Automatic Detection of Predatory Online Conversations using Small-scale LLMs

Aleksandar Miljković
Faculty of Technical Sciences,
University of Novi Sad,
Novi Sad, Serbia,
aleksandar.miljkovic@mup.gov.rs

Milan Gnjatović
Department of Information Technology,
University of Criminal Investigation
and Police Studies, Belgrade, Serbia
milan.gnjatovic@kpu.edu.rs

Slaviša Ilić
Applied Mathematics and Electronics
Centre, Belgrade, Serbia
ilic.slavisa@gmail.com

Darko Stefanović
Faculty of Technical Sciences,
University of Novi Sad,
Novi Sad, Serbia,
darko.stefanovic@uns.ac.rs

Srđan Sladojević
Faculty of Technical Sciences,
University of Novi Sad,
Novi Sad, Serbia
sladojevic@uns.ac.rs

Zvonimir Ivanović
Department of Criminalistics
University of Criminal Investigation
and Police Studies, Belgrade, Serbia
zvonimir.ivanovic@kpu.edu.rs

Jordan Atanasijević
Applied Mathematics and Electronics
Centre, Belgrade, Serbia
jordan.atanasijevic@vs.rs

Abstract — This paper examines the possibility of applying and fine-tuning small-scale large language models for the task of predatory online conversation classification. The reported study considers a set of pretrained small-scale models (Qwen3-4B-Instruct-2507, Qwen2.5-7B-Instruct-4bit, Phi-4-mini-instruct-4bit, Mistral-7B-Instruct-v0.3, and Llama-3.2-3B-Instruct-4bit). First, the baseline performance of these models in classifying predatory online conversations was evaluated. Then, the considered models were fine-tuned on workstation-level hardware, using the Low-Rank Adaptation method, its quantized variant QLoRA, and the VTPAN chat corpus produced for early sexual predator detection. The results show that fine-tuning made a significant impact on the models' performance on relatively limited, consumer-grade local hardware resources.

Keywords — harassment, llm, fine tuning, classification

I. Introduction

With the ever-growing number of platforms for online communication, the need for protection from malicious users becomes increasingly important. This especially holds true for protecting children from online harassment, where malicious actors apply sophisticated communication methods to create emotional dependence and control over the victim in the pursuit of online grooming [1].

The automatic detection of predatory online conversations faces sevaral challenges. The first challenge is that malicious actors dynamically refine their communication methods and adapt them to overcome safety restrictions set by online communication platforms. Thus, rule-based methods would not be effective [2], [3]. Another major challenge is related to the restricted availability of real-world predatory conversation examples. For both ethical and legal reasons, cases that have confirmed the presence of online grooming are not widely available [2]. This paper examines the possibility of applying and fine-tunning small-scale large language models (LLMs) for the task of predatory online conversation classification.

The paper makes the following contributions. We evaluate the baseline performance of a set of pretrained small-scale LLMs in classifying predatory online conversations: Qwen3-4B-Instruct-2507, Qwen2.5-7B-Instruct-4bit, Phi-4-mini-instruct-4bit, Mistral-7B-Instruct-v0.3, and Llama-3.2-3B-Instruct-4bit. All the considered LLMs were fine-tuned on workstation-level hardware, using the Low-Rank Adaptation (LoRA) method, its quantized variant QLoRA, and the VTPAN chat corpus produced for early sexual predator detection. The obtained results showed that the fine-tuning process significantly improved the predatory conversation classification performance.

II. Related work

Research on the automatic detection of online harassment has evolved from linguistic studies to machine learning, deep learning, and LLMs. Early research established that online grooming is a progressive communicative process of building access, trust, and control [1]. Frameworks like ChatCoder [4] and subsequent analyses [5]—[7] demonstrated that predatory behavior extends beyond explicit sexual vocabulary to subtle relationship formation, underscoring the need for context-aware NLP models.

A significant part of computational research in the field relies on transcriptions collected within the Perverted Justice project, establishing the PAN12 [8] sexual predator identification corpus as a foundational benchmark. To reduce noise, Villatoro-Tello et al. introduced a filtered variant of this corpus [9], later refined by Vogt et al. as VTPAN alongside PANC for early sexual predator detection [10]. Harassment datasets have since expanded to specialized benchmarks like #HackHarassment [11], multilingual corpora such as Bangla Facebook comments [12], institutional reporting datasets [13], and dating violence taxonomies [14].

