyCoMark data have already replaced URLs in the texts by a tag of [URL]. Similarly, we have used a text-prepocessing procedure (available in the original mdok) to replace the regex-identified email addresses, user mentions, and phone numbers by the tags of [EMAIL], [USER], [PHONE]. In comparison to the original mdok, we are not using this anonymization procedure for preprocessing all the texts, but for modification of copied data for data augmentation.

Lower-casing and Upper-casing. We consider conspiracy beliefs to be invariant on the casing of the text (especially in social media, often containing informal style). Therefore, to make the detection more focused on meaning rather than on the visual form, we integrate both, lowercase and uppercase copies of the original texts. This augmentation not only increases the number of training samples, but it also makes the detection case insensitive.

Homoglyphication. Homoglyph attacks (swapping visually similar characters of different scripts) are quite successful in confusing text classifiers (dependent on internal representation of characters and words, such as in tokenization). In our previous work on machine-generated text detection (Macko et al., 2024), we have found out that this confusion effect on classifiers can be effectively dealt with by inclusion of homoglyphied samples during training. As an data augmentation, we have used such homoglyphication of copied texts, increasing the train-set size while also making the conspiracy detector more robust.

3.2 Self-Training

In the first round, the selected LLM has been trained purely on the provided train set. Afterwards, it provided predictions of labels of the dev and test sets, while for each prediction we have dumped also a probability of the positive class ("Yes"). Based on such probabilities, we have kept as silver labels only those that had very high probability ($\geq$0.99) as positive and those that had very low probability ($\leq$0.01) as negative (in order to minimize propagation of errors). Such silver-labeled samples have been afterwards combined with the training data for retraining the detector.

3.3 Finetuning Process

Since the provided dev split of the data does not contain labels, we have used a hold-out training data (100 samples per class) for validation during finetuning. Inclusion of 10% of data from each

data-augmentation technique and subsequent deduplication resulted in 2,126 negative and 1,517 positive samples for training. After inclusion of silver-labeled dev and test samples, the training set contained 2,575 negative and 1,881 positive samples.

For finetuning, we have used QLoRA (Dettmers et al., 2023) parameter-efficient finetuning technique (PEFT) (4-bit quantization). As a framework, we have used the transformers[2] python library. We have used paged adamw optimizer with cosine learning rate of 2e-5 and a warmup ratio of 0.03. We have used a batch size of 1 sample without gradient accumulation and validation each 100 steps. The training process has run a single epoch and the final checkpoint selection was based on the best Macro F1 score.

In the footnote of the first page, we have provided a link to Github repository, where we have published the full source code for replication purposes. To install the dependencies, just install the corresponding conda environment of the IMGTB[3] framework (Spiegel and Macko, 2024a) and update the transformers library (for the support of the newest models).

3.4 Base Model Selection

We have exploited the mdok-style efficient finetuning process to train various LLMs, up to the size of 32B parameters. We have focused on Qwen3 (Yang et al., 2025) and Gemma-3 (Team et al., 2025) families. Based on the results in development phase (without the self-training component) using dev-set evaluation by the organizers, we have finally selected Qwen3-32B, as the best performing model for the task.

4 Experimental Setup

We have used only the officially provided data in the PsyCoMark shared task, which have been augmented as described in Section 3.1. We have trained the detectors without self-training and with self-training (using the silver labels predicted by Qwen3-32B). The evaluation is done only using the official shared task Codabench site[4], by the organizers, since the ground truth is not release yet. Upon the release, the error analysis might be executed.

The official metric in subtask 2 is the macro-averaged F1-score (representing a harmonic mean of precision and recall while invariant to class imbalance).

[2] https://github.com/huggingface/transformers
[3] https://github.com/kinit-sk/IMGTB
[4] https://www.codabench.org/competitions/10749

Detector	Macro F1
Qwen3-32B_ST_th0.7	**0.78**
Qwen3-32B_ST	0.77
Qwen3-32B_th0.7	0.77
Qwen3-32B	0.76
DeBERTa-Large	0.75
Qwen3-14B-Base	0.75
Gemma-3-1B-PT_ST	0.75
Qwen3-4B-Base	0.75
Gemma-3-12B-PT	0.74
Gemma-3-1B-PT	0.73
Qwen3-4B-Base_ST	0.72
random baseline	0.50

Table 1: The performance of the various system alternatives using the official test set for the PsyCoMark subtask 2.

5 Results

The results of various system alternatives are provided in Table 1. The results of the compared alternatives are very close to each other. The submitted highlighted system is using self-training (denoted _ST) and classification threshold moved to 0.7 of positive-class probability (denoted _th0.7). However, even the smallest tested DeBERTa model (<0.5B parameters) achieved 0.75 of Macro F1 score, offering a better tradeoff of performance for the cost.

The unofficial ranking (provided currently in the Codabench shared task site, some teams might be disqualified yet if not submitting system-description papers) of the submitted system is provided in Table 2. As illustrated, the mdok-style system ranked competitively, in **the top 20%** of the submissions (85th percentile).

