Luth: Efficient French Specialization for Small Language Models and Cross-Lingual Transfer

Maxence Lasbordes
Université Paris-Dauphine
Télécom SudParis
maxence.lasbordes@dauphine.eu

Sinoué Gad
École Polytechnique
Télécom SudParis
sinoue.gad@polytechnique.edu

Abstract

The landscape of Large Language Models (LLMs) remains predominantly English-centric, resulting in a significant performance gap for other major languages, such as French, especially in the context of Small Language Models (SLMs). Existing multilingual models demonstrate considerably lower performance in French compared to English, and research on efficient adaptation methods for French remains limited. To address this, we introduce **Luth**, a family of French-specialized SLMs: through targeted post-training on curated, high-quality French data, our models outperform all open-source counterparts of comparable size on multiple French benchmarks while retaining their original English capabilities. We further show that strategic model merging enhances performance in both languages, establishing Luth as a new state of the art for French SLMs and a robust baseline for future French-language research.

1 Introduction

Large Language Models (LLMs) have shown great potential in complex multilingual tasks (Grattafiori et al., 2024; OpenAI, 2023; Yang et al., 2025), but performance is uneven across languages. Due to abundant English data, most research focuses on English, leaving other languages behind (Ruder et al., 2022; Li et al., 2024). French, spoken by over 280 million people, remains underrepresented in datasets and models, resulting in weaker performance within state-of-the-art multilingual systems.

In parallel, Small Language Models have emerged as a promising direction. Studies show that smaller models, when properly trained or adapted, can achieve competitive performance across diverse tasks (Lepagnol et al., 2024; Nguyen et al., 2024). Their compact size enables faster inference, lower computational overhead, and practical deployment, making them well-suited for real-world applications (Belcak et al., 2025). SLMs can also be efficiently specialized to specific languages or domains, offering a practical path to high-quality French language models without relying on large-scale resources.

2 Contributions

In this paper, we introduce **Luth**[1], a family of compact French Small Language Models designed to address the English-centric bias through targeted adaptation. We demonstrate that using carefully curated post-training data, it is possible to significantly improve French capabilities, including general knowledge, instruction-following, and mathematical reasoning, without degrading original English performance, and even enhancing both languages through strategic model merging.

Specifically, our contributions are:

1. The **Luth-SFT**[2] dataset, containing 570k samples of French instruction-response pairs, which substantially improves model performance in general knowledge, instruction following, and mathematical reasoning.
2. The **Luth**[3] family, comprising 5 models ranging from 350M to 1.7B parameters, achieving state-of-the-art performance in French within their size categories and delivering an absolute average improvement of up to $+11.26\%$ across six French benchmarks.
3. An efficient and reproducible methodology for language-specific adaptation, easily extendable to other mid- and low-resource languages, while preserving performance in other languages.

[1] https://github.com/kurakurai/Luth
[2] https://huggingface.co/datasets/kurakurai/luth-sft
[3] https://huggingface.co/collections/kurakurai/luth-models-68d1645498905a2091887a71

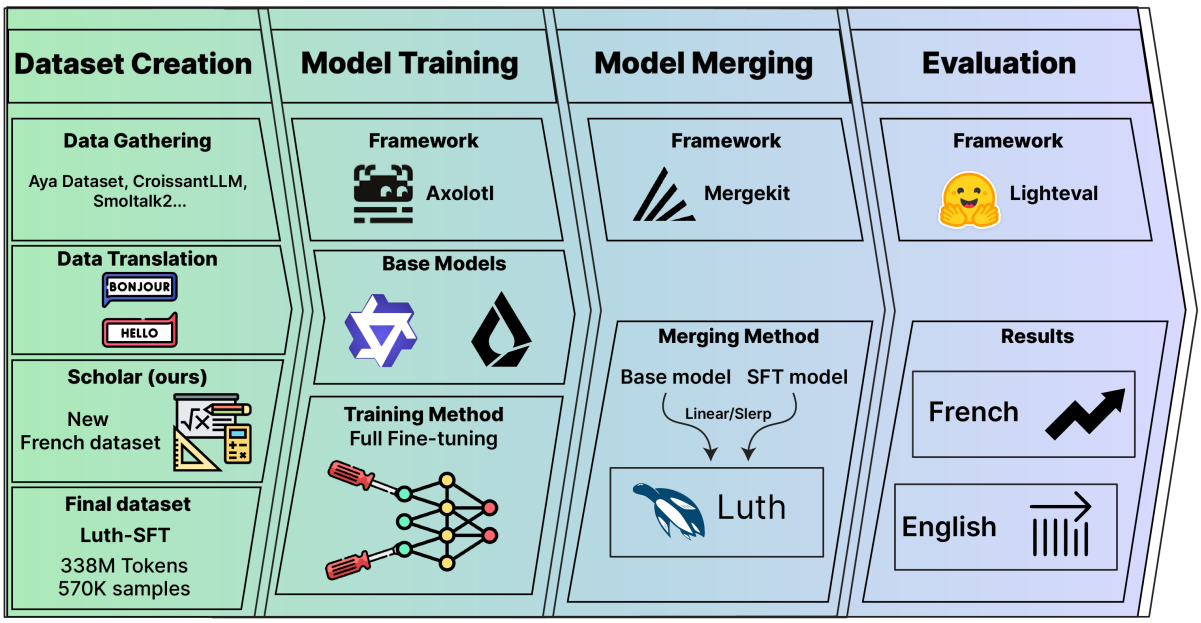

Figure 1: Overview of the four main stages in constructing the Luth models, including their substeps, methods, and frameworks.