Early detection relied on manual feature engineering, using multi-step pipelines [9], support vector machines [15], chained classifiers [16], and multi-resolution concept representations [17]. Later work introduced fine-grained taxonomies [18] and ensemble frameworks to address classification ambiguity [19], [20]. Deep learning introduced transformer representations like BERT and MobileBERT [10], data augmentation via backtranslation [21], and turn-level optimization for early detection [22].

Recent efforts evaluated generative LLMs via zero-shot prompting [23], fine-tuning models like Llama 2 [24], or dense embeddings [25]. Active LLM pipelines also explored conversational honeypots [26] and context-aware messaging cascades [27]. Most recently, Xiao applied full-parameter

fine-tuning with a linear classification head to Llama and Qwen on VTPAN [28]. However, full-parameter tuning requires heavy computing infrastructure and departs from pure instruction-following. To our best knowledge, this study is the first to systematically evaluate 4-bit Quantized Low-Rank Adaptation (QLoRA) across multiple LLM families on consumer-grade hardware.

III. Methodologies

This study is supported by the VTPAN dataset [10], a processed derivative of the PAN12 Sexual Predator Identification benchmark. VTPAN retains predator logs originally harvested from Perverted Justice decoy operations and non-predatory logs collected from public IRC channels and Omegle [8]. We retained the predefined training and test partitions of VTPAN, comprising 6,588 conversational segments for training (798 of which were predatory) and 15,329 segments for testing (1,466 of which were predatory). No class balancing or resampling was applied, and all samples were used in their original partitions.

To remove uninformative and noisy conversation segments, VTPAN applied four deterministic filtering criteria to the raw PAN12 corpus [10]:

- Elimination of conversations with only one recorded author.
- Discarding segments containing fewer than six interactions per participant.
- Stripping sequences dominated by special characters or ASCII art.
- Retaining segments generated under PAN12's 25-minute temporal inactivity split (maximum 150 messages per segment).

Applying these rules reduced the raw corpus to 21,918 conversation segments (i.e., 9.9% of the original PAN12 volume), preserving a high-quality subset for model training and evaluation. As illustrated in Fig. 1, conversational exchanges include high degrees of informal slang, typos, incomplete sentences, and short turns, requiring language models capable of context-aware semantic reasoning. An important feature of the VTPAN dataset is its real world representation, where non-predatory conversations significantly outnumber predatory instances.

Person A: hi	Person B: my name is ankit
Person B: hi	Person A: 25
Person A: my name maria wat urs	Person A: u want it
Person B: m	Person B: just for friendship
Person B: u?	Person A: lwts get 2 know
Person B: i m ankit	eachotha first u could b weird
Person A: wat ur name im female	4 all i know
Person B: 19	Person B: for chat on yahoo
Person B: u?	Person B: means
Person A: wat ta	Person B: ?

Fig. 1. *A conversation segment from the VTPAN dataset*

The selection of language models considered in this study covers several widely used model families while comparing models of different capacities, ranging from 3–4B to 7B parameters. This made it possible to examine whether increased model capacity provides an advantage, or whether compact models can achieve comparable results after fine-tuning. The models were also selected from different developers and architectural families in order to reduce the extent to which the results depend on characteristics of a single model family.

The Qwen2.5-7B-Instruct model was selected as one of the larger models in the comparison. It introduces improvements in instruction following and provides multilingual support, which make it relevant for a classification task. The larger capacity of 7B parameters was initially assumed to have more occurrences of individual explicit terms and relationships between multiple messages, so gradual changes in conversational intent, manipulation, solicitation, or other contextual patterns could be better recognized. This model also serves as an important higher-capacity reference point in the experiments [29].

The Mistral-7B-Instruct-v0.3 model represents the second instruction-trained 7B-class model and provides a useful comparison with the Qwen2.5 model while maintaining a similar parameter scale. Including Mistral makes it possible to investigate whether the performance observed for a 7B model is characteristic of the parameter scale itself or is more strongly associated with a particular model family [30].

