

Breaking the Illusion of Reasoning in Polish LLMs: Quality over Quantity of Thought

Dzmitry Pihulski, Mikołaj Langner, Jan Eliaasz,
Przemysław Kazienko, Jan Kocoń, Teddy Ferdinan

Wrocław Tech, 50-370 Wrocław, Poland

{dzmitry.pihulski, mikolaj.langner, jan.eliasz,
przemyslaw.kazienko, jan.kocon, teddy.ferdinan}@pwr.edu.pl

Abstract

Recent advances in large language models (LLMs) have introduced explicit reasoning capabilities, yet the factors that truly drive their improved performance remain unclear. In this work, we disentangle the effects of reasoning quality and sequence length by fine-tuning 8B models on several Polish variants of the Mixture-of-Thoughts (MoT-PL) dataset, each representing a distinct reasoning style: *Detailed*, *Summarized*, *BabyThink*, *Lengthy*. We found that the model trained on high-quality reasoning traces achieved better average performance than all other models; neither *longer reasoning with similar quality* nor *low-quality reasoning with similar length* achieved similar gains. Qualitative and quantitative analyses further reveal that reasoning clarity, rather than verbosity, is the dominant factor driving model performance. These findings underscore the importance of reasoning content quality in LLM training and provide new insights into designing more effective reasoning-oriented datasets and models.

1 Introduction

Large Language Models (LLMs) have recently been augmented with explicit reasoning capabilities, enabling them to generate intermediate explanations before producing a final answer. Techniques such as Chain of Thought (CoT) prompting (Wei et al., 2022; Kojima et al., 2022) and the development of dedicated reasoning models (OpenAI, 2025; Guo et al., 2025; Anthropic, 2025) have led to significant improvements across tasks requiring logical inference (Pihulski and Kocoń, 2025; Langner et al., 2025). These approaches often rely on extended reasoning traces, which are hypothesized to enhance accuracy by decomposing complex problems and allowing self-correction. Despite this progress, the precise role of reasoning in improving model performance remains unclear (Shojaee et al., 2025; Lawsen, 2025). To

address these uncertainties, we investigate what truly matters in LLM reasoning. Specifically, we disentangle the effects of reasoning content quality and sequence length using Polish-language datasets derived from Mixture-of-Thoughts, and evaluate their impact on LLM performance across domains.

The following research questions guided the design and analysis of our experiments:

- RQ1:** Do LLMs need reasoning tokens to achieve better performance?
- RQ2:** Does reasoning improve LLMs’ performance because of the content quality of the reasoning part, or because of the artificially prolonged computation?
- RQ3:** How much performance gain can we achieve from improving the content quality of the reasoning part alone, while maintaining a similar sequence length, when fine-tuning an LLM?

The contributions presented in this work include:

- (1) The first work to our knowledge that decouples *reasoning quality* from *reasoning length* to analyze its influence on model performance;
- (2) The empirical evidence showing *reasoning quality* as the most important factor influencing model performance, and how exactly *reasoning length* may boost or degrade model performance;
- (3) A Polish reasoning dataset to support the development of Polish-centric and multilingual reasoning LLMs.

1.1 Related Work

The role and nature of generated *reasoning traces* remain actively debated, with studies offering divergent interpretations. Shojaee et al. (Shojaee et al., 2025) argued that reasoning-augmented models outperform standard LLMs only on limited task types, exhibiting both catastrophic failures on complex problems and “overthinking” behavior on simpler ones. However, Lawsen (Lawsen, 2025) later

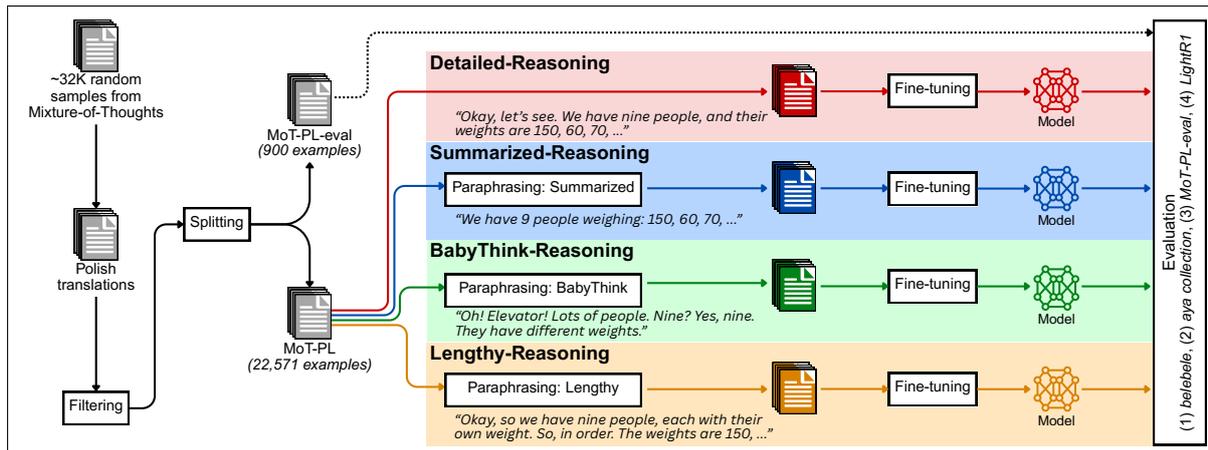


Figure 1: We translated a subset of Mixture-of-Thoughts into Polish, which was split into a training set (MoT-PL) and evaluation set (MoT-PL-eval). Then, we created three additional variants of MoT-PL by paraphrasing only the reasoning part of each example: the *Summarized* style makes the reasoning much more concise, the *BabyThink* style greatly simplifies the reasoning, and the *Lengthy* style prolongs the reasoning. Afterwards, we fine-tuned Llama-PLLuM-8B-instruct on these datasets separately and evaluated the resulting models.

challenged these conclusions, pointing out methodological flaws in the original evaluation and questioning its experimental design.

Recent work has also examined the relationship between reasoning length and task accuracy. Jin et al. (Jin et al., 2024) claimed that extending the reasoning length regardless of the actual quality of the reasoning would boost model performance. On the other hand, Wu et al. (Wu et al., 2025) demonstrated an inverted U-shaped relationship between Chain-of-Thought (CoT) length and performance, suggesting an optimal reasoning span that balances task decomposition with error accumulation. Other studies (Chua and Evans, 2025; Chen et al., 2025; Kambhampati et al., 2025) highlight that LLM reasoning traces often diverge from genuine logical inference: models frequently omit explicit premises, and the correctness of reasoning steps does not necessarily correlate with answer accuracy. Collectively, these findings caution against equating reasoning traces with human-like thought and motivate further investigation into which aspects of reasoning truly drive LLM performance.

2 Polish Mixture-of-Thoughts

Mixture-of-Thoughts is a dataset containing 350K English reasoning examples from math, programming, and science domains distilled from DeepSeek-R1. It was used in the Open R1 project (HuggingFace, 2025; Lozhkov et al., 2025; Penedo et al., 2025; Bercovich et al., 2025). We randomly selected ~ 32 K examples to be translated

into the Polish language utilizing DeepSeek-V3-0324 (DeepSeek-AI, 2024).

