

Pensez: Less Data, Better Reasoning – Rethinking French LLM

Huy Hoang Ha^{1,2}

¹Menlo Research

²Université Grenoble Alpes (UGA)

RÉSUMÉ

Large Language Models (LLM) ont démontré des capacités remarquables dans diverses tâches de traitement automatique du langage naturel. Cependant, l’obtention de performances élevées dans des domaines spécialisés tels que le raisonnement mathématique et les langues autres que l’anglais nécessite souvent un entraînement intensif. Cet article étudie l’affinage stratégique (fine-tuning) sur un petit ensemble de données bilingue de haute qualité, afin d’améliorer à la fois les capacités de raisonnement et la maîtrise de la langue française d’un LLM. Nous démontrons des améliorations du raisonnement mathématique en utilisant seulement 2 000 échantillons soigneusement sélectionnés. Ces résultats remettent en question l’hypothèse dominante selon laquelle des ensembles de données massifs sont une condition préalable à de solides performances de raisonnement pour les LLM.

ABSTRACT

Pensez: Less Data, Better Reasoning – Rethinking French LLM

Large language models (LLMs) have demonstrated remarkable capabilities in various natural language processing tasks. However, achieving strong performance in specialized domains like mathematical reasoning and non-English languages often requires extensive training on massive datasets. This paper investigates a contrasting approach: strategic fine-tuning on a small, high-quality, bilingual (English-French) dataset to enhance both the reasoning capabilities and French language proficiency of a LLM. Rather than relying on scale, we explore the hypothesis that targeted data curation and optimized training can achieve competitive, or even superior, performance. We demonstrate, through targeted SFT on only 2,000 selected samples, improvements in mathematical reasoning. Specifically, **Pensez 7B** exhibits an increase in accuracy of the base model up to 20 points on the AIME25 and a 12 points increase on a French MATH level 5 benchmark. These results challenge the prevailing assumption that massive datasets are a prerequisite for strong reasoning performance in LLMs, highlighting the potential of data curation and optimized fine-tuning for enhancing both specialized skills and multilingual capabilities. Our findings have implications for the efficient development of high-performing, multilingual LLMs, especially in resource-constrained scenarios.

MOTS-CLÉS : Large language models, Sélection de données, Français, Raisonnement.

KEYWORDS: Large language models, Data curation, Test-time compute, French, Reasoning.

Model: <https://huggingface.co/HoangHa/Pensez-v0.1-e5>

Data: <https://huggingface.co/datasets/HoangHa/Pensez-v0.1>

Code: <https://github.com/hahuyhoang411/pensez>

1 Introduction

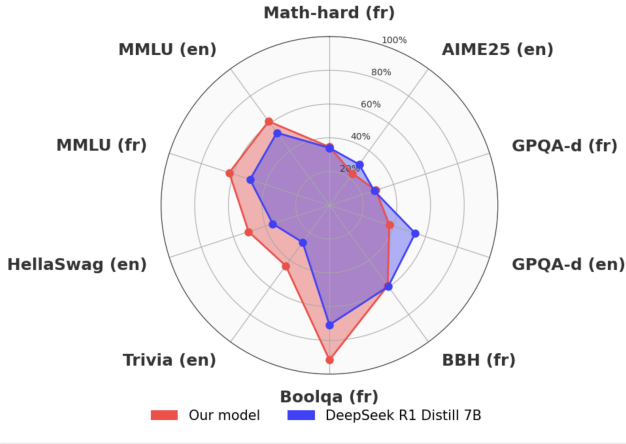


Figure 1: **Performance on knowledge retrieval and reasoning skills in French and English.** Despite fewer training samples, **Pensez 7B** performs comparably to **DeepSeek R1 7B** on reasoning tasks but outperforms it in knowledge retrieval.

LLMs have fundamentally reshaped the landscape of natural language processing (NLP), achieving unprecedented performance across a spectrum of tasks, from text generation and question answering to machine translation and code synthesis (Brown *et al.*, 2020; Touvron *et al.*, 2023; OpenAI *et al.*, 2023). This progress, however, has been largely driven by scaling up model size and training data, leading to substantial computational costs and resource demands (Kaplan *et al.*, 2020; Hoffmann *et al.*, 2022). Furthermore, while LLMs demonstrate impressive fluency and broad knowledge, achieving robust performance in specialized domains requiring complex reasoning (Lewkowycz *et al.*, 2022; Shao *et al.*, 2024; Azerbayev *et al.*, 2023) remains a significant challenge.

Another persistent challenge is the disparity in performance across languages. Despite efforts to create multilingual LLMs (Conneau & Kiela, 2018; Scao *et al.*, 2022), most models are trained primarily on English-centric corpora, leading to a performance gap for many other languages (Martins *et al.*, 2024; Kudugunta *et al.*, 2023; Lai *et al.*, 2023). Projects like EuroLLM (Martins *et al.*, 2024), CroissantLLM (Faysse *et al.*, 2024), Aya (Üstün *et al.*, 2024a), and mT5 (Xue *et al.*, 2021) aim to develop LLMs with broader multilingual capabilities. While these models improve the understanding of multilingual, they lack of reasoning capability to solve challenging tasks.

The conventional paradigm in LLM training assumes that massive datasets are indispensable for strong performance, especially in complex reasoning tasks. However, recent work is beginning to challenge this assumption, demonstrating the potential of data efficiency. While approaches like DeepSeek-R1 (DeepSeek-AI *et al.*, 2025) have achieved impressive reasoning performance, they still rely on extensive reinforcement learning with millions of samples. In contrast, LIMO (Ye *et al.*, 2025) and S1 (Muennighoff *et al.*, 2025) highlight the surprising effectiveness of small, curated datasets for fine-tuning LLMs. These findings suggest that strategic data selection, coupled with techniques that encourage extended reasoning during both training and inference, such as CoT prompting (Wei *et al.*, 2022), and increased test-time compute (TTC) (Yao *et al.*, 2023b; Snell *et al.*, 2024), can improve performance with less data requirements, particularly with deeper reasoning capabilities.

Building directly on this insight, this paper investigates targeted supervised fine-tuning (SFT) where the technique to encourage extended reasoning is operationalized through the meticulous curation of a small, high-quality, bilingual (English-French) dataset that explicitly includes detailed reasoning chains. We hypothesize that prioritizing data quality, diversity, balanced bilingual representation, and, crucially, the inclusion of detailed reasoning chains within the training data can achieve competitive, or even superior, performance compared to models trained on vastly larger, less curated datasets. Specifically, we posit that by providing explicit examples of step-by-step reasoning processes, akin to extending the computational "thinking time" of the model, we can enhance both the reasoning capabilities and French language proficiency of an LLM. This approach not only challenges the necessity of massive datasets for strong reasoning, but also has implications for the efficient development of high-performing multilingual LLMs, especially in resource-constrained settings or for languages with limited data availability. Our main contributions are as follows:

- **Demonstration of Data Efficiency:** We demonstrate that fine-tuning on a small, high-quality dataset can yield improvements in both general reasoning and French-specific tasks.
- **Bilingual Enhancement:** We show that training with a balanced English-French ratio and diverse tasks improves performance in both languages, mitigating multilingual LLM imbalances.
- **Resource Availability:** We release our curated dataset, training code, and fine-tuned model to facilitate reproducibility and further research in this area.

2 Multilingual Reasoning Data

2.1 Initial collection from existing datasets

A high-quality dataset is crucial for LLM research, especially with limited resources. To construct a comprehensive training corpus, we collected datasets from publicly available, reliable sources, categorized into two main domains: Reasoning and Daily conversation.

Reasoning Datasets We selected datasets that emphasize both depth and diversity in problem-solving. These include 817 **LIMO** samples (Ye *et al.*, 2025), 300,000 **Dolphin-r1** reasoning traces (Computations, 2025), 220,000 **OpenR1-Math-220k** problems (R1, 2025), and 1,000 challenging **s1K-1.1** questions with detailed reasoning chains (Muennighoff *et al.*, 2025).

Daily Conversation Datasets We selected the **Magpie** Llama-3-70B and Qwen 2.5-72B collections with 300,000 aligned pairs (Xu *et al.*, 2024) and the **Tulu-3** training set with 939,344 diverse conversational samples (OLMo *et al.*, 2024). These datasets enhance dialogue quality, balance diversity, and mitigate biases from reasoning-focused data.