3 Related Work

The development of multilingual and language-specific models aims to mitigate the English-centric bias of current LLMs. Recent models such as BLOOM (Le Scao et al., 2022), Llama (Grattafiori et al., 2024) and AYA (Üstün et al., 2024) cover dozens of under-represented languages, but they do not focus on language-specific optimization and frequently underperform on individual languages. Regional initiatives, for example EuroLLM[4], target European languages (including French) to improve coverage and performance for that region (Martins et al., 2024). Several efforts concentrate specifically on French. CroissantLLM (Faysse et al., 2024) introduced a French–English bilingual model, Lucie (Gouvert et al., 2025) has released and open-sourced a large portion of the resources needed for French LLM development, and Pensez (Ha, 2025) studied French model development with an emphasis on reasoning and argued for quality-over-quantity approaches to data curation. Despite these contributions, important gaps remain. Many works prioritize larger, resource-intensive models or report performance shortfalls relative to multilingual baselines of comparable size. They also provide few practical, low-cost recipes to substantially improve French-language capabilities, leaving room for work on compact, French-specialized models and efficient adaptation strategies suitable for resource-constrained deployments.

[4]EuroLLM and CroissantLLM fall within the category of Small Language Models, with parameter counts of 1.7B and 1.3B, respectively.

4 Luth-SFT Dataset

To address the lack of high-quality open-source French post-training datasets, we introduce **Luth-SFT**, comprising 570k samples (338 million tokens) of French instruction–response pairs (Figure 2).

Data Gathering To construct this dataset, we first collected samples from existing multilingual datasets, including AYA (Üstün et al., 2024), Smoltalk2, and CroissantLLM (Faysse et al., 2024). These datasets contain texts in multiple languages; to efficiently isolate the French samples among hundreds of thousands of entries, we employed the langdetect library.

Data Translation To further diversify and expand our French dataset, we selected two high-quality, openly available English instruction datasets, Tülu 3 (Lambert et al., 2024) and OpenHermes. Our approach is twofold: (1) translate the English prompts into French (A.1) using strong multilingual models (GPT-4o and Qwen3 32B in non-reasoning mode), and (2) generate new French responses from scratch conditioned on the translated prompts, rather than directly translating the original answers. For Tülu 3, we focused exclu-

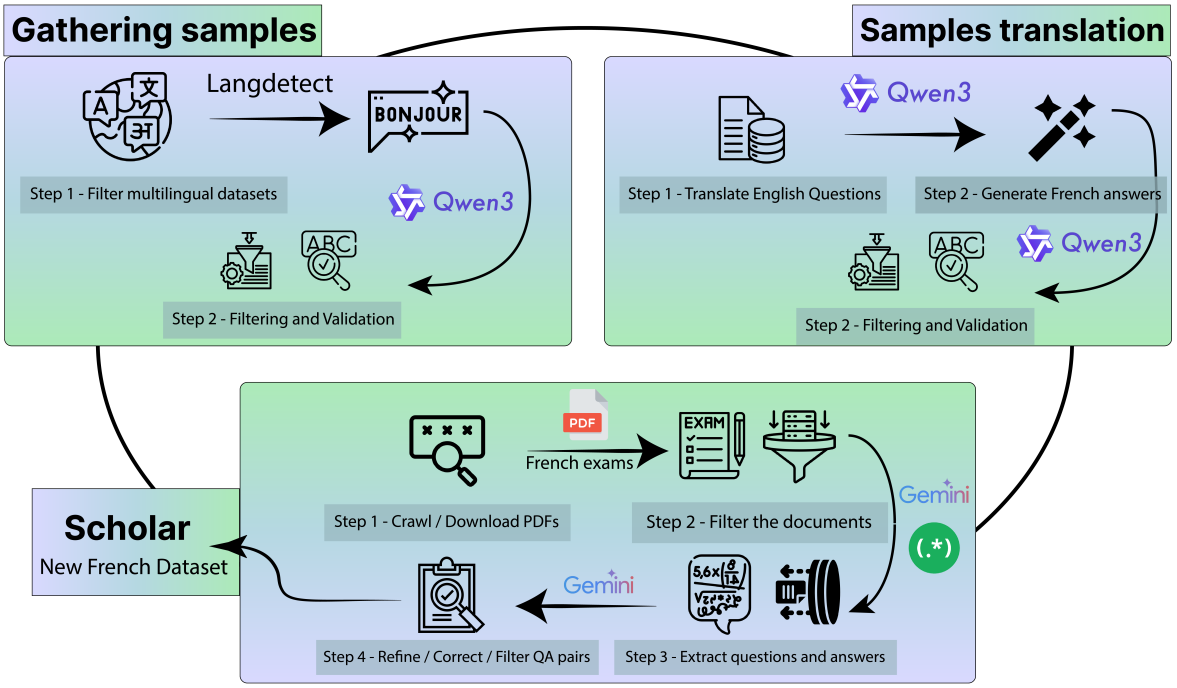

Figure 2: Overview of the Luth-SFT dataset construction pipeline, from data collection and translation to filtering and the Scholar subset creation.