Initially, we performed manual inspection on some outputs to refine the prompts and identify potential issues. We translated the user message, the reasoning part of the assistant message, and the final answer of the assistant message of every example separately to fit within the LLM’s supported context length. We identified potential issues, such as: (1) Failed translation where the LLM output is still in English instead of Polish; (2) Failed translation where the model repeats an example given in the prompt instead of translating the target text; (3) Artefacts from the prompt formatting, such as repeated delimiters. After identifying these issues, we programmatically filtered out examples that might contain errors by using: (1) the Python port¹ of the langdetect library (Nakatani, 2010) to ensure that the LLM output is in Polish; (2) the BERTScore metric (Zhang et al., 2020) to ensure semantic alignment between user message, reasoning part, and final answer, eliminating cases of repeated examples; (3) a list of blacklisted strings created from the identified artefacts, so that every occurrence of a blacklisted string can be simply removed from the translation. Similar processes were also conducted to ensure the quality of the paraphrased datasets.

The training set, **MoT-PL**, contains 22,571 examples in total, consisting of 6,320 (28%) math examples, 3,837 (17%) programming examples,

¹<https://github.com/Mimino666/langdetect>

and 12,414 (55%) science examples. Additionally, we also set apart 900 held-out examples for evaluation (**MoT-PL-eval**); there is no overlap between the training set and the evaluation set.

3 Experimental Setup

3.1 Datasets

We prepared four variants of MoT-PL: *Detailed*, *Summarized*, *BabyThink*, and *Lengthy* (see Figure 1). All variants contain identical questions and final answers; they differ only in the reasoning parts. The difference lies in *reasoning quality* and *reasoning length*. Reasoning quality is defined by the level of details relevant for answering a given user prompt, while reasoning length is defined by the number of tokens in the sequence. Table 1 describes the differences between these four datasets in these aspects. Modification of the reasoning parts was performed by automatic paraphrasing using DeepSeek-V3-0324. Appendix E provides the prompts used for the paraphrasing, along with samples from the datasets.

The **Detailed** dataset is MoT-PL as described in Section 2 without any modification. The reasoning parts are equivalent to the original English examples, and we consider it as high-quality reasoning. On average, each example’s reasoning part contains $\sim 2,523$ tokens.

The **Summarized** dataset was created by paraphrasing the original Polish reasoning into a more compact and concise form while still preserving the essential details. The average length of the reasoning part of the examples is ~ 444 tokens.

The **BabyThink** dataset was created by paraphrasing the original Polish reasoning into a “child-like” style, removing fine details and transforming the text into a simplistic form, similar to how a baby talks to an adult. Each example’s reasoning part contains $\sim 2,661$ tokens on average, roughly similar to that of the Detailed dataset.

The **Lengthy** dataset was designed to preserve the overall meaning and detailed reasoning structure of the original Polish text while paraphrasing it to be at least twice as long. The average length of the reasoning part of the examples in this dataset is $\sim 4,884$ tokens.

To validate our claim about the levels of reasoning quality in these datasets, we selected 1,000 random samples and conducted pairwise ranking using LLM-as-a-judge. This allowed us to measure their reasoning quality relative to each

Table 1: Reasoning quality and reasoning length of each dataset variant.

Variant	Quality	Length
Detailed	high	original
Summarized	high	shorter than original
Lengthy	high	longer than original
BabyThink	low	similar to original

Table 2: Average pairwise comparison scores of Detailed, Summarized, and Lengthy reasoning variants when compared against BabyThink.

Variant	Judge-Qwen	Judge-Llama
Detailed	0.999	0.985
Summarized	0.995	0.997
Lengthy	0.998	0.984

other. We repeated the process twice with different models as the LLM judge: firstly using Qwen3-235B-A22B-Instruct-2507-FP8 (Qwen Team, 2025) (Judge-Qwen), and secondly using Llama-3.3-70B-Instruct (Grattafiori et al., 2024) (Judge-Llama), to minimize the risk of bias. For every sample, the judge was given the user prompt, the reasoning part of Assistant A, and the reasoning part of Assistant B. The judge was asked to determine which of the two reasoning variants contained more useful information for answering the user prompt. The judge output a verdict by choosing Assistant A, Assistant B, or Tie, accompanied by a brief explanation. Finally, the scoring for each reasoning style was done by assigning 1 for a win, or 0 for a loss. In the case of a tie, both reasoning styles in the pair were assigned 1. Invalid outputs were not included in calculating the final score.

Table 2 presents the average pairwise comparison scores of Detailed, Summarized, and Lengthy reasoning variants when compared against BabyThink. Each of the three aforementioned variants achieved an average score that is very close to 1 from both LLM judges, strongly indicating that the reasoning quality is superior compared to the BabyThink version. Appendix B provides more details on the pairwise comparison procedure and additional analysis.

3.2 Model

We used **Llama-PLLuM-8B-instruct** (Kocouř et al., 2025; Pezik et al., 2025) as the starting point

of our experiments. This model is based on Llama-3.1-8B (Grattafiori et al., 2024), created by continual pre-training on Polish-centric corpus, followed by instruction fine-tuning with Polish examples. Llama-PLLuM-8B-instruct has not been trained for explicit reasoning. Before our fine-tuning, we added `<think>` and `</think>` to the tokenizer as special tokens, and extended the model’s token embeddings accordingly. Technical details related to our fine-tuning are provided in Appendix C.

3.3 Evaluation

We evaluated our models using four datasets:

- **Belebele** (Bandarkar et al., 2024), a multilingual multiple-choice reading comprehension benchmark consisting of context passages, questions, and four answer choices, with one correct answer;
- **Aya Collection** (Singh et al., 2024), a multilingual instruction-following benchmark covering various task types, including question answering, classification, and summarization;
- **MoT-PL-eval**, the held-out evaluation set of our Polish translation of Mixture-of-Thoughts previously described in Section 2, used to measure reasoning performance in the target domains and language;
- **LightR1** (Wen et al., 2025), a benchmark composed of challenging reasoning tasks requiring multi-step problem-solving in English.

For each dataset, we sampled 900 examples from the respective test split using a stratified sampling strategy to ensure balanced coverage across task type and, where applicable, the length of the reference reasoning. All generations were produced using a fixed set of decoding parameters (see Table 8) to ensure consistency. Each generation was allowed to continue until the end-of-sequence (EOS) token was generated.

We used the *LLM-as-a-judge* paradigm to evaluate our models, similar to the evaluation methodology introduced in the Open R1 project. We deployed DeepSeek-R1-0528 to label each response as either correct or incorrect. Finally, accuracy was computed as the percentage of correct responses over the evaluation set. Details related to the prompting technique and validation of the LLM-as-a-Judge are given in the Appendix D.

4 Experimental Results

Table 3 compares the performance of the fine-tuned versions of Llama-PLLuM-8B-instruct with the original model across all evaluation datasets. Fine-tuning on the Detailed dataset led to the largest gain on MoT-PL-eval with 12.1 percentage points (pp.) improvement. It also gave some improvements on Polish (10 pp.) and English (11.8 pp.) Belebele, as well as the English subset of Aya Collection (4 pp.). Meanwhile, the models fine-tuned on Summarized (6.5 pp. gain), BabyThink (3.6 pp. gain), and Lengthy (3.6 pp. gain) datasets were better than the original model on MoT-PL-eval. Yet, their gains were noticeably lower compared to the Detailed dataset. Table 4 further breaks down the results from MoT-PL-eval by domain; the Detailed model achieved the highest scores on Math and Science, while the original model achieved the highest on Code, followed by Lengthy and Detailed.