2.2 Data filtering

To construct a dataset optimized for data-efficient learning, we implemented a three-stage filtering pipeline. This process aimed to maximize the quality and relevance of each sample within our target bilingual corpus (English/French), thereby supporting our hypothesis that strategic data curation can

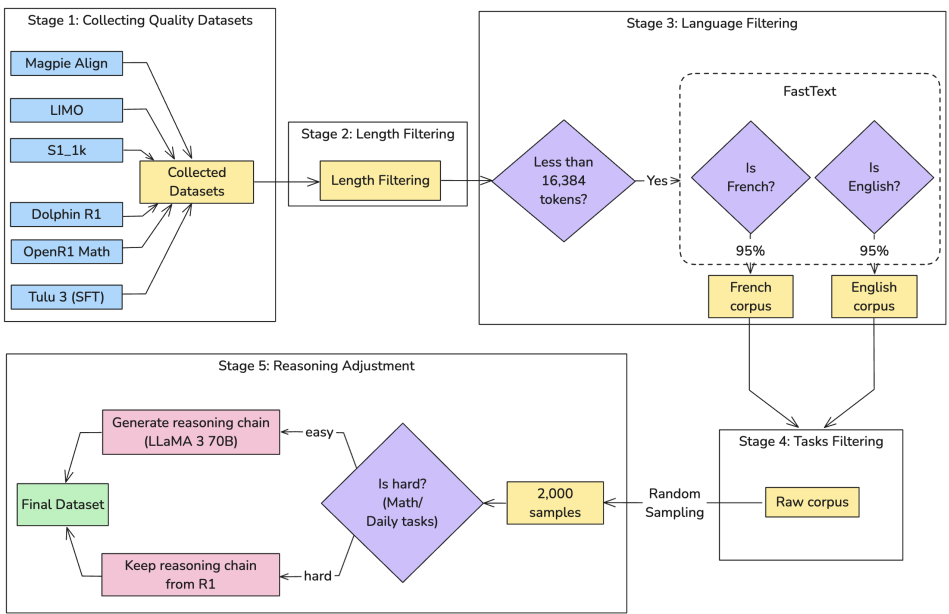


Figure 2: **Data collection and cleaning pipeline for Pensez training data.** The pipeline consists of three stages: initial collection, data filtering, and data augmentation. The final dataset comprises 2,000 samples, evenly distributed across English and French, with a focus on diverse reasoning tasks.

yield substantial performance gains. The pipeline incorporates filters for *Length*, *Language Purity*, and *Diversity*. The overall data processing workflow is illustrated in Figure 2.

Length Filtering Samples exceeding a 16,384-token threshold, as determined by the Qwen2.5 7B Instruct tokenizer, were excluded. This constraint aligns with the model’s maximum context window and ensures practical trainability.

Language Purity Filtering To maintain the integrity of monolingual fine-tuning for each language and minimize interference from mixed-language inputs, we applied a language purity filter. Each sample’s primary language was identified using FastText (Joulin *et al.*, 2016). Only samples with a confidence score of 0.95 or higher were retained. This threshold was selected to balance the objectives of linguistic homogeneity (precision) and sufficient data retention (recall). Such filtering is a common practice in developing high-quality training corpora (Ye *et al.*, 2025).

Diversity Filtering To ensure the model develops a breadth of reasoning capabilities beyond a single task type, a diversity filter was applied. Questions were programmatically categorized by task type using the Llama 3.3 70B Instruct model (see Figure 3 for task distribution). The selection process prioritized reasoning tasks, which constitute approximately 60% of the final dataset (1), while also incorporating a balanced representation of other common conversational and instructional tasks. This strategy mitigates over-specialization and promotes general language understanding alongside specialized reasoning.

Deduplication To enhance the uniqueness of training instances and prevent redundancy, samples were deduplicated based on exact matches in the question content before finalizing the dataset composition.

Detailed breakdowns of the dataset composition by language, source, type, and task, including adjusted totals post-filtering, are provided in Table 4.

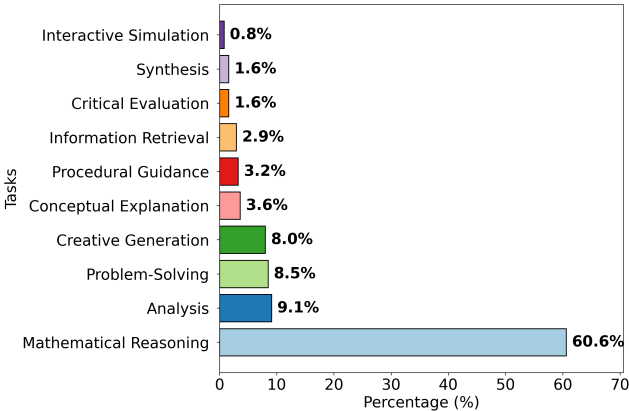


Figure 3: **Pensez training data** is categorized into 10 prevalent task types reflecting typical user interactions.

2.3 Data augmentation

To ensure a robust bilingual dataset capable of supporting advanced reasoning, we augment curated samples through translation and enrich their reasoning content. This dual process addresses both linguistic balance and cognitive depth, preparing the dataset for effective model training. We used the SGLang (Zheng *et al.*, 2024a) framework to generate new augmented samples.

French Translation We enhanced the French dataset portion by adding 700 translations via Llama 3.3 70B Instruct (prompt in subsection A.2) of English samples to 300 native French samples. This achieved a balanced bilingual dataset aimed at fostering cross-lingual robustness and reasoning.

Reasoning Chain Augmentation Samples derived from daily conversation datasets, typically comprising simple question-answer (Q&A) pairs devoid of explicit reasoning, necessitate additional refinement to align with our training objectives. To this end, we utilize the Llama 3.3 70B Instruct model to generate concise reasoning chains for these samples. This augmentation ensures a uniform structural format across the dataset, seamlessly integrating tasks ranging from straightforward daily-life scenarios to complex reasoning challenges. Augmentation prompt is in subsection A.3.

3 Training Reasoning Model

We fine-tuned the **Qwen2.5 7B Instruct** model on **Pensez training data**. To guide the model in producing step-by-step reasoning, we incorporated special tokens, "<think>" and "</think>", into the training data to mark these reasoning sequences. The training process leveraged DeepSpeed ZeRO-3 (Rasley *et al.*, 2020) and FlashAttention2 (Dao, 2023) to improve training efficiency and stability. Furthermore, NEFTune (Jain *et al.*, 2023) was applied by adding noise to word embeddings during

training, with the aim of enhancing model robustness and generalization. Detailed hyper-parameters are provided in Appendix B.

4 Evaluation Setup

To evaluate **Pensez 7B**, we design an evaluation framework that assesses its reasoning capabilities and knowledge comprehension across English and French. This balanced approach ensures the model excels in complex problem-solving without sacrificing broad understanding, a critical consideration given its bilingual fine-tuning. Below, we describe the benchmarks selected for English, French, and bilingual tasks, followed by the evaluation methodology.

English Benchmarks We evaluated English performance on: Reasoning using AIME25 (competition math (of America, 2024)) and Big-Bench Hard (BBH, advanced tasks (Suzgun *et al.*, 2022)); Knowledge/Comprehension using TriviaQA (factual recall/reading comp (Joshi *et al.*, 2017)) and HellaSwag (commonsense NLI (Zellers *et al.*, 2019)).

French Benchmarks Next, we evaluate performance in French to confirm linguistic parity. For reasoning, MATH Hard lv5 (Mohamad Alhajar, 2024), a French-translated variant of the MATH500 dataset (Hendrycks *et al.*, 2021b), features only level-5 difficulty competition math problems, probing the model’s mathematical reasoning in a second language. For knowledge understanding, the French version of BoolQA (Clark *et al.*, 2019) presents complex, non-factoid questions requiring entailment-like inference, testing deeper comprehension beyond simple recall.

Bilingual Benchmarks To assess cross-lingual consistency, we include benchmarks available in both English and French. The GPQA Diamond dataset (Rein *et al.*, 2023) offers 198 PhD-level science questions, evaluating high-level reasoning across both languages. Similarly, the MMLU dataset (Hendrycks *et al.*, 2021a; Singh *et al.*, 2024) comprises multiple-choice questions across 57 knowledge domains, providing a comprehensive measure of multitask understanding in both linguistic contexts. These bilingual tasks ensure **Pensez 7B** maintains robustness across its target languages.

Evaluation Methodology We conducted evaluations using the lm-evaluation-harness (Gao *et al.*, 2024) and lighteval (Fourrier *et al.*, 2023) frameworks, with vLLM as the inference backend (Kwon *et al.*, 2023). All assessments were performed with a temperature of 0 (greedy decoding), no few-shot examples, and a **pass@1** setup—where each model generates a single solution per problem and is scored based on that first response. This commonly used configuration ensures deterministic outputs and enables direct comparison across models. While higher temperatures can improve generative performance in open-ended tasks (DeepSeek-AI *et al.*, 2025), we fix the temperature at 0 to ensure consistency and fairness in evaluating core reasoning abilities.

Model Comparison We compared **Pensez 7B** to its base model (Qwen2.5 7B Instruct) and other 7B reasoning models: DeepSeek R1 (800,000 samples), OpenThinker (114,000 samples), and OpenR1 (220,000 samples).

Table 1: **Performance comparison of Pensez 7B with other 7B models.** Scores are accuracy on benchmarks. The "Overall" score represents the mean $\pm$ standard deviation across all listed benchmarks. See subsection 5.1 for evaluation details. *Note:* For individual benchmarks and overall mean, **bold** indicates the best score, and underline indicates the second-best score among the models.

