Empowering Persuasion Detection in Slavic Texts through Two-Stage Generative Reasoning

Xin Zou, Chuhan Wang, Dailin Li, Yanan Wang, Jian Wang, Hongfei Lin
Dalian University of Technology
{zouxin, wangchuhan, lidailin, wangyanan}@mail.dlut.edu.cn
{wangjian, hflin}@dlut.edu.cn

Abstract

This paper presents our submission to Subtask 2 (multi-label classification of persuasion techniques) of the Shared Task on Detection and Classification of Persuasion Techniques in Slavic Languages at SlavNLP 2025. Our method leverages a teacher–student framework based on large language models (LLMs): a Qwen3 32B teacher model generates natural language explanations for annotated persuasion techniques, and a Qwen2.5 32B student model is fine-tuned to replicate both the teacher's rationales and the final label predictions. We train our models on the official shared task dataset, supplemented by annotated resources from SemEval 2023 Task 3 and CLEF 2024 Task 3 covering English, Russian, and Polish to improve cross-lingual robustness. Our final system ranks 4th on BG, SI, and HR, and 5th on PL in terms of micro-F1 score among all participating teams.

1 Introduction

Persuasion techniques (Piskorski et al., 2023a) are widely employed in both formal and informal discourse, ranging from parliamentary debates to emotionally charged social media posts. These techniques leverage rhetorical devices—such as exaggeration, scapegoating, or appeals to authority—to manipulate opinions or obscure critical thinking (Nikolaidis et al., 2023). Automatically identifying such techniques is an essential step in combating misinformation and promoting media literacy.

The SlavNLP 2025 Shared Task on "Detection and Classification of Persuasion Techniques in Slavic Languages" presents a challenging multilingual, multi-label classification problem (Piskorski et al., 2025). In Subtask 2, participants are asked to identify all applicable persuasion techniques in paragraph-level texts drawn from two distinct domains—political debates and social media—and across five Slavic languages (Bulgarian, Croatian, Polish, Slovene, and Russian).

LLMs (Brown et al., 2020) have shown impressive reasoning abilities, producing detailed reasoning steps that enhance input prompts and boost few-shot or zero-shot performance (Wei et al., 2022; Kojima et al., 2022). Reasoning steps have also been utilized during further fine-tuning to enable LLMs to self-improve (Zelikman et al., 2022; Huang et al., 2022). LLMs have shown remarkable performance in generation tasks but have been less widely applied to classification tasks. Therefore, we leverage generative reasoning to enhance the classification capabilities of large models.

To address this task, we propose an explanation-guided teacher–student training framework (Hinton et al., 2015; Pintrich and Schunk, 1996) that first uses the Qwen3 32B model (Yang et al., 2025) as a teacher to generate detailed natural language rationales explaining the presence of persuasion techniques, providing an intermediate layer of supervision beyond surface annotations. Then, a student model Qwen2.5 32B (Yang et al., 2024) is fine-tuned to mimic the teacher's reasoning patterns and labels. Additionally, we apply a voting strategy (Breiman, 1994) leveraging the stochasticity of the fine-tuned model by generating multiple prediction samples and aggregating results via self-consistency (Wang et al., 2022) voting.

Our exploration integrating explanation-guided rationale generation, cross-lingual data augmentation, and ensemble voting mechanisms suggests potential pathways for addressing the challenges of multilingual persuasion technique classification.

2 Methodology

2.1 Stage 1: Abductive Reasoning from the Teacher Model

To enable the student model to acquire interpretability and contextual reasoning abilities in persuasion technique detection, we adopt a large language model (LLM), specifically **Qwen3-32B**, as

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 177–182
July 31, 2025 ©2025 Association for Computational Linguistics

the teacher model. Leveraging its strong generalization and causal reasoning capabilities, we activate its latent knowledge through carefully designed prompts, encouraging it to generate natural language rationales for multi-label persuasion decisions. These rationales are then used to guide the student model via distillation.

Given a sample text input X (e.g., a parliamentary debate speech or a social media post) and its corresponding multi-label annotation Y, we prompt the teacher model to produce a rationale R that explains why the given text implies the presence of one or more persuasion techniques. This prompt is designed to elicit rich world knowledge and argumentative reasoning from the LLM. Formally, the prompt p is structured as:

> *The following [language] sentence [text] employs persuasion techniques: [labels]. Please explain the reasons why.*

Through this prompting strategy, the LLM generates a rationale R, which often includes background knowledge, discourse clues, and inferred intentions that are implicit in the input text. These rationales are then paired with the original input X to form training samples (X, R) for the student model in the next stage.