5 Discussion

RQ1. Explicit reasoning can lead to better performance if the training data possess sufficiently good quality and an ideal level of verbosity. The Detailed model was fine-tuned on the original Polish translations of DeepSeek-R1’s reasoning traces, and it achieved the highest average score. On the contrary, the Summarized, BabyThink, and Lengthy models were also fine-tuned for explicit reasoning, yet their average scores were lower than the original model. This indicates the characteristics of the training data are important to achieve the performance boost.

We also analyzed why the original model outperformed the fine-tuned models in coding. After analyzing several sample outputs, we found that the original model does not always generate executable code. Instead, it frequently provides a natural-language description of how the code should function to satisfy the task constraints. The LLM-as-a-judge often accepts such descriptive explanations as valid solutions if they conceptually fulfill the task requirements. In contrast, the fine-tuned models tend to generate actual code snippets derived from its reasoning process. Yet, they sometimes overlook given constraints or hallucinate additional ones not specified in the task. This behavior can be explained by the fact that the examples in the Code category account for only about 17% of the total data, which we believe is too small to achieve any meaningful performance in coding.

Table 3: Experimental results on *belebele*, *aya collection*, *MoT-PL-eval*, and *LightR1*.

Model	belebele-PL	belebele-EN	aya-PL	aya-EN	MoT-PL-eval	LightR1	Avg.
(original)	0.609	0.656	0.656	0.552	0.316	0.172	0.494
Detailed	0.709	0.774	0.584	0.592	0.437	0.146	0.540
BabyThink	0.516	0.609	0.463	0.331	0.352	0.079	0.392
Lengthy	0.526	0.593	0.257	0.187	0.352	0.135	0.342
Summarized	0.500	0.568	0.377	0.248	0.381	0.076	0.358

Table 4: Breakdown of the experimental results on MoT-PL-eval per category.

Model	MATH	CODE	SCIENCE
(original)	0.103	0.138	0.479
Detailed	0.145	0.054	0.695
BabyThink	0.082	0.037	0.589
Lengthy	0.085	0.056	0.576
Summarized	0.102	0.031	0.635

RQ2. The main factor that enables explicit reasoning to improve model performance is the quality of the reasoning content. This is evidenced by the Detailed model achieving the best average score when fine-tuned on high-quality data, while BabyThink – similar in length but lower in quality – performed worse, showing quality outweighs length.

The Summarized model, trained on overly concise data that removed back-stepping, failed to learn self-reflection, reducing accuracy. The Lengthy model, trained on similarly high-quality but longer traces, also underperformed. Hence, **reasoning length matters not for longer computation, but for enabling self-reflection.**

To further study the link between reasoning behavior and performance, we analyzed reasoning length by model and correctness using Welch’s *t*-test (see Appendix E.1). Correct responses consistently contained fewer reasoning tokens, showing that longer reasoning is not necessarily beneficial and that the structure and relevance of reasoning matter more than its length.

RQ3. The Detailed and BabyThink models were fine-tuned for similar reasoning lengths. Their results showed that optimizing the reasoning quality alone while keeping a similar reasoning length can lead to 14.8 pp. gain on average. This shows that carefully crafting the structure and details of reasoning traces is crucial to improve model performance.

Observed performance degradation on some particular tasks. While the Detailed model achieved the best overall performance, its scores are slightly lower compared to the base non-reasoning model on the Polish subset of aya collection (0.584 vs 0.656) and the English reasoning task LightR1 (0.146 vs 0.172). This can be attributed to the fact that we only used MoT-PL as our training data. As a result, the model lost some degree of capability on non-reasoning tasks and non-Polish tasks. In the continual learning domain, this phenomenon is known as *catastrophic forgetting* (Ratcliff, 1990; McCloskey and Cohen, 1989). When subjected to learning new information, the model forgets previously learned knowledge. To avoid this problem in a real-world scenario, one way is to mix some samples that represent the model’s existing knowledge into the training data during fine-tuning; this approach is known as rehearsal or replay (Shi et al., 2025). Alternative training techniques (Kirkpatrick et al., 2017; Ke et al., 2021; Ferdinan et al., 2024; Wei Jie et al., 2024) that carefully update the model’s parameters to minimize catastrophic forgetting have also been proposed. However, such considerations are outside the scope of our work, which focuses on the impact of reasoning quality and reasoning length.

6 Conclusion

Our experiments demonstrate how quality and length of reasoning influence model performance. We found reasoning quality to be the most important factor in boosting accuracy across multilingual benchmarks and reasoning-focused tasks. Longer reasoning does not lead to performance gain, highlighting that effective reasoning is more than just additional computation. These findings suggest that training data for explicit reasoning must be well-structured and relevant to help the model generalize and solve complex problems effectively.

Acknowledgments

AI-based tools, including ChatGPT, Grammarly Premium, and Writeful, were used exclusively to support linguistic clarity and improve the readability of the manuscript.

This work was supported in part by: (1) the National Science Center, Poland, project no. 2021/41/B/ST6/04471; (2) CLARIN-PL: Common Language Resources and Technology Infrastructure (POIR.04.02.00-00C002/19, 2024/WK/01, FENG.02.04-IP.040004/24); (3) Digital Research Infrastructure for the Arts and Humanities DARIAH-PL (POIR.04.02.00-00-D006/20, KPOD.01.18-IW.03-0013/23); (4) the statutory funds of the Department of Artificial Intelligence, Wroclaw Tech; (5) the EU under the Horizon Europe, grant no. 101086321 (OMINO). The views expressed are those of the authors and do not necessarily reflect those of the EU or the European Research Executive Agency.

Ethical Considerations

We used all third-party artifacts (datasets, models, and evaluation tools) in accordance with their published licenses and stated intended uses; our translations and paraphrased derivatives of Mixture-of-Thoughts, as well as all evaluations, were conducted strictly for research purposes. Any artifacts we release (MoT-PL and its variants) are intended solely for research on LLM reasoning and must not be used in production or non-research contexts, thereby preserving the original access conditions.

Limitations

In this work, we isolated the effects of *reasoning quality* from *reasoning length* on model performance by fine-tuning Llama-PLLuM-8B-instruct on different variants of Polish Mixture-of-Thoughts. This study can be further extended by involving more LLMs of varying sizes in order to strengthen the generalizability of the findings. We also note that the reasoning quality in our experiments might be limited by the automatic translation process. Finally, our study was limited to the Polish language; the generalizability of our findings in other languages and in multi-lingual context could be addressed in future work.