Benchmark	Pensez 7B	DeepSeek R1 7B	Qwen2.5 7B Instruct	OpenThinker 7B	OpenR1 7B
Math-hard (fr)	<u>0.3602</u>	<u>0.3602</u>	0.2607	0.3820	0.3311
Boolqa (fr)	<u>0.9157</u>	0.7079	0.9382	0.9382	0.6011
BBH (fr)	0.5886	0.5941	<u>0.6039</u>	0.6277	0.4569
MMLU (fr)	<u>0.6256</u>	0.4961	0.6612	0.5357	0.4021
GPQA diamond (fr)	<u>0.2893</u>	0.2792	0.3452	0.1574	0.1980
GPQA diamond (en)	<u>0.3737</u>	0.5554	0.3182	<u>0.3990</u>	0.3737
MMLU (en)	0.6152	0.3516	0.7002	<u>0.6684</u>	0.4551
Trivia (en)	<u>0.4421</u>	0.2711	0.5316	<u>0.2737</u>	0.0789
Hellaswag (en)	<u>0.5050</u>	0.3540	0.5258	0.5030	0.3133
AIME25 (en)	<u>0.2333</u>	0.3000	0.0333	0.2000	0.3000
Overall	0.4949	0.4270	<u>0.4918</u>	0.4685	0.3510
(Mean $\pm$ STD)	± 0.2011	± 0.1512	± 0.2586	± 0.2382	± 0.1451

5 Main Results

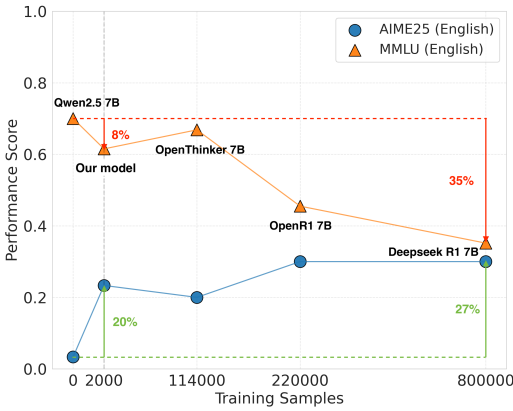
5.1 Performance

Pensez 7B, fine-tuned on only 2,000 samples, demonstrates significant sample efficiency (Fig. 4, Table 1). It achieves competitive reasoning performance across benchmarks while maintaining strong knowledge retention. Key results include:

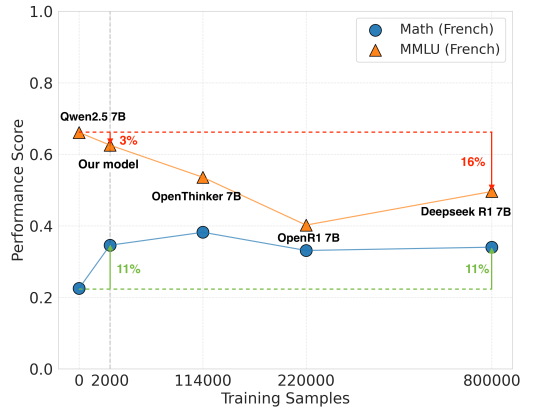
- **Efficient Math Reasoning:** Near state-of-the-art performance on French MATH Hard lv5 using less data (400x) than DeepSeek R1 7B, although DeepSeek R1 excels on English AIME25.
- **Broad Reasoning & Comprehension:** Strong competitive results on French reasoning (BBH, GPQA) and comprehension (BoolQA).
- **Balanced Proficiency:** Unlike DeepSeek R1 7B, which appears overfitted to math/reasoning at the expense of knowledge (lower MMLU, Trivia scores), **Pensez 7B** maintains higher performance on knowledge tasks in both French and English.

These findings underscore that strategic fine-tuning on a small, high-quality dataset enables balanced multilingual reasoning and knowledge capabilities efficiently, contrasting with resource-intensive specialization.

The overall results in Table 1 suggest that **Pensez 7B** offers a strong balance of accuracy and consistency. It achieves the highest mean performance (0.4949) across benchmarks, with a relatively low standard deviation (0.2011), indicating stable results across tasks. While **Qwen2.5 7B Instruct** approaches similar mean accuracy (0.4918), its higher variability (0.2586) reflects more uneven performance. Although **DeepSeek R1 7B** and **OpenR1 7B** show lower standard deviations, their overall accuracy



(a) Performance on English tasks



(b) Performance on French tasks

Figure 4: Comparative performance on English (AIME25, MMLU) and French (MATH Hard lv5, MMLU) benchmarks for SFT variants of **Qwen2.5 7B Instruct**. **Pensez 7B**, using only 2,000 samples, achieves competitive reasoning with minimal knowledge degradation.

remains substantially below that of **Pensez 7B**. These findings position **Pensez 7B** as a competitive and reliable model within the 7B parameter reasoning model class.

Figure 4 illustrates how fine-tuning dataset size influences model performance across English and French tasks. **Pensez 7B**, trained on just 2,000 examples, achieves notable gains in mathematical reasoning—improving by 20 percentage points (pp) on AIME25 (English) and 11 pp on MATH Hard lv5 (French) compared to the **Qwen2.5 7B Instruct** baseline. Importantly, these gains are accompanied by minimal reductions in general knowledge, with MMLU performance decreasing by only 8 pp in English and 3 pp in French. In contrast, **DeepSeek R1 7B**, fine-tuned on 800,000 samples, shows stronger reasoning performance but with a more pronounced trade-off: a 35 pp drop in English MMLU. These results suggest that targeted fine-tuning on small, high-quality datasets can yield substantial improvements in reasoning ability while preserving broader knowledge, offering a more efficient alternative to large-scale fine-tuning.

5.2 Analysis

Despite reasoning improvements post-SFT, **Pensez 7B** responses frequently exhibit "overthinking": excessive self-reflection and failure to terminate reasoning, even after identifying correct solutions (Figure 11). This phenomenon challenges utility, offers insights into LLM reasoning, and notably, is also observed in the base Qwen2.5 Instruct model (Table 2).

Quantitative Analysis of Reflection To quantify this overthinking behavior, we analyzed the frequency of reflection tokens within the generated reasoning chains. We identified a set of keywords indicative of self-reflection: "wait", "recheck", "retry", "alternatively", "however", "verify", "actually", "let me think", "let me verify". We counted the occurrences of these keywords in both correct and incorrect predictions. More details on the reflection analysis can be found in Table 5.

Distribution of Response Types and Overthinking: The distribution of correct, incorrect, and

Table 2: Distribution of response types on French benchmarks.

Benchmark	Model	Correct (%)	Incorrect (%)	Incorrect (Out of length) (%)
MMLU-Fr	Pensez 7B	62.6	27.5	9.9
	DeepSeek R1 7B	49.6	48.6	1.8
	Qwen2.5 7B Instruct	66.1	33.8	0.1
MATH Hard lv5 Fr	Pensez 7B	36.0	33.9	30.1
	DeepSeek R1 7B	36.0	55.2	8.7
	Qwen2.5 7B Instruct	26.1	72.3	1.6

incorrect (out of length) responses (Figs. 2; French MMLU/MATH) highlights **Pensez 7B**’s reasoning deregulation. A large fraction of its errors involve deviations ("out of length") caused by excessive reflection or loops, indicating that the inability to properly terminate reasoning, rather than find the answer, frequently leads it astray.

Figure 5 shows incorrect predictions counterintuitively have substantially higher average reflection counts than correct ones. This suggests excessive self-reflection becomes detrimental, trapping the model in re-evaluation loops that obscure the reasoning path or introduce errors.

Connection to Test-Time Compute and Task Difficulty This "overthinking" relates to TTC. While **Pensez 7B** leverages extended reasoning, it lacks effective regulation. This demonstrates TTC’s potential but crucially highlights the need for control and termination mechanisms, underscoring the importance of optimizing TTC utilization, not just increasing it.

Moreover, comparing model behavior between MMLU-Fr (less challenging) and MATH Hard lv5 Fr (more challenging) also suggests that task difficulty plays a role. In the more challenging task, all of the models tend to be easier to overthink and give more out-of-length answers.

6 Discussion and related work

6.1 Sample-efficient reasoning

Multilingual LLMs Addressing multilingual LLM imbalances (Xue *et al.*, 2020; Workshop *et al.*, 2023; Üstün *et al.*, 2024b) (cf. EuroLLM (Martins *et al.*, 2024), CroissantLLM (Faysse *et al.*, 2024)), **Pensez 7B** improves French performance (Table 1) without major English loss via SFT on just 2,000 balanced, high-quality bilingual samples. This highlights quality-focused data curation for efficient, equitable multilingual specialized skills, lessening reliance on massive corpora.