The Qwen3-4B-Instruct-2507 model was selected as a more compact alternative to Qwen2.5-7B which enables an intra-family comparison that regards size and the release date importance. The Phi-4-mini-instruct and Llama-3.2-3B-Instruct models were included as additional compact language models intended to provide a computationally efficient baseline [31].

The experimental environment consisted of an Apple Mac Book Pro equipped with an Apple M5 CPU, 48 GB of unified memory, Python 3.12 and the MLX-LM version 0.31.3 software environment. MLX-LM is Apple's ML framework for fine-tuning and inference which provides support for computation on both the CPU and integrated GPU.

To adapt the selected instruction-tuned language models to the task of sexual predator detection, parameter-efficient fine-tuning was performed using the Low-Rank Adaptation (LoRA) method [32] and its quantized variant QLoRA [33].

LoRA is a parameter-efficient fine-tuning technique that keeps the original model weights frozen instead of updating all parameters of a pretrained language model. It introduces additional trainable low-rank matrices into selected layers of the network. Consequently, only a comparatively small number of parameters need to be optimized during task-specific adaptation.

When a quantized model is supplied to the LoRA fine-tuning procedure, the resulting training process corresponds to QLoRA-based adaptation. In the original QLoRA formulation, gradients are propagated through a frozen 4-bit quantized language model into the trainable LoRA adapters. QLoRA is appropriate for the considered experiments because selected models are represented by 4-bit quantized MLX checkpoints.

The models considered in this work received an instruction describing the classification task together with the conversation to be classified and were expected to generate one of the predefined class labels. The original conversations were transformed into an instruction-based format suitable for supervised fine-tuning of instruction-tuned language models. The dataset was prepared using the same input-output structure for all models, thus differences in classification could be attributed to the underlying model. Conceptually, every training example followed this structure:

You are a text classification system for online child-safety research. Analyze the supplied online conversation and classify the entire conversation as exactly one of the following labels: PREDATOR or NON_PREDATOR. Return only the label. Do not provide an explanation.

To prevent unnecessary computations, the *--mask-prompt* option was enabled during fine-tuning. The prompt masking

excluded tokens belonging to the input prompt from the loss calculation, so that the loss was calculated only for the target completion. This behavior is especially appropriate for the classification task used in this study. The model does not need to learn to reproduce the conversation or the classification instructions themselves. Instead, optimization should encourage it to generate the correct target class after processing the provided conversation.

The same training configuration was applied across the evaluated models in order to ensure comparability between experiments. The principal QLoRA training parameters are presented in Table I. Training was performed for 1000 iterations with a batch size of one. A relatively small learning rate of $1x10^{-5}$ was used because the objective of parameter-efficient fine-tuning was to gradually adapt the pretrained representations to the classification task, rather than to substantially alter the behavior learned during the pretraining.

TABLE I. QLoRA fine-tuning parameters

Parameter	**Value**
Training iterations (iters)	1000
Batch size (batch-size)	1
Learning rate (learning-rate)	$1x10^{-5}$
Number of adapted layers (num-layers)	16
Reporting interval (steps-per-report)	10
Validation interval (steps-per-eval)	100
Adapter saving interval (save-every)	100
Maximum sequence length (max-seq-length)	8192 tokens
Prompt masking	Enabled
Gradient checkpointing	Enabled

The adaptation was restricted to 16 transformer layers to reduce both the number of trainable parameters and the memory required for the backpropagation. The 8192-token limit provided a compromise between preserving conversational context and maintaining manageable memory requirements. The training statistics were reported every 10 iterations, providing frequent information about the optimization process. The validation was performed every 100 iterations, which allowed for observing the evolution of validation loss throughout the fine-tuning process rather than relying exclusively on the final model state. The learned LoRA adapters were saved every 100 iterations, which retained the possibility to restore individual fine-tuning stages without creating complete copies of the underlying language model. Finally, the gradient checkpointing was enabled via the *--grad-checkpoint* flag to optimize memory efficiency at the small cost of computation time. Gradient checkpointing reduces memory usage by recomputing selected activations during backpropagation instead of storing them.

The combination of the 4-bit quantization, low-rank adaptation, batch size of one, adaptation of 16 layers, and gradient checkpointing made it possible to fine-tune all LLMs within the available unified-memory capacity. At the same time, using the same training configuration across the models provided a controlled experimental environment for comparing their performance.