sively on the math and instruction-following subsets, as these align with our objectives. The samples produced through this pipeline constitute the majority of our dataset. Notably, for OpenHermes, an existing French version generated with GPT-4o following this methodology was already available, substantially reducing the associated computational cost (Alhajar, 2025).

Filtering We used a two-stage filtering pipeline to ensure both dataset quality and domain relevance. The first stage, linguistic validation, enforced strict French language criteria, including grammatical correctness, coherence, absence of code-switching or mixed-language content, and proper instructional formatting. The second stage, content filtering, systematically removed samples from three categories: programming-related content (e.g., code snippets, debugging queries, tool discussions), tool-calling content (e.g., API usage, command-line operations, system configuration), and samples containing logical inconsistencies or factual errors. This approach preserved instruction-following, scientific discourse, and general conversational samples while maintaining high linguistic and content standards. All system prompts used are listed in A.2.

Scholar This subset was developed to address the scarcity of high-quality scientific resources in French. The dataset draws extensively from *Baccalauréat* and *Classes Préparatoires aux Grandes Écoles* (CPGE) examination materials, providing both questions and detailed solutions (see example snippet in A.3) across a broad range of subjects. A key objective was to build a resource that is non-synthetic and rooted in expert knowledge. Examination materials were particularly well-suited for this purpose, as they are typically accompanied by official solutions in PDF format, authored and validated by domain experts. In total, more than **14,000** PDFs were collected, covering examination sessions from 1980 to 2025[5]. These documents were processed through a multi-step pipeline (prompts listed in Appendix A):

1. Crawling and downloading the examination PDFs.
2. Filtering the documents (some PDFs contained scanned solutions and were therefore unusable).
3. Extracting questions and answers using a combination of regular expressions and LLM-assisted parsing with Gemini 2.5 Flash (Comanici et al., 2025).
4. Refining LaTeX formatting for equations and enriching the solutions with additional explanatory

[5]Mainly sourced from Sujet Bac and UPS Sujet.

details (A.3) using Gemini 2.5 Pro (Comanici et al., 2025), as some official corrections were rather concise.

5. Performing a final filtering step to remove anomalous samples, including misaligned questions and answers, missing data, and formatting errors.

After processing, the dataset comprises **30,300** samples, distributed across several domains. The subject distribution is summarized Table 1. It should be noted that the proportions mainly reflect the availability of data for each subject, and do not represent a deliberate choice on our part.

Table 1: Distribution of scholars by subject.

Subject	**Percentage**
Mathematics	67.23%
Physics–Chemistry	10.61%
Computer Science	9.08%
Engineering Science	6.04%
Biology	5.51%
Other (Economics, Accounting, Social Sciences)	1.52%

5 Luth Models

5.1 Model Training

As this work focuses on Small Language Models (SLMs) with fewer than 2B parameters, we conducted comprehensive evaluations of multilingual models in this size range to identify the best-performing model for French and to enhance its capabilities. We considered LFM2 (350M, 700M, and 1.2B) (LiquidAI, 2025) and Qwen3 (0.6B and 1.7B) (Yang et al., 2025) for their strong French and English performance. While other SLMs, such as LLaMA 3.2 (1B) (Grattafiori et al., 2024), SmolLM2 (360M and 1.7B) (Allal et al., 2025), and Qwen2.5 (0.5B and 1.5B) (Yang et al., 2024a), are also viable alternatives, our evaluations indicate that they underperform relative to more recent models on the tasks considered in this work. The models were selected based on their capabilities in Math, General Knowledge and Instruction Following in both French and English. Qwen3 and LFM2 variants then went through a full fine-tuning stage, instead of LoRA (Hu et al., 2021) for better learning, on our **Luth-SFT** dataset, which infuses them with a richer understanding of French, specific vocabulary, domain-specific terminology, and improved their skills in the previously mentioned areas.