This stage serves as a form of **interpretation-level knowledge distillation**, where the student model learns not only to predict but also to reason. The resulting rationale corpus provides fine-grained supervision that guides the student model to capture semantic patterns aligned with persuasion techniques, enhancing both accuracy and explainability in downstream classification tasks.

2.2 Stage 2: Qwen2.5 32B Fine-tuning

To perform multi-label classification of persuasion techniques on parliamentary debates and social media texts, we employ a two-phase fine-tuning strategy on a smaller language model, Qwen2.5-32B, using only textual modality. This design ensures that the model inherits reasoning capabilities while maintaining inference efficiency in real-world scenarios.

Learn from Rationale: In the first stage of training, we supervise the student model using the reasoning texts (*rationales*) previously generated by the teacher model. These rationales provide explicit explanations for why a given input contains one or more persuasion techniques, serving as valuable intermediate supervision signals.

We adopt a sequence-to-sequence learning objective: the student model takes the original text input X and generates the corresponding rationale $\hat{R}$, aiming to approximate the target rationale R. This learning process encourages the model to internalize latent reasoning patterns aligned with persuasion techniques, promoting better understanding of argumentative structures and discourse cues embedded in persuasive language.

We use a prompt-based format to elicit rationales in a natural language generation setting. Our prompt is:

> *Identify persuasion techniques used in the text and please explain the reasons why. Your answer should be a subset of the following labels: [all the labels].*

The learning objective in this stage is:

$$\mathcal{L}_{\text{rationale}} = \text{CE}(R, \hat{R}) \tag{1}$$

Learn from Label: We further fine-tune the student model to directly predict persuasion technique labels in a multi-label classification setting. Instead of relying on fixed-size classification heads, we cast this task as a generation problem. The model is prompted to generate a list of persuasion techniques from a predefined label set.

Given the same text input X, the model is trained to generate one or more applicable labels L, where each label name is separated by commas. The target output consists of all the gold labels concatenated into a natural-language-like string. This generation-based formulation allows the model to flexibly output an arbitrary number of labels without manual threshold tuning or token-level classification constraints. In this stage, we change the prompt to:

> *Identify persuasion techniques used in the text. Your answer should be a subset of the following labels: [all the labels].*

The learning objective is:

$$\mathcal{L}_{\text{label}} = \text{CE}(L, \hat{L}), \tag{2}$$

Together, these two stages ensure that the student model not only inherits the interpretive capability of the teacher model but also becomes proficient in direct label inference. The reasoning stage enhances the model's internal comprehension, while the label prediction stage adapts this understanding to the downstream multi-label task.

2.3 Self-Consistency Voting

Despite the promising capabilities of the fine-tuned student model, the open-ended nature of rationale generation and the inherent ambiguity in multi-label persuasion classification can occasionally introduce variability in the predicted labels. Such variation may stem from factors like decoder sampling stochasticity or subtle shifts in the model's attention.

To improve prediction stability and reduce uncertainty, we employ self-consistency at inference time. Concretely, for each input text X, we sample the model multiple times independently to obtain a collection of predicted label sets $\{Y_1, Y_2, ..., Y_n\}$. For each candidate label, we count its frequency across these runs and include it in the final prediction if it appears in more than half of them. The final aggregated label set Y^* is defined as:

$$Y^* = \{y \mid \sum_{i=1}^{n} I(y \in Y_i) > \frac{n}{2}\} \qquad (3)$$

This self-consistency approach helps mitigate inconsistencies across individual predictions and promotes more reliable output in the multi-label setting. By aggregating multiple decoding outcomes, it reinforces stable and representative label assignments while suppressing occasional noise.

3 Experiment and Result

3.1 Dataset and Evaluation

We constructed the evaluation set by randomly sampling 50 instances each from the organizer-provided RU, PL, and BG datasets. For SI, given its limited data availability, we selected 20 instances. The remaining data constituted our training set. Additionally, we curated supplementary training data from the RU, PL, and EN portions of SemEval 2023 Task 3 (Piskorski et al., 2023b) and CLEF 2024 Task 3 (Piskorski et al., 2024). Complete dataset statistics are presented in Table 1 and 2.