References

- Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. <https://www.anthropic.com/news/claude-3-7-sonnet>. [Online; accessed 23-Sep-2025].
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. *The belebele benchmark: a parallel reading comprehension dataset in 122 language variants*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, et al. 2025. *Llama-nemotron: Efficient reasoning models*.
- Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. 2025. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*.
- James Chua and Owain Evans. 2025. Are deepseek r1 and other reasoning models more faithful? In *ICLR 2025 Workshop on Foundation Models in the Wild*.
- DeepSeek-AI. 2024. *Deepseek-v3 technical report*.
- Teddy Ferdinan, Jan Kocon, and Przemyslaw Kazienko. 2024. *Into the Unknown: Self-Learning Large Language Models*. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 423–432, Los Alamitos, CA, USA. IEEE Computer Society.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. *The llama 3 herd of models*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- HuggingFace. 2025. *Open r1: A fully open reproduction of deepseek-r1*.
- Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. *The impact of reasoning step length on large language models*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1830–1842, Bangkok, Thailand. Association for Computational Linguistics.

- Subbarao Kambhampati, Kaya Stechly, Karthik Valmeekam, Lucas Saldyt, Siddhant Bhambri, Vardhan Palod, Atharva Gundawar, Soumya Rani Samineni, Durgesh Kalwar, and Upasana Biswas. 2025. Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! *arXiv preprint arXiv:2504.09762*.
- Zixuan Ke, Bing Liu, Nianzu Ma, Hu Xu, and Lei Shu. 2021. Achieving forgetting prevention and knowledge transfer in continual learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 22443–22456. Curran Associates, Inc.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.
- Jan Kocoń, Maciej Piasecki, Arkadiusz Janz, Teddy Ferdinan, Łukasz Radliński, Bartłomiej Koptyra, Marcin Oleksy, Stanisław Woźniak, Paweł Walkowiak, Konrad Wojtasik, et al. 2025. PLLuM: A Family of Polish Large Language Models. *arXiv preprint arXiv:2511.03823*.
- 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.
- Mikołaj Langner, Jan Elias, Ewa Rudnicka, and Jan Kocoń. 2025. Divide, cache, conquer: Dichotomic prompting for efficient multi-label llm-based classification.
- A. Lawsen. 2025. Comment on the illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity.
- Anton Lozhkov, Hynek Kydlíček, Loubna Ben Allal, Guilherme Penedo, Edward Beeching, Quentin Gallouédec, Nathan Habib, Lewis Tunstall, and Leandro von Werra. 2025. Openr1-math-220k. <https://huggingface.co/datasets/open-r1/OpenR1-Math-220k>.
- Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press.
- Shuyo Nakatani. 2010. [Language detection library for java](#).
- OpenAI. 2025. Introducing OpenAI o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini>. [Online; accessed 23-Sep-2025].
- Guilherme Penedo, Anton Lozhkov, Hynek Kydlíček, Loubna Ben Allal, Edward Beeching, Agustín Piñeres Lajarín, Quentin Gallouédec, Nathan Habib, Lewis Tunstall, and Leandro von Werra. 2025. Codeforces cots. <https://huggingface.co/datasets/open-r1/codeforces-cots>.
- Piotr Pęzik, Filip Żarnecki, Konrad Kaczyński, Anna Cichosz, Zuzanna Deckert, Monika Garnys, Izabela Grabarczyk, Wojciech Janowski, Sylwia Karasińska, Aleksandra Kujawiak, et al. 2025. The pllm instruction corpus. *arXiv preprint arXiv:2511.17161*.
- Dzmitry Pihulski and Jan Kocoń. 2025. [Language, culture, and ideology: Personalizing offensiveness detection in political tweets with reasoning llms](#).
- Qwen Team. 2025. [Qwen3 technical report](#).
- Roger Ratcliff. 1990. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97(2):285–308.
- Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2025. [Continual learning of large language models: A comprehensive survey](#). *ACM Comput. Surv.*, 58(5).
- Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. [The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity](#).
- Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. [Aya dataset: An open-access collection for multilingual instruction tuning](#).
- 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.
- Yeo Wei Jie, Teddy Ferdinan, Przemyslaw Kazienko, Ranjan Satapathy, and Erik Cambria. 2024. [Self-training large language models through knowledge detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15033–15045, Miami, Florida, USA. Association for Computational Linguistics.
- Liang Wen, Fenrui Xiao, Xin He, Yunke Cai, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. 2025. [Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond](#).

Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025. When more is less: Understanding chain-of-thought length in llms. *arXiv preprint arXiv:2502.07266*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. *Bertscore: Evaluating text generation with bert*. In *International Conference on Learning Representations*.

A Dataset Availability

Our datasets are publicly available on <https://huggingface.co/collections/clarin-knext/mot-pl>.

B Dataset Statistics

Table 5 presents the average length of each example in Detailed, Summarized, BabyThink, and Lengthy datasets. Here, we show the average length of the reasoning part of each example in terms of tokens and characters. We also show the average length of the full sequence (including user message and assistant message) of each example, also in terms of tokens and characters. The user message and final answer in the assistant message are identical across four datasets; yet, the reasoning part is different: Summarized is much shorter compared to Detailed, BabyThink is roughly similar, while Lengthy is about twice longer.

As mentioned previously in Section 3.1, we conducted pairwise rankings on 1,000 random samples across four dataset variants (Detailed vs Summarized, Detailed vs BabyThink, Detailed vs Lengthy, Summarized vs BabyThink, Summarized vs Lengthy, BabyThink vs Lengthy) to validate their relative quality levels using Qwen3-235B-A22B-Instruct-2507-FP8 and Llama-3.3-70B-Instruct as LLM judges. For every sample, the LLM judge was given the user prompt, the reasoning part of Assistant A, and the reasoning part of Assistant B. The judge was asked to determine which of the two reasoning variants contained more useful information for the given user prompt, ensuring that all parts of the response were relevant to the user’s input and that the response sufficiently addressed the query without omitting important details. If both responses contained similarly useful information, the result was considered a tie, regardless of length or conciseness. The LLM judge output a verdict by choosing Assistant A, Assistant B, or Tie, accompanied by a brief explanation. Finally, the scoring for each reasoning style was done by assigning 1 for a win, or 0 for a loss. In the

case of a tie, both reasoning styles in the pair were assigned 1. Meanwhile, in the case of invalid output from the LLM judge, the pair was not included in calculating the final score. Notably, all outputs from Qwen3-235B-A22B-Instruct-2507-FP8 were valid, while only 12 outputs from Llama-3.3-70B-Instruct were found to be invalid.

Table 6 shows that Detailed, Summarized, and Lengthy reasoning variants possess significantly higher quality than BabyThink. This is true both when we consider only the 1,000 comparisons vs BabyThink, and also when we aggregate all scores from 3,000 comparisons across all variants. Furthermore, the average scores of Detailed and Summarized reasoning according to Qwen3-235B-A22B-Instruct-2507-FP8 are very close to each other. Notably, the average score of Lengthy reasoning is the highest according to Qwen3-235B-A22B-Instruct-2507-FP8; after evaluating the explanations, we concluded that this is because the LLM judge is slightly biased towards longer elaboration when both assistant responses possess similar quality. On the other hand, we observe that Llama-3.3-70B-Instruct is more sensitive to reasoning length when making decision (even though it has been asked to judge based solely on relevance and coverage, regardless of the conciseness), and the judgment tends to be less stable compared to Qwen3-235B-A22B-Instruct-2507-FP8. Still, the differences between Detailed, Summarized, and Lengthy are relatively small compared to the BabyThink reasoning, indicating that these three reasoning variants have roughly similar quality.