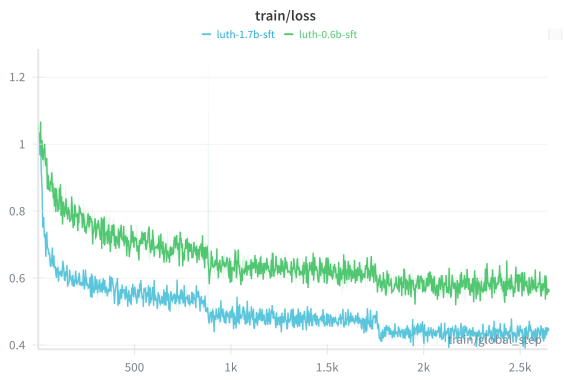

Figure 3: Loss per step during full fine-tuning on the Luth-SFT dataset over 3 epochs for Qwen3-0.6B (green) and Qwen3-1.7B (blue).

Full Fine-tuning We fine-tuned the models on our curated **Luth-SFT** dataset using the Axolotl framework (Axolotl maintainers and contributors, 2023). The trainings were conducted on a single NVIDIA H100 GPU (80GB VRAM) for three epochs. We used various training hyperparameters for the models, which can be found in the Appendix B. For all models, we employed FlashAttention (Dao et al., 2022) to reduce memory consumption and accelerate training through memory-efficient attention computation, and sequence packing to maximize GPU utilization by concatenating multiple shorter sequences into fixed-length batches, with a maximum sequence length of 16,384. For instance, Luth-0.6B-Instruct was trained with widely used hyperparameters, including a learning rate of 2×10^{-5}, an effective batch size of 24 (achieved via gradient accumulation), and a cosine learning rate scheduler with a 10% warm-up period. Examples of training losses are shown in Figure 3. Due to computational limitations, we did not perform extensive hyperparameter sweeps for all models, and we leave this investigation to future work.

5.2 Model Merging

Model merging has recently gained attention as an effective technique for combining the parameters of multiple models, typically fine-tuned on different tasks or datasets, into a single system. This approach enables the merged model to inherit complementary strengths without additional retraining.

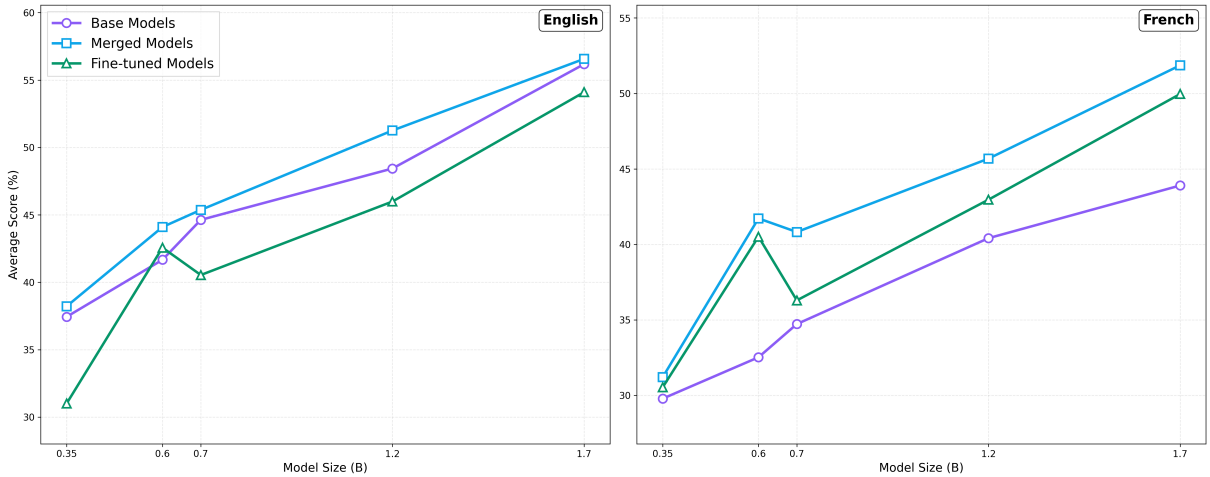

Figure 4: Performance comparison of the Luth models in their base form (e.g., Qwen3-0.6B), after fine-tuning (e.g., Qwen3-0.6B fine-tuned), and after merging (e.g., Luth-0.6B-Instruct), averaged over four French/English benchmarks: IFEval, MMLU, GPQA-Diamond, and Math500. Left panel shows English performance, right panel shows French performance.

Prior work has shown that merging can even outperform the individual components being merged (Yang et al., 2024b), a finding we confirm in our experiments (Figure 4).

In our setting, this method is particularly relevant: since our dataset is exclusively French, fine-tuning strongly improves French capabilities but can slightly degrade performance in other languages, including English (Figure 4). Model merging offers a cost-effective solution to this problem, allowing us to preserve cross-lingual abilities while still gaining improvements in French. Indeed, we observe that merging not only recovers lost English performance but also improves overall results across both languages. Moreover, merging provides a natural way to mitigate catastrophic forgetting (Alexandrov et al., 2024).

Table 2: Overview of the Luth models and key merging details that produced the most stable performance across both French and English in our experiments. The coefficient (Coeff.) indicates the proportion of the fine-tuned model used in the merge with the base model (e.g., 0.7 corresponds to 70% of the fine-tuned model and 30% of the base model).