6 Conclusion

Our participation in the shared task provided multiple insights. Our mdok approach, which has been developed for machine-generated text detection, has been successfully transferred to conspiracy detection task. We have compared multiple system alternatives, out of which the submitted system using data augmentation and self-training based on Qwen3-32B offers the best perfomance, ranking competitively among all the submissions. However, tradeoff between detection performance and computation costs must be considered, since the

Rank	Team	Macro F1
1	NJUST_KMG	0.89
2	AGAI	0.87
3	jia57	0.86
4	baishanxiaoqi	0.8
5	CSECU-DSG	0.8
6	joccerrillo	0.79
7	qinchihongye	0.79
8	**mdok-style**	**0.78**
9	dangphuduy	0.78
10	shubham_bits	0.78
11	rziaei	0.77
12	dmarhoef	0.77
13	jorgegomezn	0.77
14	srikarkashyap	0.77
15	shahmir2002	0.76
16	CuriosAI	0.76
17	dengkuihou	0.76
18	joesrwt	0.76
19	Tilde	0.76
20	Macaroni	0.76
21	davidinfotec	0.75
22	psy_detectives	0.75
23	ishaank	0.75
24	panos_span	0.75
25	autist	0.74
26	Macaroni	0.74
27	wangkongqiang	0.74
28	YNU-HPCC	0.74
29	TM	0.74
30	Sarang	0.73
31	hidetsune	0.73
32	CCNU	0.73
33	pc0907	0.73
34	pannndoo	0.73
35	zhangpeng	0.73
36	lamiaa	0.72
37	stangerine	0.72
38	123xxx	0.72
39	777lily	0.72
40	yiyu12	0.72
41	zzz666	0.72
42	bruhh	0.72
43	wagetl	0.72
44	civilwen	0.72
45	xlxxlx	0.72
46	lemontr1	0.72
47	samu721	0.72
48	emmasleghel	0.72
49	hritav_1896	0.71
50	samorymatest	0.71
51	faozia_fariha	0.59
52	bpiper02	0.41

Table 2: The unofficial ranking of the submitted system for the PsyCoMark subtask 2.

DeBERTa model achieved only slightly lower performance.

Limitations

We have explored only small set of base language models. Others could be better performing. Since only the English texts have been included in the shared task, the generalization to other languages is unevaluated. We have limited the set of texts for finetuning only to the official data of the shared task. Other publicly available datasets could be used for training as well.

Acknowledgments

This work was supported by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00068.

Computational resources. We acknowledge EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy.

References

Massih-Reza Amini, Vasilii Feofanov, Loïc Pauletto, Liès Hadjadj, Émilie Devijver, and Yury Maximov. 2025. Self-training: A survey. *Neurocomputing*, 616:128904.

Janek Bevendorff, Yuxia Wang, Jussi Karlgren, Matti Wiegmann, Maik Fröbe, Akim Tsivgun, Jinyan Su, Zhuohan Xie, Mervat T. Abassy, Jonibek Mansurov, Rui Xing, Minh Ngoc Ta, Kareem Ashraf Elozeiri, Tianle Gu, Raj Vardhan Tomar, Jiahui Geng, Ekaterina Artemova, Artem Shelmanov, Nizar Habash, and 5 others. 2025. Overview of the "voight-kampff" generative AI authorship verification task at PAN and ELOQUENT 2025. In *Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR-WS.org*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2023. Detection of conspiracy propagators using psycho-linguistic characteristics. *Journal of Information Science*, 49(1):3–17.

Zhiwei Liu, Boyang Liu, Paul Thompson, Kailai Yang, and Sophia Ananiadou. 2024. Conspemollm: Conspiracy theory detection using an emotion-based large language model. *arXiv preprint arXiv:2403.06765*.

Dominik Macko. 2025. mdok of KInIT: Robustly fine-tuned LLM for binary and multiclass AI-generated text detection. In *Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR-WS.org*.

Dominik Macko, Robert Moro, and Ivan Srba. 2025. Increasing the robustness of the fine-tuned multilingual machine-generated text detectors. *Preprint*, arXiv:2503.15128.

Dominik Macko, Robert Moro, Adaku Uchendu, Ivan Srba, Jason S Lucas, Michiharu Yamashita, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2024. Authorship obfuscation in multilingual machine-generated text detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6348–6368, Miami, Florida, USA. Association for Computational Linguistics.

Mattia Samory, Felix Soldner, and Veronika Batzdorfer. 2026. SemEval-2026 task 10: PsyCoMark – psycholinguistic conspiracy marker extraction and detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Michal Spiegel and Dominik Macko. 2024a. IMGTB: A framework for machine-generated text detection benchmarking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 172–179, Bangkok, Thailand. Association for Computational Linguistics.

Michal Spiegel and Dominik Macko. 2024b. KInIT at SemEval-2024 task 8: Fine-tuned LLMs for multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 558–564, Mexico City, Mexico. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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. *Preprint*, arXiv:2505.09388.

A Computational Resources

For experiments regarding detector training (model finetuning) and inference, we have used $1\times$ NVIDIA A100 64GB GPU, consumed approximately 100 GPU hours. Analysis has been done without the GPU acceleration.