The experiments use Macro-F1 and Micro-F1 as the main evaluation metrics to measure model performance. Macro-F1 reflects performance across classes, while Micro-F1 considers the overall label distribution. This is important in our multi-label setting, where the number of samples per persuasion technique varies greatly.

Source	Language	Samples
CLEF 24	English (EN)	8,826
	Russian (RU)	3,940
	Polish (PL)	3,730
SemEval 23	English (EN)	7,520
	Russian (RU)	1,555
	Polish (PL)	1,655
Slavic 25	Polish (PL)	145
	Bulgarian (BG)	118
	Russian (RU)	116
	Slovenian (SI)	38
Total	—	**23,693**

Table 1: Training set statistics

Source	Language	Samples
Slavic 25	Polish (PL)	50
	Bulgarian (BG)	50
	Russian (RU)	50
	Slovenian (SI)	20
Total	—	**170**

Table 2: Validation set statistics

3.2 Experimental Setup

We utilize Qwen3-32B for rationale generation and fine-tune Qwen2.5-32B for reasoning, with training deployed on 4 NVIDIA A100 GPUs. The detailed parameter configurations are presented in Table 3

3.3 Results and Analysis

Table 4 shows our final results on test dataset. Polish achieves the highest scores due to supplementary training data beyond the competition dataset. In contrast, Bulgarian, Croatian, and Slovenian demonstrate substantially lower performance with limited training samples from the competition data alone. The consistently lower Macro-F1 compared to Micro-F1 across all languages further indicates class imbalance challenges.

Since the classification task is reformulated as a generative framework, careful temperature selection becomes essential for controlling output diversity. Figure 1 illustrates how Micro-F1 and Macro-F1 scores on the validation set vary across different temperature settings, analyzing the impact of generation temperature on model performance. After comprehensive evaluation of this trade-off,

Parameter	Value
LoRA rank	8
LoRA target layers	All
Batch size	8
Learning rate	1.0×10^{-4}
Learning schedule	Cosine + 10% warmup
Training epochs	6
Checkpoint interval	25 steps
n(voting)	3

Table 3: Training Configuration Summary

Language	Micro-F1	Macro-F1
BG	0.2796	0.1504
HR	0.2968	0.1776
PL	0.3557	0.1958
SI	0.1911	0.1128

Table 4: Results on test dataset

we select temperature 0.9 as the optimal configuration.

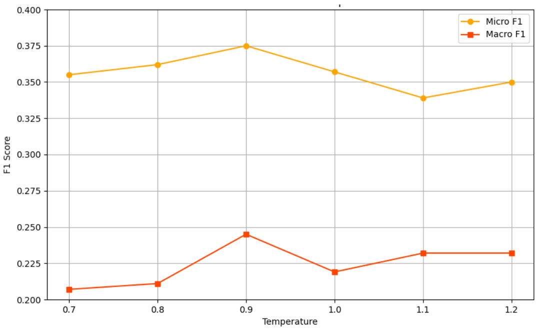

Figure 1: F1 scores under different temperature settings

To assess the impact of rationale-guided training and compare performance with LLM inference, we report results on the validation set in Table 5. The "1 stage" setting refers to directly fine-tuning the student model with label supervision only. The "2 stages" setting first fine-tunes the model using rationales generated by a 32B LLM, followed by label-based fine-tuning. Additionally, we include zero-shot predictions from both the 32B and 72B LLMs for reference.

The results show that 2-stage training yields the best performance, suggesting that rationale supervision helps the model better capture subtle persuasive cues. Moreover, both fine-tuning approaches significantly outperform direct inference from even larger LLMs (72B), underscoring the effectiveness of task-specific training over sheer model size.