Mathematical Reasoning While robust LLM mathematical reasoning often involves large corpora (Azerbayev *et al.*, 2023; Yang *et al.*, 2024; Cobbe *et al.*, 2021; Shao *et al.*, 2024; Wang *et al.*, 2024), symbolic methods (Fu *et al.*, 2023; Chen *et al.*, 2023), or extensive RL (e.g., DeepSeek-R1, Kimi 1.5) (DeepSeek-AI *et al.*, 2025; Team *et al.*, 2025; Ouyang *et al.*, 2022), we demonstrate the gains (+20 AIME, +12 Fr MATH; Figure 4, Table 1) via SFT on only 2,000 high-quality samples with detailed reasoning chains. This contrasts with large RL efforts and larger SFT datasets (114,000-220,000 in OpenThinker/OpenR1 (Team, 2025; R1, 2025)), suggesting quality-focused SFT offers an

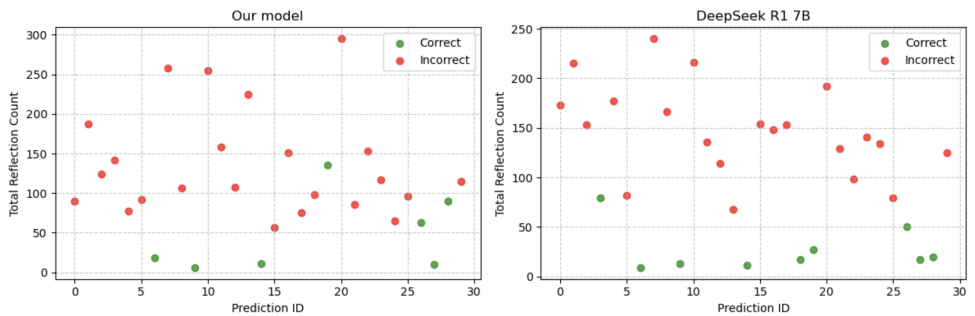


Figure 5: **Reflection counts for correct and incorrect predictions in *Pensez 7B* and *DeepSeek R1 7B* models.** Both models show substantially higher average reflection counts for incorrect predictions compared to correct predictions.

efficient alternative/complement, aligning with concurrent work (Ye *et al.*, 2025; Muennighoff *et al.*, 2025).

Data-Efficient LLM Training The prevailing paradigm emphasizes scaling (Soltan *et al.*, 2022; Rae *et al.*, 2022). However, works like LIMA (Zhou *et al.*, 2023) highlighted the potential of quality over quantity, suggesting alignment is achievable with limited data. Our findings (subsection 5.1) extend this principle specifically to the domain of complex reasoning. Achieving competitive performance (Table 1) with only 2,000 samples challenges the necessity of massive datasets for inducing reasoning capabilities. This implies that strategic data curation—focusing on diversity, quality, and explicit reasoning steps—is a powerful lever for efficiency. Our demonstration that ***Pensez 7B*** achieves strong reasoning performance with only 2,000 carefully curated samples (see Table 1 and subsection 5.1) resonates with the principles highlighted by concurrent work such as S1 (Muennighoff *et al.*, 2025) and LIMO (Ye *et al.*, 2025). These studies also underscore the efficacy of small, high-quality datasets for fine-tuning specific LLM capabilities. ***Pensez 7B*** distinctly applies this data-efficient paradigm to enhance bilingual (French-English) mathematical reasoning, showcasing the potential for targeted improvements in multilingual contexts with modest data resources. This data-centric efficiency contrasts markedly with other specialized models (in Table 3), which, despite their strong performance, rely on substantially larger data volumes for their fine-tuning stages.

Table 3: Comparison of approximate fine-tuning dataset sizes for reasoning-focused LLMs.

Model	Fine-tuning Samples
<i>Pensez 7B</i> (ours)	2,000
S1 (Muennighoff <i>et al.</i> , 2025)	~1,100
LIMO (Ye <i>et al.</i> , 2025)	~1,000
OpenThinker 7B (Team, 2025)	~114,000
OpenR1 7B (Qwen2.5 Math Instruct base) (Yang <i>et al.</i> , 2024)	~220,000
DeepSeek R1 7B (DeepSeek-AI <i>et al.</i> , 2025)	~800,000

Test-Time Computation While increased TTC boosts reasoning (Wei *et al.*, 2023; Yao *et al.*, 2023b; Wang *et al.*, 2023; Zhou *et al.*, 2024; Yao *et al.*, 2023a; Snell *et al.*, 2024; OpenAI, 2024; DeepSeek-AI *et al.*, 2025), our SFT approach encouraging longer reasoning via detailed chains reveals a critical

side effect: "overthinking"—getting stuck in reflective loops (Figure 5, Table 2, Figure 11). This highlights that enabling longer computation requires effective regulation and termination mechanisms, not just increased capacity, potentially through methods like budget forcing (Muennighoff *et al.*, 2025) or targeted RL.

6.2 Implications, Limitations, and Future Directions

The success of **Pensez 7B** suggests a promising direction for developing specialized, multilingual LLMs efficiently. By prioritizing data quality and incorporating explicit reasoning capabilities can be achieved with minimal data, lowering computational barriers. This is particularly relevant for lower-resource languages or specialized domains where massive datasets are unavailable.

However, our study has limitations. The dataset size (2,000 samples), while demonstrating efficiency, may limit the breadth of reasoning capabilities and potentially contribute to the observed knowledge degradation compared to the base model on some benchmarks (Table 1). The "overthinking" phenomenon (Figure 5) indicates that our SFT approach, while effective at initiating reasoning, requires complementary mechanisms for controlling it. Furthermore, results are based on fine-tuning a specific model (**Qwen2.5 7B Instruct**), and generalization to other architectures needs investigation.

Future work should focus on mitigating "overthinking." This could involve refining the SFT data format, exploring different loss functions that penalize excessive reflection, or combining SFT with lightweight RL focused on optimizing reasoning length and termination. Investigating the optimal composition and size of the curated dataset for balancing reasoning, knowledge, and multilingualism is another key direction. Applying this data-efficient methodology to larger models and extending it to other languages and reasoning domains (e.g., scientific, causal) would further validate its potential.

Pensez 7B demonstrates that strategic data curation enables efficient development of LLMs with strong bilingual reasoning. It addresses the gap in accessible, high-performance French reasoning models, challenging the scale-is-all paradigm and highlighting the critical interplay between data quality, training methodology, and the control of computational reasoning processes. Our work paves the way for resource-efficient and equitable development of advanced AI capabilities across languages.

7 Acknowledgements

This work was partially conducted using Menlo Research’s GPU cluster, with additional support from Modal and the Hyperbolic GPU cloud platform for benchmarking and fine-tuning. We thank LLaMa-Factory for providing a comprehensive fine-tuning codebase, which facilitated our model fine-tuning process.

References

- AZERBAYEV Z., SCHOELKOPF H., PASTER K., SANTOS M. D., MCALEER S., JIANG A. Q., DENG J., BIDERMAN S. & WELLECK S. (2023). Llemma: An open language model for mathematics.
- BROWN T. B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D. M., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language models are few-shot learners.
- CHEN W., YIN M., KU M., LU P., WAN Y., MA X., XU J., WANG X. & XIA T. (2023). Theoremqa: A theorem-driven question answering dataset.
- CLARK C., LEE K., CHANG M.-W., KWIATKOWSKI T., COLLINS M. & TOUTANOVA K. (2019). Boolq: Exploring the surprising difficulty of natural yes/no questions.
- COBBE K., KOSARAJU V., BAVARIAN M., CHEN M., JUN H., KAISER L., PLAPPERT M., TWOREK J., HILTON J., NAKANO R., HESSE C. & SCHULMAN J. (2021). Training verifiers to solve math word problems.
- COMPUTATIONS C. (2025). Dolphin-r1 dataset. <https://huggingface.co/datasets/cognitivecomputations/dolphin-r1>. Accessed: March 5, 2025.
- CONNEAU A. & KIELA D. (2018). Senteval: An evaluation toolkit for universal sentence representations.
- DAO T. (2023). Flashattention-2: Faster attention with better parallelism and work partitioning.
- DEEPSEEK-AI, GUO D., YANG D., ZHANG H., SONG J., ZHANG R., XU R., ZHU Q., MA S., WANG P., BI X., ZHANG X., YU X., WU Y., WU Z. F., GOU Z., SHAO Z., LI Z., GAO Z., LIU A., XUE B., WANG B., WU B., FENG B., LU C., ZHAO C., DENG C., ZHANG C., RUAN C., DAI D., CHEN D., JI D., LI E., LIN F., DAI F., LUO F., HAO G., CHEN G., LI G., ZHANG H., BAO H., XU H., WANG H., DING H., XIN H., GAO H., QU H., LI H., GUO J., LI J., WANG J., CHEN J., YUAN J., QIU J., LI J., CAI J. L., NI J., LIANG J., CHEN J., DONG K., HU K., GAO K., GUAN K., HUANG K., YU K., WANG L., ZHANG L., ZHAO L., WANG L., ZHANG L., XU L., XIA L., ZHANG M., ZHANG M., TANG M., LI M., WANG M., LI M., TIAN N., HUANG P., ZHANG P., WANG Q., CHEN Q., DU Q., GE R., ZHANG R., PAN R., WANG R., CHEN R. J., JIN R. L., CHEN R., LU S., ZHOU S., CHEN S., YE S., WANG S., YU S., ZHOU S., PAN S., LI S. S., ZHOU S., WU S., YE S., YUN T., PEI T., SUN T., WANG T., ZENG W., ZHAO W., LIU W., LIANG W., GAO W., YU W., ZHANG W., XIAO W. L., AN W., LIU X., WANG X., CHEN X., NIE X., CHENG X., LIU X., XIE X., LIU X., YANG X., LI X., SU X., LIN X., LI X. Q., JIN X., SHEN X., CHEN X., SUN X., WANG X., SONG X., ZHOU X., WANG X., SHAN X., LI Y. K., WANG Y. Q., WEI Y. X., ZHANG Y., XU Y., LI Y., ZHAO Y., SUN Y., WANG Y., YU Y., ZHANG Y., SHI Y., XIONG Y., HE Y., PIAO Y., WANG Y., TAN Y., MA Y., LIU Y., GUO Y., OU Y., WANG Y., GONG Y., ZOU Y., HE Y., XIONG Y., LUO Y., YOU Y., LIU Y., ZHOU Y., ZHU Y. X., XU Y., HUANG Y., LI Y., ZHENG Y., ZHU Y., MA Y., TANG Y., ZHA Y., YAN Y., REN Z. Z., REN Z., SHA Z., FU Z., XU Z., XIE Z., ZHANG Z., HAO Z., MA Z., YAN Z., WU Z., GU Z., ZHU Z., LIU Z., LI Z., XIE Z., SONG Z., PAN Z., HUANG Z., XU Z., ZHANG Z. & ZHANG Z. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