IV. Results

Table II shows that the best baseline model balanced accuracy result (of 0.6) was achieved by Qwen2.5-7B-Instruct-4bit, followed by Llama-3.2-3B-Instruct-4bit.

It can be observed that the model size is not necessarily an advantage in general, and that different model families behave differently (as seen in the result of Mistral-7B compared to Lllama-3.2-3B). Another evident result is the outperformance of the fine-tuned models compared to their corresponding base model. However, since the underlying VTPAN corpus is highly imbalanced (i.e., the non-predatory conversation class is much more dominantly represented), the accuracy is not necessarily indicative of the models' performance. Therefore, additional and more informative performance measures are reported in Table II: macroaverage precision, recall, and F1-score. Confusion matrices for the considered models are given in Table III.

TABLE II. Performance metrics for 15,329 instances

LLM	Variant	Valid predictions	Balanced accuracy	Macro precision	Macro recall	Macro F1
Qwen3	Baseline	15,329	0.53	0.52	0.53	0.52
	Fine-tuned	15,329	0.98	0.97	0.98	0.98
Qwen2.5	Baseline	15,329	0.60	0.53	0.60	0.46
	Fine-tuned	15,329	0.98	0.97	0.98	0.98
Phi	Baseline	15,329	0.50	0.50	0.50	0.47
	Fine-tuned	15,329	0.98	0.96	0.98	0.97
Mistral	Baseline	15,326	0.49	0.49	0.49	0.48
	Fine-tuned	15,327	0.99	0.99	0.96	0.99
Llama	Baseline	12,794	0.57	0.53	0.57	0.51
	Fine-tuned	15,329	0.98	0.96	0.98	0.97

TABLE III. Confusion summary across LLMs

LLM	Variant	TN	FP	FN	TP
Qwen3	Baseline	12,093	1,770	1,201	265
	Fine-tuned	13,786	77	52	1,414
Qwen2.5	Baseline	7,633	6,230	513	953
	Fine-tuned	13,787	76	38	1,428
Phi	Baseline	9,644	4,219	1,023	443
	Fine-tuned	13,738	125	57	1,409
Mistral	Baseline	11,497	2,363	1,255	211
	Fine-tuned	13,849	12	40	1,426
Llama	Baseline	8,134	3,273	798	589
	Fine-tuned	13,754	109	44	1,422

In addition to the total number of predictions, the number of valid predictions is reported because some models did not always generate one of the two predefined class labels (i.e. PREDATOR and NON_PREDATOR). In such cases, the model output could not be interpreted as a valid classification and was therefore treated as an unknown prediction. Consequently, the total number of predictions represents all evaluated instances, whereas valid predictions include only those instances for which the model produced one of the expected class labels. This distinction is particularly important when comparing models, as classification metrics are calculated only on valid predictions, while the difference between total and valid predictions reflects the model's ability to consistently follow the required classification format.

Finally, Table IV reports the performance of the models specifically for the predator class. Reporting the macroaverage precision, recall, and F1-score value for this class makes it possible to directly assess how reliably each model detected predatory conversations, how many predator instances were missed, and how often non-predatory conversations are incorrectly classified as predatory. This class-specific analysis therefore provides a more informative evaluation of model behavior for the primary detection objective of the study.

TABLE IV. Performance metrics for the PREDATOR class.

LLM	Variant	Precision	Recall	F1
Qwen3	Baseline	0.13	0.18	0.15
	Fine-tuned	0.94	0.96	0.96
Qwen2.5	Baseline	0.13	0.65	0.22
	Fine-tuned	0.94	0.97	0.96
Phi	Baseline	0.09	0.30	0.14
	Fine-tuned	0.92	0.96	0.93
Mistral	Baseline	0.08	0.14	0.10
	Fine-tuned	0.99	0.97	0.98
Llama	Baseline	0.15	0.42	0.22
	Fine-tuned	0.93	0.96	0.94

V. Discussion and Conclusion

The better classification performance of smaller models (i.e., Llama-3.2-3B, Phi-4-mini, Qwen3-4B) relative to the larger Mistral-7B-v0.3 models highlights the impact of architectural and post-training advancements over raw parameter count. Newer compact model families benefit from higher pre-training token density and refined multi-turn

reasoning capacity. These characteristics make them better suited for noisy chat logs while remaining highly adaptable under 4-bit QLoRA fine-tuning. Another conclusion is that fine-tuning made a significant impact on every tested model performance on relatively limited, consumer-grade local hardware resources.