Method	Micro-F1	Macro-F1
32B zero-shot	0.2527	0.1997
72B zero-shot	0.2799	0.2130
1 stage	0.3575	0.2418
2 stages	0.3757	0.2446

Table 5: Comparison of fine-tuning strategies and zero-shot LLM inference

Text: *The Withdrawal Agreement abrogates this fundamental contract and would place control of aspects of our national security in foreign hands.*

1 stage: Loaded Language, Appeal to Fear-Prejudice

2 stages: Appeal to Fear-Prejudice, Doubt, Flag Waving

True Label: Appeal to Fear-Prejudice, Doubt, Flag Waving

As shown in the preceding case, the 1-stage and 2-stage approaches yield distinct predictions. Notably, the model failed to detect the Doubt technique implied by the phrase "abrogates this fundamental contract", due to the absence of Stage 1 training for implicit cues, limiting its recognition of non-interrogative skepticism. It also missed Flag-Waving as a separate technique, interpreting the nationalist tone in "our national security" solely as Appeal to Fear, reflecting lexical overattribution. Finally, the model misclassified the phrase as Loaded Language alone, overlooking its role in reinforcing both Appeal to Fear-Prejudice and Flag-Waving, showing a pattern of overgeneralizing emotional cues while underrepresenting nationalist appeals.

To evaluate the effectiveness of the self-consistency voting strategy, we submitted two sets of results to the official evaluation platform. Specifically, run 1 corresponds to the prediction results using a single forward pass without voting, while run 2 applies hard voting across multiple inference outputs. The results are shown in Figure 2.

As shown in the evaluation results, the voting strategy did not consistently improve performance across all languages. In terms of Micro-F1, the differences between run 1 and run 2 are marginal or slightly negative, suggesting that the voting strategy does not bring notable gains in overall prediction accuracy. However, in some cases (Polish

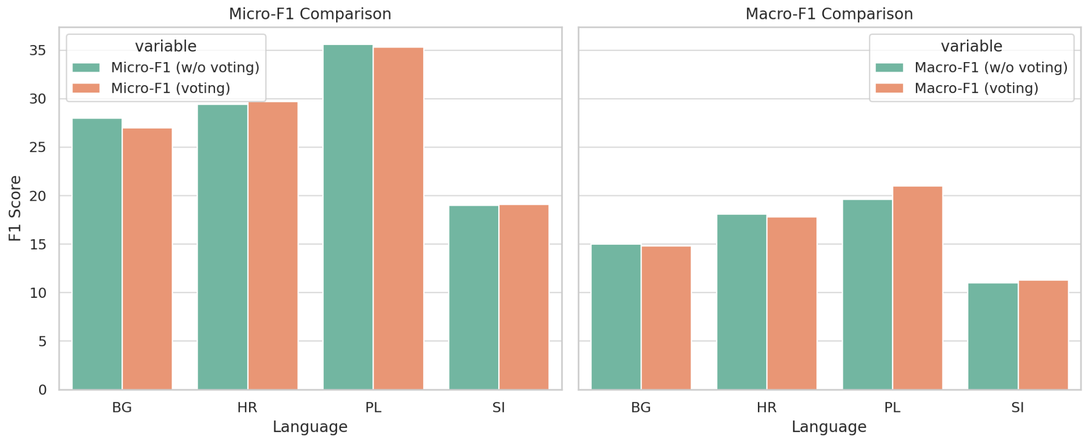

Figure 2: Ablation experiment of voting

and Slovenian), we observe minor improvements in Macro-F1, indicating that voting may help capture more diverse labels and improve robustness on underrepresented classes.

Label	Prec	Rec	F1
Appeal_to_Authority	22.22	40.00	28.57
Appeal_to_Fear-Prejudice	28.00	72.92	40.46
Appeal_to_Hypocrisy	24.81	32.32	28.07
Appeal_to_Pity	0.00	0.00	0.00
Appeal_to_Popularity	20.00	13.16	15.87
Appeal_to_Time	0.00	0.00	0.00
Appeal_to_Values	15.63	21.28	18.02
Causal_Oversimplification	20.00	4.17	6.90
Consequential_Oversimplification	100.00	2.00	3.92
Conversation_Killer	0.00	0.00	0.00
Doubt	40.35	58.82	47.86
Exaggeration-Minimisation	24.73	23.96	24.34
False_Dilemma-No_Choice	14.29	1.35	2.47
False_Equivalence	0.00	0.00	0.00
Flag_Waving	18.59	46.77	26.61
Guilt_by_Association	0.00	0.00	0.00
Loaded_Language	27.58	77.89	40.74
Name_Calling-Labeling	33.79	43.81	38.15
Obfuscation-Vagueness-Confusion	0.00	0.00	0.00
Questioning_the_Reputation	32.26	12.93	18.46
Red_Herring	0.00	0.00	0.00
Repetition	11.48	44.68	18.26
Slogans	8.77	31.25	13.70
Straw_Man	0.00	0.00	0.00
Whataboutism	50.00	1.85	3.57

Table 6: Per-label classification performance on BG, reported as percentages (%).