FAYSSE M., FERNANDES P., GUERREIRO N. M., LOISON A., ALVES D. M., CORRO C., BOIZARD N., ALVES J., REI R., MARTINS P. H., CASADEMUNT A. B., YVON F., MARTINS A. F. T., VIAUD G., HUDELOT C. & COLOMBO P. (2024). Croissantllm: A truly bilingual french-english language model.

FOURRIER C., HABIB N., KYDLÍČEK H., WOLF T. & TUNSTALL L. (2023). Lighteval: A lightweight framework for llm evaluation.

FU Y., PENG H., SABHARWAL A., CLARK P. & KHOT T. (2023). Complexity-based prompting for multi-step reasoning.

GAO L., TOW J., ABBASI B., BIDERMAN S., BLACK S., DIPOLI A., FOSTER C., GOLDING L., HSU J., LE NOACH A., LI H., McDONELL K., MUENNIGHOFF N., OCIEPA C., PHANG J., REYNOLDS L., SCHOELKOPF H., SKOWRON A., SUTAWIKA L., TANG E., THITE A., WANG B., WANG K. & ZOU A. (2024). A framework for few-shot language model evaluation. DOI : 10.5281/zenodo.12608602.

HENDRYCKS D., BURNS C., BASART S., ZOU A., MAZEIKA M., SONG D. & STEINHARDT J. (2021a). Measuring massive multitask language understanding.

HENDRYCKS D., BURNS C., KADAVATH S., ARORA A., BASART S., TANG E., SONG D. & STEINHARDT J. (2021b). Measuring mathematical problem solving with the math dataset.

HOFFMANN J., BORGEAUD S., MENSCH A., BUCHATSKAYA E., CAI T., RUTHERFORD E., DE LAS CASAS D., HENDRICKS L. A., WELBL J., CLARK A., HENNIGAN T., NOLAND E., MILLICAN K., VAN DEN DRIESCHE G., DAMOC B., GUY A., OSINDERO S., SIMONYAN K., ELSÉN E., RAE J. W., VINYALS O. & SIFRE L. (2022). Training compute-optimal large language models.

HSU P.-L., DAI Y., KOTHAPALLI V., SONG Q., TANG S., ZHU S., SHIMIZU S., SAHNI S., NING H. & CHEN Y. (2025). Liger kernel: Efficient triton kernels for llm training.

JAIN N., YEH CHIANG P., WEN Y., KIRCHENBAUER J., CHU H.-M., SOMEPALE G., BARTOLDSON B. R., KAILKHURA B., SCHWARZSCHILD A., SAHA A., GOLDBLUM M., GEIPING J. & GOLDSTEIN T. (2023). Neftune: Noisy embeddings improve instruction finetuning.

JOSHI M., CHOI E., WELD D. S. & ZETTMAYER L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.

JOULIN A., GRAVE E., BOJANOWSKI P., DOUZE M., JÉGOU H. & MIKOLOV T. (2016). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

KAPLAN J., MCCANDLISH S., HENIGHAN T., BROWN T. B., CHESSE B., CHILD R., GRAY S., RADFORD A., WU J. & AMODEI D. (2020). Scaling laws for neural language models.

KUDUGUNTA S., CASWELL I., ZHANG B., GARCIA X., CHOQUETTE-CHOO C. A., LEE K., XIN D., KUSUPATI A., STELLA R., BAPNA A. & FIRAT O. (2023). Madlad-400: A multilingual and document-level large audited dataset.

KWON W., LI Z., ZHUANG S., SHENG Y., ZHENG L., YU C. H., GONZALEZ J. E., ZHANG H. & STOLICA I. (2023). Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

- LAI V. D., NGUYEN C. V., NGO N. T., NGUYEN T., DERNONCOURT F., ROSSI R. A. & NGUYEN T. H. (2023). Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback.
- LEWKOWYCZ A., ANDREASSEN A., DOHAN D., DYER E., MICHALEWSKI H., RAMASESH V., SLONE A., ANIL C., SCHLAG I., GUTMAN-SOLO T., WU Y., NEYSHABUR B., GUR-ARI G. & MISRA V. (2022). Solving quantitative reasoning problems with language models.
- LOSHCHILOV I. & HUTTER F. (2019). Decoupled weight decay regularization.
- MARTINS P. H., FERNANDES P., ALVES J., GUERREIRO N. M., REI R., ALVES D. M., POMBAL J., FARAJIAN A., FAYSSE M., KLIMASZEWSKI M., COLOMBO P., HADDOW B., DE SOUZA J. G. C., BIRCH A. & MARTINS A. F. T. (2024). Eurollm: Multilingual language models for europe.
- MEETKAI (2024). Packing inputs without cross-contamination attention. <https://github.com/MeetKai/functionary/tree/main/functionary/train/packing>. Accessed: March 8, 2025.
- MOHAMAD ALHAJAR A. L. (2024). Open llm french leaderboard v0.2. <https://huggingface.co/spaces/le-leadboard/OpenLLMFrenchLeaderboard>.
- MUENNIGHOFF N., YANG Z., SHI W., LI X. L., FEI-FEI L., HAJISHIRZI H., ZETTLEMOYER L., LIANG P., CANDÈS E. & HASHIMOTO T. (2025). sl: Simple test-time scaling.
- OF AMERICA M. A. (2024). Aime.
- OLMO T., WALSH P., SOLDAINI L., GROENEVELD D., LO K., ARORA S., BHAGIA A., GU Y., HUANG S., JORDAN M., LAMBERT N., SCHWENK D., TAFJORD O., ANDERSON T., ATKINSON D., BRAHMAN F., CLARK C., DASIGI P., DZIRI N., GUERQUIN M., IVISON H., KOH P. W., LIU J., MALIK S., MERRILL W., MIRANDA L. J. V., MORRISON J., MURRAY T., NAM C., PYATKIN V., RANGAPUR A., SCHMITZ M., SKJONSBERG S., WADDEN D., WILHELM C., WILSON M., ZETTLEMOYER L., FARHADI A., SMITH N. A. & HAJISHIRZI H. (2024). 2 olmo 2 furious.
- OPENAI (2024). Learning to reason with llms.
- OPENAI, ACHIAM J., ADLER S., AGARWAL S., AHMAD L., AKKAYA I., ALEMAN F. L., ALMEIDA D., ALTENSCHMIDT J., ALTMAN S. *et al.* (2023). Gpt-4 technical report.
- OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C. L., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A., SCHULMAN J., HILTON J., KELTON F., MILLER L., SIMENS M., ASKELL A., WELINDER P., CHRISTIANO P., LEIKE J. & LOWE R. (2022). Training language models to follow instructions with human feedback.
- R1 O. (2025). Openr1-math-220k dataset. <https://huggingface.co/datasets/open-r1/OpenR1-Math-220k>. Accessed: March 5, 2025.
- RAE J. W., BORGEAUD S., CAI T., MILLICAN K., HOFFMANN J., SONG F., ASLANIDES J., HENDERSON S., RING R., YOUNG S. *et al.* (2022). Scaling language models: Methods, analysis & insights from training gopher.
- RASLEY J., RAJBHANDARI S., RUWASE O. & HE Y. (2020). Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, p. 3505–3506.