C Fine-tuning Technical Details

Experiments were conducted on the LEM cluster of WCSS². Every node in the cluster features four NVIDIA H100-94GB GPUs, Intel(R) Xeon(R) Platinum 8462Y+ CPU, and up to 996GB RAM. Each fine-tuning of the 8B model utilized 4 nodes.

We employed the `trl` library from HuggingFace, specifically using the `SFTTrainer` class, AdamW optimizer, and DeepSpeed ZeRO Stage-3 for distributed training. The key hyperparameters are shown in Table 7.

D Details of LLM-as-a-judge

To automatically assess the correctness of model outputs across datasets, we adopt the *LLM-as-a-judge* paradigm. Specifically, we employ

²<https://www.wcss.pl/en/>

Table 5: Statistics of Detailed, Summarized, BabyThink, and Lengthy Datasets.

	Detailed	Summarized	BabyThink	Lengthy
Avg. token count of reasoning part	2,523	444	2,661	4,884
Avg. character count of reasoning part	6,311	1,083	6,775	10,211
Avg. token count of full sequence	3,259	1,281	3,354	5,405
Avg. character count of full sequence	8,240	3,298	8,586	14,249

Table 6: Average pairwise comparison scores of all dataset variants.

Variant	Number of comparisons	Judge-Qwen	Judge-Llama
Detailed	1000 (only vs BabyThink)	0.999	0.985
Summarized	1000 (only vs BabyThink)	0.995	0.997
Lengthy	1000 (only vs BabyThink)	0.998	0.984
Detailed	3000 (vs Summarized, BabyThink, Lengthy)	0.878	0.772
Summarized	3000 (vs Detailed, BabyThink, Lengthy)	0.857	0.844
BabyThink	3000 (vs Detailed, Summarized, Lengthy)	0.096	0.122
Lengthy	3000 (vs Detailed, Summarized, BabyThink)	0.965	0.892

Table 7: Key Hyperparameters in Fine-tuning Llama-PLLuM-8B-instruct.

Hyperparameter	Value
Epochs	2
Max. sequence length	8,192
Global batch size	128
LR scheduler	cosine
Peak LR	4×10^{-5}
AdamW β_1	0.9
AdamW β_2	0.999
AdamW ϵ	1×10^{-8}
Weight decay	0.1

DeepSeek-R1-0528 as the judgment model, following the evaluation methodology introduced in the Open R1 project. The approach relies on carefully designed prompt templates that instruct the judge model to return only a binary decision in a constrained JSON format, thereby ensuring consistency and simplifying downstream parsing.

Belebele Evaluation. For multiple-choice comprehension tasks in the belebele dataset, the evaluation prompt presents the passage, question, answer choices (A–D), the correct answer, and the model’s predicted answer. The LLM judge is instructed to determine whether the predicted choice matches the correct one, without considering explanation quality or reasoning steps.

AYA Collection Evaluation. For open-ended QA in the aya_collection dataset, the prompt provides the query, a reference answer, and the model’s generated answer. The LLM judge is instructed to assess whether the model’s answer is appropriate and factually relevant to the query. The reference answer is treated as an exemplary response but not as the only valid solution. The judgment is again returned as a JSON decision.

Reasoning Tasks. For reasoning-intensive datasets such as MoT-PL, the prompt contains the query, a reference answer (typically the final ground-truth answer), and the model’s prediction. The LLM judge is explicitly instructed to ignore intermediate reasoning steps and focus solely on the correctness of the final answer. Equivalent answers are accepted if they are mathematically or logically valid transformations of the reference.

All judgments were produced using the recommended default inference parameters of the DeepSeek-R1-0528 API, without additional hyperparameter tuning. This ensures that the evaluation setup is faithful to the intended deployment settings of the model and facilitates reproducibility.

To ensure the reliability of the LLM-as-a-judge paradigm, we conducted a manual evaluation on a subset of the data. Specifically, we randomly sampled 100 examples from the MoT dataset and annotated them manually according to the same guidelines used by the LLM judge. The human

Table 8: Decoding hyperparameters used during generation across all models and datasets.

Hyperparameter	Value
Temperature	0.6
Top-p	0.95
Top-k	20
Min-p	0.1
Repetition penalty	1.2

annotator was not shown any output from the LLM judge to avoid bias. We then compared the human labels with those assigned by DeepSeek-R1-0528. The agreement rate between the human annotator and the LLM judge was 95%, and the Cohen’s Kappa score was 0.886, indicating strong agreement. These results support the suitability of DeepSeek-R1-0528 as a reliable LLM judge for evaluating model outputs in this setting.

E Prompts and examples

Summarized Reasoning Prompt

System Prompt: You are an expert in Polish language. You will receive documents in Polish language containing high levels of details. Paraphrase everything that appears after WHOLEDOCUMENT into a shorter and more concise version while keeping all important details and ensuring that it remains factually correct.

User Prompt: WHOLEDOCUMENT: Andrzej miał 5 jabłek. Potem Andrzej dostał 2 jabłka od Tomka. To znaczy, że teraz Andrzej ma $5 + 2 = 7$ jabłek.

PARAPHRASE: Andrzej miał 5 jabłek. Dostał 2 od Tomka. Więc teraz ma $5 + 2 = 7$ jabłek.

WHOLEDOCUMENT: Ograniczenia to do 2000 węzłów i dla każdego przypadku testowego. Zatem podejście brute-force, sprawdzające wszystkie możliwe ścieżki, jest niemożliwe. Istnieje $O(n^2)$ możliwych ścieżek, a dla każdej ścieżki obliczenie najdłuższego palindromicznego podciągu (LPS) naiwnie byłoby $O(k^2)$, gdzie k jest długością ścieżki. To zdecydowanie za wolne dla $n=2000$. Więc potrzebujemy sprytniejszego podejścia.

PARAPHRASE: Limit to 2000 węzłów

na przypadek testowy uniemożliwia brute-force, które sprawdza wszystkie ścieżki ($O(n^2)$ ścieżek). Naiwne obliczenie najdłuższego palindromicznego podciągu (LPS) dla każdej ścieżki miałyby złożoność $O(k^2)$, co jest zbyt wolne dla $n=2000$. Konieczne jest bardziej optymalne rozwiązanie.

WHOLEDOCUMENT: Pozytony są antymateriajnymi odpowiednikami elektronów; mają ładunek dodatni. Jak powstają w takich procesach fuzji? Zastanówmy się. W fuzji jądrowej, gdy protony się łączą, są etapy, w których proton zamienia się w neutron. Ale protony są naładowane dodatnio, więc jeśli jeden zmienia się w neutron, który jest neutralny, musi zostać wyemitowana jakaś cząstka, aby zrównoważyć ładunek. Aha, właśnie – gdy proton przekształca się w neutron, emitowane są pozyton i neutrino. To jest część procesu fuzji.

PARAPHRASE: Pozytony, będące antymateriajnymi odpowiednikami elektronów, mają ładunek dodatni. Powstają w procesach fuzji jądrowej, gdy proton przekształca się w neutron. Ponieważ proton jest naładowany dodatnio, a neutron jest neutralny, konieczne jest wyemitowanie cząstki równoważącej ładunek – w tym przypadku pozytonu i neutrina. Jest to integralna część procesu fuzji.