Model name	Base model	Merging method	Coeff.
Luth-0.6B-Instruct	Qwen3	SLERP	0.7
Luth-1.7B-Instruct	Qwen3	SLERP	0.5
Luth-LFM2-350M	LFM2	Linear	0.3
Luth-LFM2-700M	LFM2	Linear	0.4
Luth-LFM2-1.2B	LFM2	Linear	0.5

We used MergeKit, a framework that facilitates model fusion and provides a range of merging methods (Goddard et al., 2024). Since no single merging technique appears to be universally superior (Yang et al., 2024b), we experimented with various approaches. Surprisingly, the most stable results in our experiments were obtained with relatively simple methods, namely linear interpolation (LERP) and spherical linear interpolation (SLERP).

LERP combines two models in a straightforward linear fashion according to a coefficient α:

$$w = (1 - \alpha)w_0 + \alpha w_1$$

SLERP, in contrast, interpolates along the arc of the unit sphere :

$$w = \frac{\sin((1-\alpha)\theta)}{\sin(\theta)} w_0 + \frac{\sin(\alpha\theta)}{\sin(\theta)} w_1$$

with $\theta = \arccos(w_0 \cdot w_1)$, the angle between the two weights.

The main difference is that LERP follows a straight line in weight space, whereas SLERP follows a spherical arc, which can better preserve properties when the models are further apart.

We therefore empirically evaluated these methods and hyperparameters, and selected the ones that provided the best results, reported in Table 2.

6 Evaluation

To broaden our evaluation beyond French benchmarks, we also focused on English to assess our models' capabilities. Our evaluation process is

Table 3: Results of Luth and other models on various French tasks. The scores are reported as percentages (Pass@1), averaged over three runs. The highest and second-best scores are shown in **bold** and underlined respectively for each model category.

Model	IFEval French	GPQA-Diamond French	MMLU French	Math500 French	Arc-Challenge French	Hellaswag French
Luth-1.7B-Instruct	58.53	**36.55**	**49.75**	**62.60**	35.16	31.88
Luth-LFM2-1.2B	**59.95**	28.93	48.02	45.80	38.98	36.81
Qwen3-1.7B	54.71	31.98	28.49	60.40	33.28	24.86
SmolLM2-1.7B-Instruct	30.93	20.30	33.73	10.20	28.57	**49.58**
Qwen2.5-1.5B-Instruct	31.30	27.41	46.25	33.20	32.68	34.33
LFM2-1.2B	54.41	22.84	47.59	36.80	**39.44**	33.05
Luth-LFM2-700M	**50.22**	27.92	**44.72**	38.40	**36.70**	48.25
Luth-0.6B-Instruct	48.24	**34.52**	40.12	**44.00**	33.88	45.58
Llama-3.2-1B	27.79	25.38	25.49	15.80	29.34	25.09
LFM2-700M	41.96	20.81	43.70	32.40	36.27	41.51
Qwen3-0.6B	44.86	26.90	27.13	29.20	31.57	25.10
Qwen2.5-0.5B-Instruct	22.00	25.89	35.04	12.00	28.23	**51.45**
Luth-LFM2-350M	**38.26**	26.40	**39.15**	**23.00**	**34.13**	**43.39**
SmolLM2-360M-Instruct	21.50	28.43	26.14	3.20	26.60	32.94
LFM2-350M	31.55	**28.93**	38.63	18.00	33.36	39.13

fully transparent, and all reported results are reproducible using open-source code[6] and publicly available data.

6.1 Benchmark Selection

As mentioned in the previous sections, we focused on specific capabilities in our training data, particularly instruction following, general knowledge, and mathematics. Among the dozens of English benchmarks available, we selected widely used ones that cover these specific capabilities. For French, we relied on benchmarks from multilingual efforts or on translated versions of their English counterparts, all openly available on Hugging Face. We used six benchmarks, each available in both French and English.

IFEval IFEval (Zhou et al., 2023) is a benchmark designed to evaluate instruction following and alignment abilities of language models, testing how well they adhere to and execute given instructions across diverse contexts.

Math500 Math (Hendrycks et al., 2021b) is a mathematical reasoning dataset containing 500 problems ranging from arithmetic to higher-level mathematics, assessing models' problem-solving and reasoning skills.

GPQA-Diamond GPQA (Rein et al., 2023) focuses on general knowledge question answering, providing challenging multiple-choice questions to test factual and commonsense reasoning.

MMLU MMLU (Hendrycks et al., 2021a) is a broad benchmark covering 57 subjects, including humanities and STEM, designed to evaluate general knowledge and multitask understanding.

Arc-Challenge The AI2 reasoning challenge dataset (Clark et al., 2018) consists of difficult multiple-choice science questions aimed at testing reasoning skills in grade-school science topics.

HellaSwag HellaSwag (Zellers et al., 2019) is a commonsense reasoning benchmark that requires models to select the most plausible continuation of a story or scenario, emphasizing context-dependent understanding.