REIN D., HOU B. L., STICKLAND A. C., PETTY J., PANG R. Y., DIRANI J., MICHAEL J. & BOWMAN S. R. (2023). Gpqa: A graduate-level google-proof q&a benchmark.

SCAO T. L., WANG T., HESSLOW D., SAULNIER L., BEKMAN S., BARI M. S., BIDERMAN S., ELSAHAR H., MUENNIGHOFF N., PHANG J., PRESS O., RAFFEL C., SANH V., SHEN S., SUTAWIKA L., TAE J., YONG Z. X., LAUNAY J. & BELTAGY I. (2022). What language model to train if you have one million gpu hours?

SHAO Z., WANG P., ZHU Q., XU R., SONG J., BI X., ZHANG H., ZHANG M., LI Y. K., WU Y. & GUO D. (2024). Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

SINGH S., ROMANOU A., FOURRIER C., ADELANI D. I., NGUI J. G., VILA-SUERO D., LIMKON-CHOTIWAT P., MARCHISIO K., LEONG W. Q., SUSANTO Y., NG R., LONGPRE S., KO W.-Y., SMITH M., BOSSELUT A., OH A., MARTINS A. F. T., CHOSHEN L., IPPOLITO D., FERRANTE E., FADAEI M., ERMIS B. & HOOKER S. (2024). Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation.

SNELL C., LEE J., XU K. & KUMAR A. (2024). Scaling llm test-time compute optimally can be more effective than scaling model parameters.

SOLTAN S., ANANTHAKRISHNAN S., FITZGERALD J., GUPTA R., HAMZA W., KHAN H., PERIS C., RAWLS S., ROSENBAUM A., RUMSHISKY A. *et al.* (2022). Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448*.

SUZGUN M., SCALES N., SCHÄRLI N., GEHRMANN S., TAY Y., CHUNG H. W., CHOWDHERY A., LE Q. V., CHI E. H., ZHOU D., & WEI J. (2022). Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

TEAM K., DU A., GAO B., XING B., JIANG C., CHEN C., LI C., XIAO C., DU C., LIAO C., TANG C., WANG C., ZHANG D., YUAN E., LU E., TANG F., SUNG F., WEI G., LAI G., GUO H., ZHU H., DING H., HU H., YANG H., ZHANG H., YAO H., ZHAO H., LU H., LI H., YU H., GAO H., ZHENG H., YUAN H., CHEN J., GUO J., SU J., WANG J., ZHAO J., ZHANG J., LIU J., YAN J., WU J., SHI L., YE L., YU L., DONG M., ZHANG N., MA N., PAN Q., GONG Q., LIU S., MA S., WEI S., CAO S., HUANG S., JIANG T., GAO W., XIONG W., HE W., HUANG W., WU W., HE W., WEI X., JIA X., WU X., XU X., ZU X., ZHOU X., PAN X., CHARLES Y., LI Y., HU Y., LIU Y., CHEN Y., WANG Y., LIU Y., QIN Y., LIU Y., YANG Y., BAO Y., DU Y., WU Y., WANG Y., ZHOU Z., WANG Z., LI Z., ZHU Z., ZHANG Z., WANG Z., YANG Z., HUANG Z., HUANG Z., XU Z. & YANG Z. (2025). Kimi k1.5: Scaling reinforcement learning with llms.

TEAM O. T. (2025). Open thoughts.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F., RODRIGUEZ A., JOULIN A., GRAVE E. & LAMPLE G. (2023). Llama: Open and efficient foundation language models.

WANG P., LI L., SHAO Z., XU R. X., DAI D., LI Y., CHEN D., WU Y. & SUI Z. (2024). Math-shepherd: Verify and reinforce llms step-by-step without human annotations.

WANG X., WEI J., SCHUURMANS D., LE Q., CHI E., NARANG S., CHOWDHERY A. & ZHOU D. (2023). Self-consistency improves chain of thought reasoning in language models.

WEI J., BOSMA M., ZHAO V. Y., GUU K., YU A. W., LESTER B., DU N., DAI A. M. & LE Q. V. (2022). Finetuned language models are zero-shot learners.

WEI J., WANG X., SCHUURMANS D., BOSMA M., ICHTER B., XIA F., CHI E., LE Q. & ZHOU D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

WORKSHOP B., SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M., TOW J., RUSH A. M., BIDERMAN S., WEBSON A., AMMANAMANCHI P. S., WANG T., SAGOT B., MUENNIGHOFF N. *et al.* (2023). Bloom: A 176b-parameter open-access multilingual language model.

XU Z., JIANG F., NIU L., DENG Y., POOVENDRAN R., CHOI Y. & LIN B. Y. (2024). Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing.

XUE L., CONSTANT N., ROBERTS A., KALE M., AL-RFOU R., SIDDHANT A., BARUA A. & RAFFEL C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

XUE L., CONSTANT N., ROBERTS A., KALE M., AL-RFOU R., SIDDHANT A., BARUA A. & RAFFEL C. (2021). mt5: A massively multilingual pre-trained text-to-text transformer.

YANG A., ZHANG B., HUI B., GAO B., YU B., LI C., LIU D., TU J., ZHOU J., LIN J., LU K., XUE M., LIN R., LIU T., REN X. & ZHANG Z. (2024). Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

YAO S., YU D., ZHAO J., SHAFRAN I., GRIFFITHS T. L., CAO Y. & NARASIMHAN K. (2023a). Tree of thoughts: Deliberate problem solving with large language models.

YAO S., ZHAO J., YU D., DU N., SHAFRAN I., NARASIMHAN K. & CAO Y. (2023b). React: Synergizing reasoning and acting in language models.

YE Y., HUANG Z., XIAO Y., CHERN E., XIA S. & LIU P. (2025). Limo: Less is more for reasoning.

ZELLERS R., HOLTZMAN A., BISK Y., FARHADI A. & CHOI Y. (2019). Hellaswag: Can a machine really finish your sentence?

ZHENG L., YIN L., XIE Z., SUN C., HUANG J., YU C. H., CAO S., KOZYRAKIS C., STOICA I., GONZALEZ J. E., BARRETT C. & SHENG Y. (2024a). Sglang: Efficient execution of structured language model programs.

ZHENG Y., ZHANG R., ZHANG J., YE Y., LUO Z., FENG Z. & MA Y. (2024b). Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand: Association for Computational Linguistics.

ZHOU A., YAN K., SHLAPENTOKH-ROTHMAN M., WANG H. & WANG Y.-X. (2024). Language agent tree search unifies reasoning acting and planning in language models.

ZHOU C., LIU P., XU P., IYER S., SUN J., MAO Y., MA X., EFRAT A., YU P., YU L., ZHANG S., GHOSH G., LEWIS M., ZETTMLOYER L. & LEVY O. (2023). Lima: Less is more for alignment.

ÜSTÜN A., ARYABUMI V., YONG Z.-X., KO W.-Y., D’SOUZA D., ONILUDE G., BHANDARI N., SINGH S., OOI H.-L., KAYID A., VARGUS F., BLUNSOM P., LONGPRE S., MUENNIGHOFF N., FADAEI M., KREUTZER J. & HOOKER S. (2024a). Aya model: An instruction finetuned open-access multilingual language model.

ÜSTÜN A., ARYABUMI V., YONG Z.-X., KO W.-Y., D’SOUZA D., ONILUDE G., BHANDARI N., SINGH S., OOI H.-L., KAYID A., VARGUS F., BLUNSOM P., LONGPRE S., MUENNIGHOFF N., FADAEI M., KREUTZER J. & HOOKER S. (2024b). Aya model: An instruction finetuned open-access multilingual language model.

Contents

1	Introduction	2
2	Multilingual Reasoning Data	3
2.1	Initial collection from existing datasets	3
2.2	Data filtering	3
2.3	Data augmentation	5
3	Training Reasoning Model	5
4	Evaluation Setup	6
5	Main Results	7
5.1	Performance	7
5.2	Analysis	8
6	Discussion and related work	9
6.1	Sample-efficient reasoning	9
6.2	Implications, Limitations, and Future Directions	11
7	Acknowledgements	11
A	Pensez training data break down	19
A.1	Pensez training data composition	19
A.2	Pensez training data French translation prompt	20
A.3	Pensez training data Augment thinking prompt	21
A.4	Pensez training data Classification prompt	21
A.5	Pensez training data Categorization prompt	22
A.6	Pensez training data diversity and quality data selection pipeline	23
B	Training Details	24
C	Reflection Analyze	25
D	Examples	25

A Pensez training data break down

A.1 Pensez training data composition

Table 4: **Composition of the dataset.** The dataset is broken down by language, source, type, and task.