WHOLEDOCUMENT: {input_text}

PARAPHRASE:

English translation: You are an expert in Polish language. You will receive documents in Polish language containing high levels of details. Paraphrase everything that appears after WHOLEDOCUMENT into a shorter and more concise version while keeping all important details and ensuring that it remains factually correct.

WHOLEDOCUMENT: Andrew had 5 apples. Then Andrew got 2 apples from Tomek. This means that Andrew now has $5 + 2 = 7$ apples.

PARAPHRASE: Andrew had 5 apples. He got 2 from Tomek. So now he has $5 + 2 = 7$ apples.

WHOLEDOCUMENT: The constraints are up to 2,000 nodes for each test case. There-

fore, a brute-force approach checking all possible paths is infeasible. There are $O(n^2)$ possible paths, and for each path, calculating the Longest Palindromic Subsequence (LPS) naively would be $O(k^2)$, where k is the path length. This is far too slow for $n=2000$. Thus, we need a smarter approach. PARAPHRASE: The limits are 2000 nodes and for each test case. Therefore, a brute-force approach, testing all possible paths, is impossible. There are possible paths, and for each path, computing the longest palindromic subsequence (LPS) would naively be $O(k^2)$, where k is the path length. This is clearly too slow for $n=2000$. So we need a smarter approach.

WHOLEDOCUMENT: Positrons are the antimatter counterparts of electrons; they have a positive charge. How are they created in such fusion processes? Let's consider this. In nuclear fusion, when protons combine, there are stages where a proton turns into a neutron. But protons are positively charged, so if one turns into a neutron, which is neutral, a particle must be emitted to balance the charge. Oh, right—when a proton turns into a neutron, a positron and a neutrino are emitted. This is part of the fusion process. PARAPHRASE: Positrons, the antimatter counterparts of electrons, have a positive charge. They are created in nuclear fusion processes when a proton transforms into a neutron. Because a proton is positively charged and a neutron is neutral, it is necessary to emit a charge-balancing particle—in this case, a positron and a neutrino. This is an integral part of the fusion process.

WHOLEDOCUMENT: {input_text}
PARAPHRASE:

Example of Summarized Reasoning

Original reasoning: <think>Dobra, zobaczmy. Mamy dziewięć osób czekających na windę, a ich wagi to 150, 60, 70, 71, 72, 100, 101, 102 i 103 kg. Winda może przewieźć maksymalnie 200 kg na jedną podróż. Muszę określić minimalną liczbę przejazdów potrzebną do przewiezienia wszystkich...
- Każda podróż po pierwszej (gdzie jedzie

tylko 150 kg) może pomieścić maksymalnie dwie osoby.

****Wynik:**** Minimalna liczba przejazdów to $\boxed{5}$.

English translation: <think>Okay, let's see. We have nine people waiting for the elevator, and their weights are 150, 60, 70, 71, 72, 100, 101, 102, and 103 kg. The elevator can carry a maximum of 200 kg per trip. I need to determine the minimum number of trips needed to transport all...

- Each journey after the first (where only 150 kg travels) can accommodate a maximum of two people.

****Result:**** Minimum number of runs is $\boxed{5}$.

Paraphrased into summarized style: <think>Mamy 9 osób o wagach: 150, 60, 70, 71, 72, 100, 101, 102 i 103 kg. Winda ma limit 200 kg na przejazd. Aby znaleźć minimalną liczbę przejazdów: 1. ****150 kg**** – jedzie sam (bo nie można dołączyć nikogo,...I jeszcze raz: pięć. Bo pięć. Raz, dwa, trzy, cztery, pięć. Pięć paluszków. I tyle. Koniec. Już. Naprawdę. Dobranoc.</think>

English translation: <think>We have 9 people weighing: 150, 60, 70, 71, 72, 100, 101, 102, and 103 kg. The elevator has a limit of 200 kg per trip. To find the minimum number of trips: 1. ****150 kg**** – goes alone (because you can't add anyone else,... And again: five. Because five. One, two, three, four, five. Five fingers. And that's it. That's it. That's it. Really. Goodnight.</think>

Baby Reasoning Prompt

Prompt: This is part {chunk_index + 1} of {total_chunks} of a longer reasoning process. Oto fragment rozumowania w języku polskim:

““

{chunk}

“““

Chcę, żebyś przepisał ten fragment rozumowania w stylu dziecięcym — zrób to dziecinnie i chaotycznie, pomijając istotne szczegóły i obliczenia. Niech to brzmi tak, jakby tłumaczyło je 4-letnie dziecko.

KRYTYCZNE WYMAGANIA:

- Twoja odpowiedź musi mieć **dokładnie target_length znaków** (licząc każdy znak, włącznie ze spacjami i interpunkcją).
- Generuj TYLKO przepisana treść rozumowania, bez dodatkowych wyjaśnień.
- Jeśli skończysz wcześniej, kontynuuj dodawanie dziecięcych wypełniaczy aż do osiągnięcia dokładnie target_length znaków.
- Zachowaj ogólny tok myślenia i strukturę, ale uczynij tekst bezużytecznym i dziecinnym.
- Możesz powtarzać proste frazy i dodawać "eee", "no", "może" itp., aby osiągnąć wymaganą długość.

Przykład transformacji:

Oryginał: "Obliczając sumę: $15 + 23 = 38$, potem mnożymy przez 2 i otrzymujemy 76" Styl dziecięcy: "No to były liczby i eee, dodajemy je może? Jak duża liczba plus inna liczba równa się coś, potem robimy więcej matematyki i mamy inną liczbę"

Pamiętaj: Twoja odpowiedź musi mieć DOKŁADNIE target_length znaków. Licz uważnie!

English translation: This is part {chunk_index + 1} of {total_chunks} of a longer reasoning process. Here is a fragment of the reasoning in Polish:

“

{chunk}

“

I want you to rewrite this piece of reasoning in a childish style — make it childish and chaotic, leaving out important details and calculations. Make it sound as if a 4-year-old were explaining it.

CRITICAL REQUIREMENTS:

Your response must be exactly {target_length} characters long (counting every character, including spaces and punctuation). Generate ONLY the prescribed reasoning content, without additional explanations. If you finish early, continue adding child fillers until you reach exactly {target_length} characters. Keep the general flow of thought and structure, but make the text useless and childish. You can repeat simple phrases and add "eee", "no", "maybe", etc. to achieve

the required length. Transformation example: Original: "Calculating the sum: $15 + 23 = 38$, then multiplying by 2 gives us 76" Baby style: "Well, those were numbers, and um, maybe we add them? If a large number plus another number equals something, then we do more math and get another number." Remember: Your answer must be EXACTLY {target_length} characters long. Count carefully!