The study has several limitations. First, it was focused to the English-language VTPAN dataset. Generalization to other languages remains untested and constrained by the limited availability of comparable annotated datasets. Second, the considered predatory conversations involve adult decoys posing as children, which may not fully reflect interactions with actual child victims. And third, classification used an entire conversational segment as input, without assessing early detection during an ongoing interaction.

ACKNOWLEDGMENT

This study was partially conducted within the SALVUS (Ensuring SAfer justice outcomes in onLine, including undercoVer, child sexUal abuse inveStigations) project, under the Grant agreement ID: 101225758, and partially within the research project AMEC/01/25-27 of the Applied Mathematics and Electronics Centre.

REFERENCES

[1] L. N. Olson, J. L. Daggs, B. L. Ellevold, and T. Rogers, "Entrapping the innocent: Toward a theory of child sexual predators' luring communication," *Communication Theory*, vol. 17, no. 3, pp. 231–251, Aug. 2007, doi: 10.1111/j.1468-2885.2007.00294.x.

[2] N. Mylonas et al., "Online child grooming detection: Challenges and future directions," in *Paradigms on Technology Development for Security Practitioners*. Cham, Switzerland: Springer, 2025, pp. 237-247, doi: 10.1007/978-3-031-62083-6_19.

[3] P. R. Borj, K. Raja, and P. Bours, "Online grooming detection: A comprehensive survey of child exploitation in chat logs," *Knowledge-Based Systems*, vol. 259, Art. no. 110039, 2023, doi: 10.1016/j.knosys.2022.110039.

[4] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski, "Learning to identify Internet sexual predation," *International Journal of Electronic Commerce*, vol. 15, no. 3, pp. 103–122, 2011, doi: 10.2753/JEC1086-4415150305.

[5] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski, " Learning to identify Internet sexual predation," *International Journal of Electronic Commerce*, vol. 15, no. 3, pp. 103–122, 2011, doi: 10.2753/JEC1086-4415150305.

[6] S. Gupta, R. Kaushal, and L. Nath, "Linguistic analysis of grooming stages in online predator conversations," IEEE Trans. Comput. Social Syst., vol. 5, no. 3, pp. 783–794, Sep. 2018, doi: 10.1109/TCSS.2018.2858169.

[7] P. Black, M. Wollis, M. Woodworth, and J. T. Hancock, "A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world," *Child Abuse & Neglect*, vol. 44, pp. 140–149, 2015, doi: 10.1016/j.chiabu.2014.12.004.

[8] G. Inches and F. Crestani, "Overview of the international sexual predator identification competition at PAN-2012," in CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, Rome, 2012.

[9] E. Villatoro-Tello, A. P. López-Monroy, M. Montes-y-Gómez, and L. Villaseñor-Pineda, "A two-step scheme for detecting sexual predators in chat logs," Inf. Process. Manage., vol. 49, no. 3, pp. 720–731, May 2013.

[10] M. Vogt, U. Leser, and A. Akbik, "Early detection of sexual predators in chats," in *Proc. 59th Annual Meeting of the Association for Computational Linguistics and 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 4985–4999, doi: 10.18653/v1/2021.acl-long.386.

[11] A. Bastidas, E. Dixon, C. Loo, and J. Ryan, "Harassment detection: A benchmark on the #HackHarassment dataset," arXiv preprint arXiv:1609.02809, 2016.

[12] M. F. Ahmed, Z. Mahmud, Z. T. Biash, A. A. N. Ryen, A. Hossain, and F. B. Ashraf, "Bangla text dataset and exploratory analysis for online harassment detection," arXiv preprint arXiv:2102.02478, 2021.

[13] R. Divya and V. G. Ranga Nath, "Sexual harassment complaints across BSE100 companies: A longitudinal dataset (2013–2023)," BMC Res. Notes, vol. 18, no. 1, p. 227, doi: 10.1186/s13104-025-07294-0.

[14] N. Satomi and O. Keita, "Classification of forms of dating violence-harassment and relationship of characteristics of the perpetrator-victim," *The Proceedings of the Annual Convention of the Japanese Psychological Association*, vol. 79, Art. no. 2EV-053, 2015.