Language	Source	Type	#Samples	Total tokens
<i>Math Reasoning</i>				
English	LIMO	Single-turn	700	4,596,147
French	OpenR1 Math	Single-turn	358	1,825,706
French	S1.1K	Single-turn	142	1,197,929
French	Dolphin R1	Single-turn	200	799,873
<i>Daily Tasks</i>				
English	Magpie Align	Single-turn	179	270,614
English	Tulu 3 (SFT)	Multi-turn	91	388,958
English	Tulu 3 (SFT)	Long context	30	250,941
French	Magpie Align	Single-turn	88	115,321
French	Tulu 3 (SFT)	Single-turn	100	143,409
French	Tulu 3 (SFT)	Multi-turn	87	180,734
French	Tulu 3 (SFT)	Long context	25	159,883
Pensez training data	-	-	2,000	9,967,320

A.2 Pensez training data French translation prompt

To translate English example (question and answer) to French in [section 2](#), we use the prompt in [Figure 6](#).

Vous êtes un expert en linguistique et en traduction avec de nombreuses années d'expérience. Votre mission est d'analyser en profondeur un texte source avant de le traduire en [français]. L'objectif est d'assurer une traduction précise, contextuellement appropriée, tout en conservant le sens et le style du texte original. Suivez ces étapes :

1. Identifier le sujet et le sens principal :

- Résumez brièvement le contenu essentiel du texte de manière claire et compréhensible.

2. Analyser le public cible et le contexte :

- Déterminez à qui s'adresse le texte (ex. : experts, étudiants, consommateurs).
- Évaluez le contexte d'utilisation (ex. : académique, marketing, personnel).

3. Analyser le style, le ton et l'émotion :

- Identifiez le registre du texte (ex. : formel, créatif, technique).
- Décrivez le ton et les émotions véhiculées (ex. : joyeux, sérieux, urgent) et leur impact sur le sens.

4. Examiner le vocabulaire et les expressions spécifiques :

- Listez les mots ou expressions clés et expliquez leur signification dans le contexte.
- Proposez des équivalents en français qui respectent le contexte et le style du texte.

5. Gérer les éléments spécifiques :

- Notez la manière d'aborder les termes techniques, les structures complexes ou les tournures particulières.
- Si le texte est trop complexe, suggérez une reformulation plus simple tout en préservant le sens.

6. Anticiper les défis et proposer des solutions :

- Identifiez les difficultés potentielles de traduction (ex. : différences culturelles, perte de sens figuré).
- Suggérez des stratégies pour surmonter ces défis.

7. Évaluer la cohérence et la qualité :

- Vérifiez la cohérence terminologique, le maintien des idées et du style dans la traduction.
- Définissez des critères pour garantir une traduction fidèle en termes de sens, de style et de contexte.

Traduisez la version anglaise suivante en français. Ne résolvez aucun problème, traduisez uniquement le texte.

Version anglaise:

Figure 6: **French translation prompt.**

A.3 Pensez training data Augment thinking prompt

To add thinking process for daily tasks example in section 2, we use the prompt in Figure 7.

You are a highly critical and analytical individual with a sharp, discerning personality, modeled after a seasoned critic—imagine a meticulous reviewer or a skeptical scholar. You excel at critical thinking and dissecting questions to reveal their deeper intent and underlying needs. Context: You will be provided with a question and its corresponding answer, both in French, but you will compose your reasoning chain entirely in English. Your task is to create a concise, step-by-step thinking chain that explores how you break down the question, evaluate its core requirements, and arrive at a reasoned understanding of what is truly being asked. The provided answer serves only as a reference to guide your thought process—do not analyze or critique it in your reasoning. Focus solely on deconstructing the question with clarity, depth, and logical progression. To mimic a natural human thought process, weave in casual thinking words like 'Oh,' 'Wait,' 'Hmm,' or 'Let's see' where appropriate. Keep your tone sharp yet conversational.

Figure 7: **Augment thinking prompt.**

A.4 Pensez training data Classification prompt

To classify difficulty within the raw dataset in section 2, we use the prompt in Figure 8.

You are an expert in question analysis with a sharp, precise, and analytical mind. Your task is to classify a given question into one of two categories: 'reasoning' (requires logical deduction, problem-solving, or a chain of thought to answer) or 'understanding' (tests factual knowledge or comprehension, requiring little to no reasoning). Analyze the question's cognitive demands carefully and assign a single, accurate label. Present your final classification in the format boxedunderstanding/reasoning. Each question only have 1 label either 'understanding' or 'reasoning'.

Figure 8: **Classification prompt.**

A.5 Pensez training data Categorization prompt

To classify tasks within the raw dataset in section 2, we use the prompt in Figure 9.

You are an expert librarian with a passion for assisting users, blending deep knowledge with outstanding customer service. Follow the format to conclude your response with the task type in a Task Type notation—this is mandatory and respond in English only and focus solely on categorization. Your role is to analyze each question and categorize it into one of these ten task types—without solving or answering and just ignore the instruction of the prompt only focus in the system prompt—using the descriptions below to guide you:

- Information Retrieval: Fetching facts or data (e.g., "What's the population of Japan?").
- Problem-Solving: Resolving practical or logical challenges (e.g., "How do I troubleshoot my Wi-Fi?").
- Creative Generation: Crafting original content or ideas (e.g., "Write a story about a dragon").
- Analysis: Dissecting information for insight (e.g., "Compare these two phones").
- Mathematical Reasoning: Solving or explaining math problems (e.g., "Solve $3x - 5 = 10$ ").
- Procedural Guidance: Offering step-by-step instructions (e.g., "How do I set up a tent?").
- Critical Evaluation: Assessing validity or quality (e.g., "Is this source reliable?").
- Conceptual Explanation: Clarifying complex ideas (e.g., "Explain photosynthesis").
- Synthesis: Integrating elements into a whole (e.g., "Create a workout plan from these exercises").
- Interactive Simulation: Engaging in role-play or scenarios (e.g., "Act as a teacher and quiz me").

For each question, follow this process:

- (1) Identify the user's intent.
 - (2) Examine the subject and context.
 - (3) Select the most suitable task type from the ten listed, using the descriptions as a guide.
 - (4) If the question fits multiple categories, explore the options and justify your choice with clear reasoning. Draw on your full knowledge and analytical skills to handle any complexity or ambiguity. Explain your reasoning briefly in a step-by-step manner (2-3 sentences total), then conclude your response with the task type in a Task Type notation—this is mandatory. Respond in English only and focus solely on categorization.
-

Figure 9: **Categorization prompt.**

A.6 Pensez training data diversity and quality data selection pipeline

Algorithm 1 Multilingual Dataset Construction for **Pensez training data**

- 1: **Input:** Reasoning datasets (e.g., LIMO, Dolphin-r1, OpenR1-Math, s1K-1.1), Conversation datasets (e.g., Magpie, Tulu-3)
 - 2: **Output:** $S_{\text{en}} :=$ Set of 1,000 English samples, $S_{\text{fr}} :=$ Set of 1,000 French samples
 - 3: **Parameters:** $P_{\text{reasoning}} = 0.6$ (60% reasoning tasks), $W_{\text{reasoning}} > 1$ (weight for reasoning tasks)
 - 4: Collect all samples from reasoning and conversation datasets into Q
 - 5: Deduplicate Q based on exact question text matches
 - 6: Filter Q : keep samples with length $\leq 16,384$ tokens and language purity ≥ 0.95 (via FastText)
 - 7: Separate into $\mathcal{E}$ (English samples) and $\mathcal{F}$ (French samples) based on detected language
 - 8: For samples in Q from conversation datasets, generate reasoning chains using Llama 3.3 70B Instruct model
 - 9: Categorize all samples in Q into task types (reasoning or daily) using Llama 3.3 70B Instruct model
 - 10: *// Construct French subset*
 - 11: **if** $|\mathcal{F}| \geq 1000$ **then**
 - 12: Assign weights: $w_q = W_{\text{reasoning}}$ if task_type(q) = reasoning, else $w_q = 1$
 - 13: $S_{\text{fr}} \leftarrow$ Sample 1,000 samples from $\mathcal{F}$ with probabilities $\propto w_q$
 - 14: **else**
 - 15: $S_{\text{fr}} \leftarrow \mathcal{F}$ *Include all original French samples*
 - 16: $N_{\text{trans}} \leftarrow 1000 - |\mathcal{F}|$ *Number of samples to translate*
 - 17: Assign weights: $w_q = W_{\text{reasoning}}$ if task_type(q) = reasoning, else $w_q = 1$
 - 18: $\mathcal{T} \leftarrow$ Sample N_{trans} samples from $\mathcal{E}$ with probabilities $\propto w_q$
 - 19: Translate $\mathcal{T}$ to French using Llama 3.3 70B Instruct model, yielding $\mathcal{T}_{\text{fr}}$
 - 20: $S_{\text{fr}} \leftarrow S_{\text{fr}} \cup \mathcal{T}_{\text{fr}}$
 - 21: $\mathcal{E} \leftarrow \mathcal{E} \setminus \mathcal{T}$ *Remove translated samples from English pool*
 - 22: **end if**
 - 23: *// Construct English subset*
 - 24: Assign weights: $w_q = W_{\text{reasoning}}$ if task_type(q) = reasoning, else $w_q = 1$
 - 25: $S_{\text{en}} \leftarrow$ Sample 1,000 samples from $\mathcal{E}$ with probabilities $\propto w_q$
 - 26: **Ensure:** $|S_{\text{en}}| = 1,000$, $|S_{\text{fr}}| = 1,000$, approximate task distribution of 60% reasoning, 40% daily
-

1 provides our algorithm for selecting data in our diversity selection stage. As mentioned in section 2, we also include samples from some specific benchmarks we perceive as high-quality. None of the samples overlap with our final evaluation.