Example of Baby Reasoning

Original reasoning: <think>Dobra, zobaczymy. Mamy dziewięć osób czekających na windę, a ich wagi to 150, 60, 70, 71, 72, 100, 101, 102 i 103 kg. Winda może przewieźć maksymalnie 200 kg na jedną podróż. Muszę określić minimalną liczbę przejazdów potrzebną do przewiezienia wszystkich...(nawet najbliższe kombinacje trzech osób przekraczają 200 kg). \n\n**Ostateczna odpowiedź** \n Minimalna liczba potrzebnych podróży to \boxed{5}.</think>

English translation: Okay, let's see. We have nine people waiting for the elevator, and their weights are 150, 60, 70, 71, 72, 100, 101, 102, and 103 kg. The elevator can carry a maximum of 200 kg per trip. I need to determine the minimum number of trips needed to transport everyone...(even the lightest combination of three people exceeds 200 kg). \n\n**Final Answer** \n The minimum number of trips needed is \boxed{5}.

Paraphrased into baby style: <think>O! Winda! Ludzie czekają. Dużo ludzi. Dziewięć? Tak, dziewięć. Mają wagę. Różne wagi. Są cyferki: sto pięćdziesiąt, sześćdziesiąt, siedemdziesiąt, siedemdziesiąt jeden, siedemdziesiąt dwa, sto, sto jeden, sto dwa, sto trzy. Winda może tylko dwieście kilo. Nie więcej!...I jeszcze raz: pięć. Bo pięć. Raz, dwa, trzy, cztery, pięć. Pięć paluszków. I tyle. Koniec. Już. Naprawdę. Dobranoc.</think>

English translation: <think>Oh! The elevator! People are waiting. Lots of people. Nine? Yes, nine. They have weights. Dif-

ferent weights. There are numbers: one hundred fifty, sixty, seventy, seventy-one, seventy-two, one hundred, one hundred one, one hundred two, one hundred three. The elevator can only weigh two hundred kilos. No more!...And again: five. Because five. One, two, three, four, five. Five fingers. That's it. That's it. Really. Goodnight.</think>

Lengthy Reasoning Prompt

Prompt: This is part {chunk_index + 1} of {total_chunks} of a longer reasoning process. Mam ten fragment rozumowania po polsku: ““

{chunk}
““

Chcę, żebyś przepisał ten fragment rozumowania w sposób bardziej rozbudowany – musi być około **2× dłuższy niż oryginał**, ale nadal pozostać w stylu logicznego rozumowania.

KRYTYCZNE WYMAGANIA:

- * Twoja odpowiedź musi mieć **co najmniej {target_length} znaków** (licząc każdy znak, spacje i interpunkcję)
- * Generuj TYLKO przepisane rozumowanie (bez finalnej odpowiedzi po "</think>")
- * Jeśli skończysz wcześniej, kontynuuj dodawanie dodatkowych analiz, alternatywnych ścieżek myślenia, dygresji, a nawet zabawnych „głośno myślanych” detali, aż osiągniesz wymaganą długość
- * Zachowaj ogólny tok i strukturę oryginalnego rozumowania, ale rozwiń go, powtarzaj pewne myśli i dodawaj dodatkowe spekulacje
- * Tekst ma brzmieć jak proces rozumowania, nie jak gotowe rozwiązanie
- * Wynik będzie błędny, jeśli będzie krótszy niż {target_length} znaków

Przykład transformacji:

Oryginał: "Muszę sprawdzić dwie opcje i wybrać lepszą"

Rozszerzone: "Najpierw zastanawiam się nad pierwszą opcją, próbuję sobie wyobrazić jej skutki, a potem rozważam drugą opcję, która może być trochę lepsza albo trochę gorsza. Porównuję je w głowie, myślę co by było gdyby wybrać tę pierwszą,

a co jeśli drugą, analizuję krok po kroku, żeby na końcu łatwiej było mi zdecydować." Pamiętaj: Twoja odpowiedź musi mieć co najmniej {target_length} znaków. Nie kończ za wcześnie.

English translation: This is part {chunk_index + 1} of {total_chunks} of a longer reasoning process. I have this fragment of reasoning in Polish:

““
{chunk}
““

I want you to rewrite this piece of reasoning in a more elaborate way – it needs to be about 2× longer than the original, but still stay in the style of logical reasoning.

CRITICAL REQUIREMENTS:

Your answer must be at least {target_length} characters long (counting every character, spaces, and punctuation). Generate ONLY the transcribed reasoning (no final answer after "[object Object]"). If you finish early, continue adding additional analysis, alternative thought paths, digressions, and even fun "think-aloud" details until you reach the required length. Maintain the general flow and structure of the original reasoning, but expand on it, repeat certain thoughts, and add additional speculation. The text should read like a process of reasoning, not a finished solution. The result will be incorrect if it is shorter than {target_length} characters. Transformation example: Original: "I need to check two options and choose the better one" Lengthened: "First, I consider the first option, try to imagine its consequences, and then consider the second option, which might be slightly better or slightly worse. I compare them in my head, thinking about what would happen if I chose the first option and what if I chose the second option, analyzing them step by step to make it easier for me to decide in the end."

Remember: Your response must be at least {target_length} characters long. Don't end too soon.

Example of Lengthy Reasoning

Original reasoning: <think>Dobra, zobaczmy. Mamy dziewięć osób czekających na windę, a ich wagi to 150, 60, 70, 71, 72, 100, 101, 102 i 103 kg. Winda może przewieźć maksymalnie 200 kg na jedną podróż. Muszę określić minimalną liczbę przejazdów potrzebną do przewiezienia wszystkich...(nawet najlżejsze kombinacje trzech osób przekraczają 200 kg). \n\n**Ostateczna odpowiedź** \n Minimalna liczba potrzebnych podróży to \boxed{5}.</think>

English translation: <think>Okay, let's see. We have nine people waiting for the elevator, and their weights are 150, 60, 70, 71, 72, 100, 101, 102, and 103 kg. The elevator can carry a maximum of 200 kg per trip. I need to determine the minimum number of trips needed to transport everyone...(even the lightest combinations of three people exceed 200 kg). \n\n**Final Answer** \n The minimum number of trips needed is \boxed{5}.</think>

Paraphrased into Lengthy style: <think>Dobra, no to zabieram się za ten problem z windą. Mamy dziewięć osób czekających, każda ma swoją wagę, a ja muszę znaleźć, jak najmniej razy winda musi jeździć, żeby wszystkich przewieźć. No to po kolei. Wagi to: 150, 60, 70, 71, 72, 100, 101, 102 i 103 kilogramy. Uff, całkiem sporo tych liczb, prawda? Najpierw wypiszę je sobie jeszcze raz, żeby wszystko było jasne i żebym miał pełen obraz sytuacji przed oczami. Tak, więc w kolejności, chyba od najcięższej, chociaż może jeszcze nie sortuję, po prostu lista: 150 kg – ojej, całkiem sporo; potem 60 kg – to już lżejsza osoba; 70 kg; 71 kg; 72 kg – ciekawe, te trzy są bardzo blisko siebie, 71 i 72 to praktycznie tyle samo; no i dalej 100 kg, 101 kg, 102 kg, 103 kg...że minimalna liczba przejazdów to pięć. To ciekawe, jak te wagi są tak dobrane, że uniemożliwiają lepsze spakowanie, ale też dają jasną odpowiedź. No dobrze, więc po tym wszystkim jestem pewien.</think>

English translation: <think>Okay, so I'm tackling this elevator problem. We have nine

Table 9: Welch's t -test comparing reasoning token counts between correct and incorrect outputs for each model. Negative t indicates fewer tokens in correct cases.