Overall, although self-consistency voting does not yield significant improvement in this task, it offers a simple and generalizable approach to slightly enhance performance, particularly in multi-label classification with varying label distributions.

In Table 6, we can see model performance varies widely across different persuasion techniques. Frequent labels like Loaded_Language, Name_Calling-Labeling, and Doubt exhibit relatively strong recall and F1 scores, reflecting the benefit of ample training examples. In contrast, low-resource labels such as False_Equivalence, Appeal_to_Pity, and Straw_Man receive near-zero performance, underscoring the model's limitations under few-shot or zero-shot conditions. Notably, some low-frequency classes like Whataboutism show modest precision, suggesting that certain well-defined rhetorical patterns may still be captured despite data sparsity.

4 Conclusion

This paper presents our approach to the Slavic NLP 2025 Workshop, focusing on multi-label persuasion technique classification in parliamentary debates and social media texts. We adopt a two-stage framework: a teacher model first generates contextual rationales via prompt-based reasoning, which guide fine-tuning of a student model. The student is optimized with both rationale and label supervision. To improve prediction robustness, we perform self-consistency voting over multiple decoding runs to produce the final label set. However, the current prompt template assumes input instances contain at least one persuasion technique, which aligns with the training data distribution of Subtask 2 and other additional datasets (where empty-label instances are absent), but fails to account for the test distribution containing non-persuasive content.

In future work, we plan to explore more effective prompting strategies for handling non-persuasive content, investigate architectures that integrate persuasion detection modules, and develop calibration techniques to enhance robustness for multi-label classification in open-domain scenarios.

References

Leo Breiman. 1994. Bagging predictors. Technical Report TR-460, University of California, Berkeley, CA, USA.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Computing Research Repository*, arXiv:1503.02531.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Nikolaos Nikolaidis, Nicolas Stefanovitch, and Jakub Piskorski. 2023. On experiments of detecting persuasion techniques in polish and russian online news: Preliminary study. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing (SlavicNLP 2023)*, Dubrovnik, Croatia. Association for Computational Linguistics.

Paul R. Pintrich and Dale H. Schunk. 1996. *Motivation in Education: Theory, Research, and Applications*. Merrill, Englewood Cliffs, NJ.

J. Piskorski, N. Stefanovitch, V-A Bausier, N. Faggiani, J. Linge, S. Kharazi, N. Nikolaidis, G. Teodori, B. De Longueville, B. Doherty, J. Gonin, C. Ignat, B. Kotseva, E. Mantica, L. Marcaletti, E. Rossi, A. Spadaro, M. Verile, G. Da San Martino, F. Alam, and P. Nakov. 2023a. News categorization, framing and persuasion techniques: Annotation guidelines. Technical Report JRC132862, European Commission, Ispra.

Jakub Piskorski, Dimitar Dimitrov, Filip Dobrani’c, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubeši’c, Michał Marci’nczuk, Arkadiusz Modzelewski, Ivo Moravski, and Roman Yangarber. 2025. SlavicNLP 2025 Shared Task: Detection and Classification of Persuasion Techniques in Parliamentary Debates and Social Media. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing 2025 (SlavicNLP 2025)*, Vienna, Austria. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Firoj Alam, Ricardo Campos, Dimitar Dimitrov, Alípio Mário Jorge, Senja Pollak, Nikolay Ribin, Zoran Fijavz, Maram Hasanain, Purificação Silvano, Elisa Sartori, Nuno Guimarães, Ana Zwitter Vitez, Ana Filipa Pacheco, Ivan Koychev, Nana Yu, Preslav Nakov, and Giovanni Da San Martino. 2024. Overview of the clef-2024 checkthat! lab task 3 on persuasion techniques. In *Conference and Labs of the Evaluation Forum*.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023b. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multilingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. Technical report, Alibaba Group. ArXiv preprint arXiv:2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. Technical report, Alibaba Group. ArXiv preprint arXiv:2412.15115.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.