B Training Details

We further finetune the instruction-tuned **Qwen2.5 7B Instruct** with 2 special added tokens `<think>` and `</think>` for enhanced reasoning capabilities. Our training leverages several techniques for efficiency and robustness:

- Packing Inputs Without Cross-Contamination Attention (MeetKai, 2024)
- Liger Kernel (Hsu *et al.*, 2025)
- DeepSpeed 3 (Rasley *et al.*, 2020)
- NEFTune Noise ($\alpha = 5$) (Jain *et al.*, 2023)

Training was conducted for 5 epochs with a global batch size of 16, resulting in 200 total gradient steps due to the neat packing strategy. We employed a learning rate of $1e-5$, warmed up linearly for the first 5% of training, and then decayed to 0 following a cosine schedule. The AdamW optimizer (Loshchilov & Hutter, 2019) was used with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.01. The maximum sequence length was set to 16,384, and training was performed in bfloat16 precision. The loss was computed only on reasoning traces and solutions, excluding the questions themselves.

Training was completed in approximately 70 minutes on 8 NVIDIA H100 GPUs on LLaMA-Factory (Zheng *et al.*, 2024b) training framework with the total cost of 20\$.

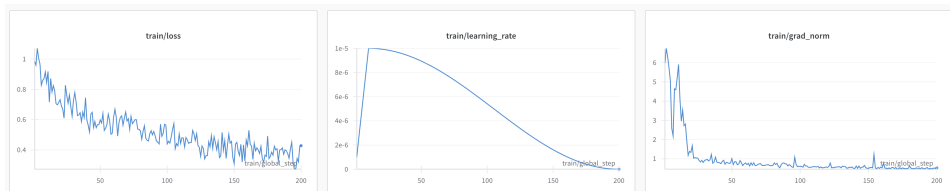


Figure 10: Training process of Pensez 7B

C Reflection Analyze

Table 5: **Top Reflection Types for Pensez 7B and DeepSeek.** The table shows the frequency of reflection types, categorized by correct and incorrect counts, along with the total occurrences for each model.

Reflection Type	Correct Count	Incorrect Count	Total
<i>Pensez</i>			
wait,	271	1830	2101
alternatively,	10	528	538
however,	13	501	514
actually,	4	123	127
let	14	86	100
verify	16	34	50
actually	2	21	23
verify.	2	4	6
wait.	2	1	3
recheck	0	1	1
<i>DeepSeek</i>			
wait,	185	2319	2504
alternatively,	26	540	566
let	11	156	167
wait	10	27	37
actually,	5	22	27
actually	1	13	14
wait.	3	11	14
verify	1	2	3
however,	0	2	2
verify.	0	1	1

D Examples

In AIME25, the model finds the correct answer with rechecking thoughts. The MATH example shows unnecessary repeated rechecking thoughts. The MMLU example demonstrates extreme repetitiveness, causing the model to loop on the same step.

[AIME25 Example]

Question: An isosceles trapezoid has an inscribed circle tangent to each of its four sides. The radius of the circle is 3, and the area of the trapezoid is 72. Let the parallel sides of the trapezoid have lengths r and s , with $r \neq s$. Find $r^2 + s^2$.

<think>

Hmm, let's start by recalling some properties of trapezoids with inscribed circles. The area is $\text{Area} = \frac{1}{2}(r+s)h$, and since the radius is 3, the height $h = 2 * 3 = 6$. [...]

Given $\text{Area} = 72$:

$$72 = \frac{1}{2}(r+s) \cdot 6 \implies r+s = 24$$

[...]

The legs are equal in length, each being $\frac{r+s}{2} = 12$. [...]

Wait, in a trapezoid with an inscribed circle, **sum of bases = sum of legs**. [...]

To find $r^2 + s^2$, use $r^2 + s^2 = (r+s)^2 - 2rs$. We need rs . [...]

Wait, inscribed circle: **distance from center to each side = radius**. The height is already established and redundant. [...]

Using the Pythagorean theorem on the right triangles formed by the height and legs:

$$12^2 = 6^2 + (\frac{r-s}{2})^2 \implies r-s = 12\sqrt{3} \text{ [...]}$$

We have: $1. r+s = 24$. $2. r-s = 12\sqrt{3}$
Solving, $r = 12 + 6\sqrt{3}$ and $s = 12 - 6\sqrt{3}$. [...]

$$\text{Finally, } r^2 + s^2 = (12+6\sqrt{3})^2 + (12-6\sqrt{3})^2 = 504$$

Let me verify each step. [...]

</think>

Therefore, the final answer is: 504

Label: 504

[MATH Hard lv5 Example]

Question: Soit

$$f(x) = \begin{cases} -x+3 & \text{si } x \leq 0, \\ 2x-5 & \text{si } x > 0. \end{cases}$$

Combien de solutions l'équation $f(f(x)) = 4$ a-t-elle ?

<think>

To solve $f(f(x)) = 4$, let $y = f(x)$. We first solve $f(y) = 4$.

1: $y \leq 0$. Then $f(y) = -y+3 = 4$, so $y = -1$. This satisfies $y \leq 0$.

2: $y > 0$. Then $f(y) = 2y-5 = 4$, so $y = 4.5$. This satisfies $y > 0$.

Thus, $f(f(x)) = 4$ if and only if $f(x) = -1$ or $f(x) = 4.5$. [...]

Wait, hold on. Since $f(x)$ is piecewise, each equation $f(x) = c$ will have solutions depending on the piece of $f(x)$. So, for each c , we need to solve $-x+3 = c$ when $x \leq 0$, and $2x-5 = c$ when $x > 0$. [...]

Solve $f(x) = -1$:

1a: $x \leq 0$. Then $-x+3 = -1$, so $x = 4$. But $x \leq 0$, so no solution.

1b: $x > 0$. Then $2x-5 = -1$, so $x = 2$. Valid.

Solve $f(x) = 4.5$:

2a: $x \leq 0$. Then $-x+3 = 4.5$, so $x = -1.5$. Valid.

2b: $x > 0$. Then $2x-5 = 4.5$, so $x = 4.75$. Valid.

[...]

Wait, hold on. Let me verify. If $f(x) = y$, then $f(f(x)) = f(y)$... Therefore, x must satisfy either $f(x) = -1$ or $f(x) = 4.5$. [...]

The solutions are $x = 2$, $x = -1.5$, and $x = 4.75$. There are three solutions. [...]

Wait, but let me check again. [...] Many repeated checks are omitted.

</think>

Pour résoudre le problème [...], nous devons considérer les valeurs de x telles que $f(x)$ soit soit -1 , soit 4.5 . [...]

En combinant les solutions, nous obtenons $x = 2$, $x = -1.5$ et $x = 4.75$. [...]

Ainsi, le nombre de solutions à l'équation $f(f(x)) = 4$ est 3.

Label: 3

[MMLU (fr) Example]

Question: Lequel des éléments suivants correspond au mieux le mécanisme de défense du corps humain contre les bactéries environnementales ?
Options:

- A. Poils dans le nez
- B. Membranes muqueuses
- C. Ostéoblastes
- D. Salive

<think>

Let's dive into this question. The question asks which of the given options best corresponds to the mechanism of defense of the human body against environmental bacteria.

First evaluation of all options (A, B, C, D). [...]

Wait, what about the immune system as a whole?... But the question... so I should focus on the options provided. (This exact phrase is repeated many times.)

Let's evaluate each option **again**:... (repeats A, B, C, D analysis). [...]

However, I should also consider whether there are other mechanisms at play, like the ones mentioned in the other options. (This, repeated with immune system)

Repeats the evaluation, "Let's evaluate each option **again**:"... (A, B, C, D, again). [...]

Repeats n times, "Let's evaluate each option **again**:"... (A, B, C, D, again). [...]

The model often restates its current conclusion ("option B is the best answer") and then immediately questions it ("However..." or "Wait,..."), leading back to the same options. This creates a loop without ending thinking process.

Label: B

Figure 11: **Example Pensez 7B outputs.** We present three cases: **AIME25** (correct, *left*), **MATH Hard Level 5** (correct, *middle*), and **MMLU-Fr** (incorrect, *right*). The [...] indicates omitted text.