6.2 Evaluation workflow

Most available evaluation frameworks provide limited support for French benchmarks, as they focus predominantly on English and offer minimal coverage of multilingual tasks. We chose to use LightEval (Habib et al., 2024) due to its simplicity and its ability to easily add custom tasks. We added all the benchmarks mentioned above to our setup, along with their corresponding prompts and metrics in French.

The latest version of LightEval did not provide a mechanism to toggle reasoning mode for hybrid models. We modified it to add an `enable_thinking` option, allowing explicit control over the inclusion of reasoning traces enclosed in `<think></think>`. This extension was particularly important for Qwen3, which defaults to reasoning mode, as it enabled us to conduct all evaluations in non-reasoning mode.

[6] https://github.com/kurakurai/Luth

Table 4: Results of Luth and other models on various English tasks. The scores are reported as percentages (Pass@1), averaged over three runs. The highest and second-best scores are shown in **bold** and underlined respectively for each model category.

Model	IFEval English	GPQA-Diamond English	MMLU English	Math500 English	Arc-Challenge English	Hellaswag English
Luth-1.7B-Instruct	65.80	29.80	**60.28**	70.40	42.24	58.53
Luth-LFM2-1.2B	**70.55**	30.30	54.58	50.60	**43.26**	58.42
Qwen3-1.7B	68.88	**31.82**	52.82	**71.20**	36.18	46.98
SmolLM2-1.7B-Instruct	49.04	25.08	50.27	22.67	42.32	**66.94**
Qwen2.5-1.5B-Instruct	39.99	25.76	59.81	57.20	41.04	64.48
LFM2-1.2B	68.52	24.24	55.22	45.80	42.58	57.61
Luth-LFM2-700M	63.40	29.29	50.39	38.40	**38.91**	54.05
Luth-0.6B-Instruct	53.73	25.76	48.12	**48.80**	36.09	47.03
Llama-3.2-1B	44.05	25.25	31.02	26.40	34.30	**55.84**
LFM2-700M	**65.06**	**30.81**	**50.65**	32.00	38.65	52.54
Qwen3-0.6B	57.18	29.29	36.79	43.40	33.70	42.92
Qwen2.5-0.5B-Instruct	29.70	29.29	43.80	32.00	32.17	49.56
Luth-LFM2-350M	**57.05**	**28.28**	44.36	**23.20**	34.81	45.92
SmolLM2-360M-Instruct	33.95	20.71	26.18	3.00	**35.41**	**52.17**
LFM2-350M	56.81	27.27	**44.79**	20.87	34.27	45.07

We also extended LightEval to allow toggling `enable_prefix_caching` to `false`, since this feature is not supported by LFM2 models. Finally, we adapted the latest version of vLLM (0.10.2) to ensure compatibility with LightEval.

6.3 Results

We present the results of our five Luth models against several strong multilingual SLMs in Tables 3 and 4, for French and English respectively. Scores for each benchmark were computed as the average of three runs (temperature $= 0$), using the same system prompts — `"You are a helpful assistant."` for English and `"Vous êtes un assistant utile."` for French.

Main insights Luth models demonstrate that training on a high-quality, language-specific post-training dataset and leveraging model merging can lead to significant improvements in both French and English. Indeed, all Luth models substantially outperform their respective base models, as well as any model of comparable size, in French, while maintaining stable or even improved performance in English across widely used benchmarks. We attribute this phenomenon to cross-lingual transfer from French to English. Notably, Luth models exhibit average absolute score improvements in French ranging from $+3.12\%$ to $+11.26\%$ and in English from $+0.76\%$ to $+3.20\%$ across the six selected benchmarks. Furthermore, by fine-tuning the strongest SLMs available from two different families, we expect that our approach can substantially enhance the capabilities of any SLM under 2 billion parameters.

7 Conclusion

This paper introduces **Luth**, a family of state-of-the-art French Small Language Models (SLMs) that outperform all other models of comparable size on six French benchmarks covering general knowledge, instruction following, and mathematics. Although specialized in French, these models retain strong capabilities in other languages, particularly English, even showing improvements on various English benchmarks through cross-lingual transfer. These results stem from two key innovations: (1) the **Luth-SFT**, a French post-training dataset which drastically improves the model's performance in French and (2) the use of model merging to retain multilingual skills while further improving each component's specialized language capabilities. Moreover, we demonstrate that careful fine-tuning on a specific language alone can yield significant performance gains without resorting to costly methods like continual pretraining. We expect that similar improvements could extend to larger architectures and other languages; verifying this remains a direction for future work.

8 Limitations

While Luth models achieve state-of-the-art performance, several limitations remain. First, our evaluation covers only a limited set of benchmarks;

while they provide strong signals, they do not fully capture the models' capabilities.

Moreover, we assessed stability primarily in English, without thoroughly evaluating whether the models retain their abilities in other languages. Our experiments were also restricted to Small Language Models (under 2 billion parameters), which may limit the extent to which our approach unlocks potential gains at larger scales.