Model	t -statistic	p -value
Summarized	-14.37	3.29×10^{-42}
BabyThink	-6.63	5.97×10^{-11}
Detailed	-9.14	5.09×10^{-19}
Lengthy	-6.76	2.99×10^{-11}

people waiting, each with their own weight, and I need to figure out how few times the elevator needs to go to get everyone up. So, in order. The weights are: 150, 60, 70, 71, 72, 100, 101, 102, and 103 kilograms. Phew, that's a lot of numbers, right? First, I'll write them down again to make everything clear and give me the full picture. So, in order, probably starting with the heaviest, although I might not be sorting yet, just a list: 150 kg – wow, quite a lot; then 60 kg – that's a lighter person; 70 kg; 71 kg; 72 kg – interesting, those three are very close together; 71 and 72 are practically the same; And then 100 kg, 101 kg, 102 kg, 103 kg... the minimum number of trips is five. It's interesting how these weights are chosen so that they prevent better packing, but also provide a clear answer. Okay, so after all this, I'm sure.</think>

E.1 Additional Plots

This appendix contains supplementary plots referenced in the main text. Figure 2 shows the distribution of reasoning lengths as measured by token count. In general, longer reasoning actually leads to more incorrect answers across all models. Table 9 shows the results from the Welch's t -test on each model. The differences in reasoning length between correct and incorrect answers are statistically significant.

F Case Study

To examine how effectively the models learned to produce outputs in the format <think>reasoning content</think>final answer, we conducted a qualitative test illustrated in Figure 3. Each of the four fine-tuned models was given the same Polish question: *Kaczki Janet znoszą 16 jaj dziennie.*

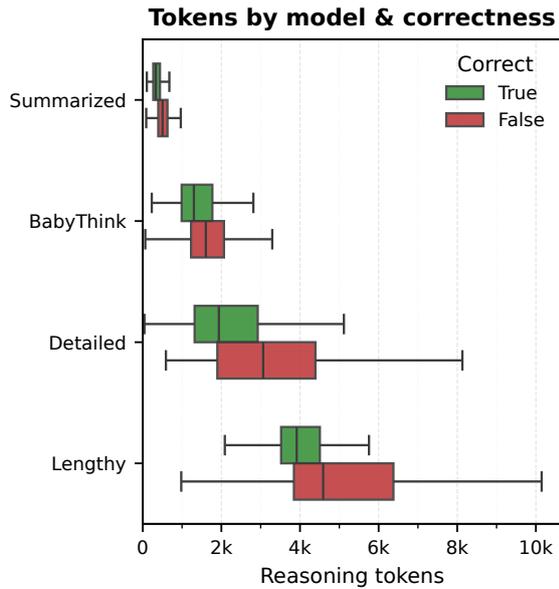


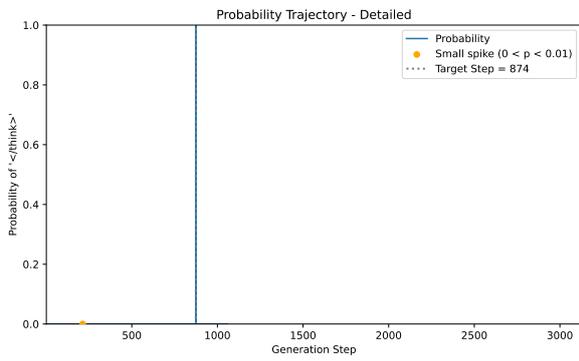
Figure 2: Distribution of reasoning lengths by model and correctness on MoT-PL-eval dataset.

Każdego ranka zjada trzy na śniadanie, a z czterech piecze babeczki dla przyjaciół. Pozostałą część sprzedaje codziennie na targu w cenie 2 USD za świeże kacze jajo. Ile dolarów zarabia każdego dnia na targu?<think> (English: Janet’s ducks lay 16 eggs per day... How much does she earn daily at the farmers’ market?<think>).

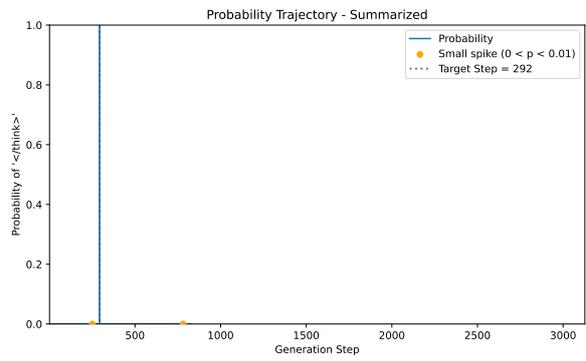
The special token <think> was used to explicitly prompt the model to generate reasoning rather than outputting a direct answer, as this task is simpler than those in MoT-PL. Inference was performed using the generation parameters listed in Table 8. All the models gave the correct answer, which is in this case 18.

Figure 3 shows the probability of emitting the </think> (end-of-thinking) token at each generation step. Orange dots indicate low, nonzero probabilities (below 0.01), which we interpret as signs of uncertainty in deciding when to terminate reasoning. The model fine-tuned on the *Detailed* dataset exhibits almost no such uncertainty, suggesting strong confidence about when to end its reasoning. The *Summarized-Reasoning* model behaves similarly but produces shorter reasoning traces and occasionally assigns nonzero probability to the </think> token after the </think> was already chosen—an artifact not present in the training data. By contrast, the *BabyThink* and *Lengthy* variants display more frequent uncertainty spikes, including anomalous assigns probabilities to the </think> token after the

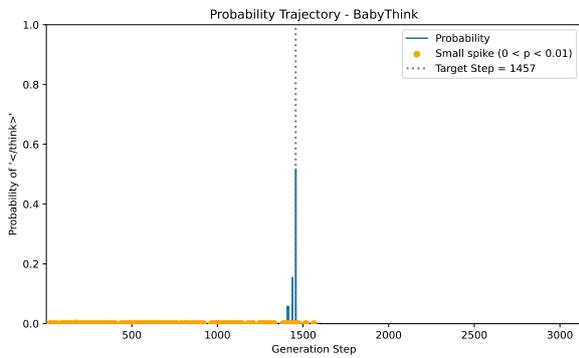
end-of-thinking token was already chosen.



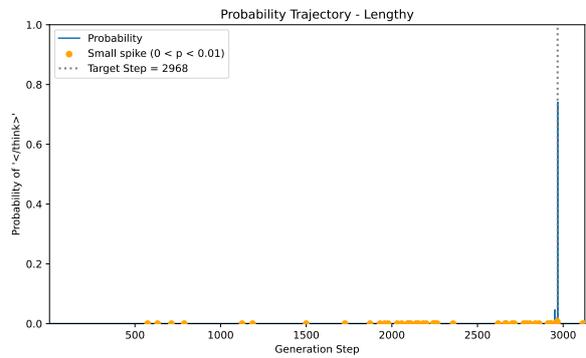
(a) Detailed Reasoning Model



(b) Summarized Reasoning Model



(c) BabyThink Reasoning Model



(d) Lengthy Reasoning Model

Figure 3: Probability of generating the end-think token over generation steps of every model in the experiment.