[15] C. Morris and G. Hirst, "Identifying sexual predators by SVM classification with lexical and behavioral features - Notebook for PAN at CLEF 2012," in *CLEF 2012 Evaluation Labs and Workshop-Working Notes Papers*, Rome, Italy, 2012.

[16] H. J. Escalante, M. Montes-y-Gómez, and A. P. López-Monroy, "Sexual predator detection in chats with chained classifiers," *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 46-54.

[17] A. P. López Monroy, F. A. González, M. Montes, H. J. Escalante, and T. Solorio, "Early text classification using multi-resolution concept representations," in Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 1216–1225, doi: 10.18653/v1/N18-1110.

[18] S. Sharifirad and S. Matwin, "When a tweet is actually sexist: A more comprehensive classification of different online harassment categories and the challenges in NLP," arXiv preprint arXiv:1902.10584, 2019.

[19] S. Abarna, J. I. Sheeba, and S. Pradeep, " A novel ensemble model for identification and classification of cyber harassment on social media platform " J. Intell. Fuzzy Syst., vol. 45, no. 4, pp. 6125–6138, Oct. 2023, doi: 10.3233/JIFS-230346.

[20] N. A. Azeez and E. Fadhal, "Classification of virtual harassment on social networks using ensemble learning techniques," *Applied Sciences*, vol. 13, no. 7, Art. no. 4570, 2023.

[21] Ł. Radliński, M. Guściora, and J. Kocoń, "Backtranslation and paraphrasing in the LLM era? Comparing data augmentation methods for emotion classification," in *Computational Science* - ICCS 2025, Lecture Notes in Computer Science, vol. 15903. Cham, Switzerland: Springer, 2025, pp. 3-17, doi: 10.1007/978-3-031-97626-1_1.

[22] J. An, S. Ryu, H. Do, Y. Kim, J. Ok, and G. Lee, "Revisiting early detection of sexual predators via turn-level optimization," in Proc. NAACL-HLT 2025 (Volume 1: Long Papers), Albuquerque, NM, USA, Apr. 2025, pp. 4713–4724.

[23] C. Wu, B. Ma, Z. Zhang, N. Deng, Y. He, and Y. Xue, "Evaluating zero-shot multilingual aspect-based sentiment analysis with large language models," *International Journal of Machine Learning and Cybernetics*, vol. 16, pp. 8079–8101, 2025.

[24] T. T. Nguyen, C. Wilson, and J. Dalins, "Fine-tuning Llama 2 large language models for detecting online sexual predatory chats and abusive texts," in *Proc. European Symp. Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN),* Bruges, Belgium, Oct. 2024, pp. 613–618.

[25] A. S. Susmitha and P. Pujari, "Sentiment analysis of cyberbullying data in social media," arXiv preprint arXiv:2411.05958, 2024.

[26] Y. Yao, B. Wang, J. Duan, K. Xu, C. Guo, Z. E. Sun, and Y. Zhang, "The imitation game: Using large language models as chatbots to combat chat-based cybercrimes," arXiv preprint arXiv:2512.21371, 2025.

[27] P. Lu, N. Ishfaq, E. Win, M. Rose, S. R. Strickland, C. L. Biernesser, J. Zelazny, and M. De Choudhury, "Context-aware detection and victim-centered response generation for online harassment in private messaging," arXiv preprint arXiv:2512.14700, 2025.

[28] Y. Xiao, "Adapting generative LLMs for high-accuracy online grooming detection," in *Proc. 6th Int. Conf. Intelligent Computing and Human-Computer Interaction* (ICHCI), 2025, pp. 206–209, doi: 10.1109/ICHCI67614.2025.11291960.

[29] Qwen et al., "Qwen2.5 technical report," 2024, arXiv:2412.15115, doi: 10.48550/arXiv.2412.15115.

[30] A. Q. Jiang et al., "Mistral 7B," 2023, arXiv:2310.06825, doi: 10.48550/arXiv.2310.06825.

[31] A. Yang et al., "Qwen3 technical report," 2025, arXiv:2505.09388, doi: 10.48550/arXiv.2505.09388.

[32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent*. (ICLR), 2022.

[33] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, pp. 10088–10115, 2023.