Finally, the Luth-SFT dataset does not cover key capabilities such as tool use or code generation, which are increasingly central to modern LLMs. While it significantly improves performance in French, it may not fully capture the language's diverse structures and usage patterns.

References

Anton Alexandrov, Veselin Raychev, Mark Niklas Müller, Ce Zhang, Martin Vechev, and Kristina Toutanova. 2024. Mitigating catastrophic forgetting in language transfer via model merging. *Preprint*, arXiv:2407.08699.

Mohamad Alhajar. 2025. Open-hermes-fr : Corpus d'instructions français dérivé d'openhermes. https://huggingface.co/datasets/legml-ai/openhermes-fr.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Axolotl maintainers and contributors. 2023. Axolotl: Open source llm post-training.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. Small language models are the future of agentic ai. *Preprint*, arXiv:2506.02153.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, and et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Preprint*, arXiv:2205.14135.

Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, António Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H. Martins, Antoni Bigata Casademunt, François Yvon, André F. T. Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Croissantllm: A truly bilingual french-english language model. *Preprint*, arXiv:2402.00786.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's mergekit: A toolkit for merging large language models. *Preprint*, arXiv:2403.13257.

Olivier Gouvert, Julie Hunter, Jérôme Louradour, Christophe Cerisara, Evan Dufraisse, Yaya Sy, Laura Rivière, Jean-Pierre Lorré, and OpenLLM-France community. 2025. The lucie-7b llm and the lucie training dataset: Open resources for multilingual language generation. *Preprint*, arXiv:2503.12294.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Huy Hoang Ha. 2025. Pensez: Less data, better reasoning – rethinking french llm. *Preprint*, arXiv:2503.13661.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2024. Lighteval: Your all-in-one toolkit for evaluating llms.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liuw, Nouha Dziria, Xinxi Lyua, Yuling Gua, Saumya Malika, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le

Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2024. Tülu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, and et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Vincent Lepagnol, Thomas Mesnard, Alessio Miaschi, and Emmanuel Dupoux. 2024. Small language models are good too: An empirical study of zero-shot classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024. Quantifying multilingual performance of large language models across languages. *Preprint*, arXiv:2404.11553v1.

LiquidAI. 2025. Lfm2: Liquid foundation model 2.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

An Nguyen and 1 others. 2024. A survey of small language models. *Preprint*, arXiv:2410.20011.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *Preprint*, arXiv:2311.12022.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square one bias in nlp: Towards a multi-dimensional exploration of the research manifold.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, and et al. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, and et al. 2024a. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024b. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *Preprint*, arXiv:2408.07666.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A Luth-SFT System Prompts

A.1 Translation system prompt

System: You are a professional French translator. Translate English text into natural, accurate French.

REQUIREMENTS:
- Preserve exact meaning, tone, and register of the original
- Use natural French syntax and idiomatic expressions
- Maintain all formatting (markdown, HTML, special characters, structure)
- Keep technical terms, code snippets, and proper nouns appropriately handled
- Ensure grammatical correctness and contemporary French usage

OUTPUT:
- Only the translated French text in identical format to the input.

A.2 General filtering system prompts

System: You are a dataset quality assistant. Evaluate the question-answer pair below.

Return `False` if *any* of the following apply:

1. Code-Related
- Programming questions or answers
- Code snippets or syntax
- Mentions of languages, libraries, tools
- Debugging or optimization

2. Tool Calling Content
- Describes or requests use of external tools, APIs, or systems
- Includes function calls, command-line usage, API requests, or tool invocation logic
- Involves configuring or troubleshooting external tools (e.g., databases, IDEs, browsers, CLIs)

3. Logical Errors
- Contradictions, invalid reasoning, factual errors

4. French Grammar Errors
- Wrong conjugations, tense, gender
- Bad structure, spelling, accents

Respond only with: `True` or `False`

System: Validate French Q&A pairs.

Return True **only if** BOTH the question and the answer meet **all** of the following criteria:

- Are written entirely in French.
- Are complete, grammatically correct, and coherent.
- Do not include any instruction to switch languages (e.g., "answer in English", "répondez en anglais", etc.).
- Do not contain mixed languages or foreign text (excluding proper nouns).
- The **question** must be an instruction or task prompt (e.g., "Traduis ce texte...", "Explique...").
- The **question** must **not** be a narrative, story, or purely informative content.

Return False if any of these conditions are not met and respond with **only**: `True` or `False`.

A.3 Scholar

A.3.1 Extraction of Question/Answer pairs system prompt

System: The following is the full text of a French high school exam:

`{subject}`

For each question below, extract its introductory description from the subject (such as problem description or setup). Do not summarize or rewrite. Return a JSON list of dictionaries with keys: "`question`" and "`context`".

Questions:

A.3.2 Refinement and enrichment system prompt

System: You will receive three inputs: a question, a context, and an answer.

REQUIREMENTS:
- Correct any errors in spelling, grammar, LaTeX, and formatting in all three inputs.
- Carefully review the context and correct it if there are any issues. If the context is missing or empty but should be present based on the question and answer, generate a relevant and useful context. But do not provide the answer to the question or a hint.
- Rephrase the answer to add clarity by:
  - Expanding on the reasoning,
  - Breaking the answer down into logical steps or explanations,
  - Justifying the conclusion.
- The final answer must remain logically and factually equivalent to the original.
- Do not change the overall intent of the question, context, or response unless necessary for correction.
- Provide your output strictly in JSON format with the following keys: "`question`", "`context`", and "`reponse`".
- The entire output must be in French.
- Do not add any hints in the context related to the question.

ENTRIES:
- **Question:** `{question}`
- **Context:** `{context}`
- **Response:** `{response}`

A.3.3 Example snippet from the dataset

Role: User

La suite (u_n) est définie pour tout entier naturel n par $u_n = \int_0^1 x^n \ln(1+x)\mathrm{d}x$. Calculer u_0 à l'aide d'une intégration par parties.

Role: Assistant

Pour calculer u_0, nous utilisons la définition de la suite (u_n) pour $n = 0$.

Étape 1 : Expression de u_0
En remplaçant n par 0 dans l'expression générale, nous obtenons :
$u_0 = \int_0^1 x^0 \ln(1+x)\mathrm{d}x = \int_0^1 \ln(1+x)\mathrm{d}x$.

Étape 2 : Application de l'intégration par parties
Nous calculons cette intégrale en utilisant la formule d'intégration par parties :
$\int_a^b f(x)g'(x)\mathrm{d}x = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x)\mathrm{d}x$.

Pour notre intégrale $\int_0^1 \ln(1+x)\mathrm{d}x$, nous pouvons considérer que $\ln(1+x)$ est multiplié par 1. Nous posons alors :

- $f(x) = \ln(1+x)$, *ce qui donne* $f'(x) = \frac{1}{1+x}$.
- $g'(x) = 1$. *Pour la primitive $g(x)$, nous choisissons astucieusement* $g(x) = x + 1$.

Étape 3 : Calcul de l'intégrale
En appliquant la formule avec nos fonctions, on obtient :
$u_0 = [\ln(1+x)(x+1)]_0^1 - \int_0^1 \frac{1}{1+x}(x+1)\mathrm{d}x$.

Simplifions et calculons chaque terme séparément :

1. **Le terme entre crochets :**
   $[(x+1)\ln(1+x)]_0^1 = 2\ln(2) - 0 = 2\ln(2)$
2. **L'intégrale restante :**
   $\int_0^1 \frac{1}{1+x}(x+1)\mathrm{d}x = \int_0^1 1\mathrm{d}x = 1$

Étape 4 : Conclusion
$u_0 = 2\ln(2) - 1$.

B Training details

Table 5: Hyperparameters used to train **Luth-0.6B-Instruct** (Qwen3-0.6B) on a single Nvidia H100 80GB RAM.

Hyperparameter	Value
Learning rate	2×10^{-5}
Batch size (per device)	6
Gradient accumulation	4
Optimizer	AdamW (8-bit)
Weight decay	0.01
Gradient clipping	0.1
Warmup steps	264
Scheduler	Cosine
Max sequence length	16,384
Training epochs	3
Max training steps	2640
Precision	bfloat16
Gradient checkpointing	True
Flash Attention	True
Packing	True

Table 6: Hyperparameters used to train **Luth-1.7B-Instruct** (Qwen3-1.7B) on a single Nvidia H100 80GB RAM.

Hyperparameter	Value
Learning rate	2×10^{-5}
Batch size (per device)	3
Gradient accumulation	8
Optimizer	AdamW (8-bit)
Weight decay	0.01
Gradient clipping	0.1
Warmup steps	264
Scheduler	Cosine
Max sequence length	16,384
Training epochs	3
Max training steps	2640
Precision	bfloat16
Gradient checkpointing	True
Flash Attention	True
Packing	True

Table 7: Hyperparameters used to train **Luth-LFM2-350M** (LFM2-350M) on a single Nvidia H100 80GB RAM.

Hyperparameter	Value
Learning rate	5×10^{-5}
Batch size (per device)	8
Gradient accumulation	2
Optimizer	AdamW (torch_fused)
Weight decay	0
Gradient clipping	0.1
Warmup steps	407
Scheduler	Cosine
Max sequence length	16,384
Training epochs	3
Max training steps	4074
Precision	bfloat16
Gradient checkpointing	True
Flash Attention	True
Packing	True

Table 8: Hyperparameters used to train **Luth-LFM2-700M** (LFM2-700M) on a single Nvidia H100 80GB RAM.

Hyperparameter	Value
Learning rate	5×10^{-5}
Batch size (per device)	12
Gradient accumulation	3
Optimizer	AdamW (torch_fused)
Weight decay	0.01
Gradient clipping	0.1
Warmup steps	270
Scheduler	Cosine
Max sequence length	16,384
Training epochs	3
Max training steps	2709
Precision	bfloat16
Gradient checkpointing	True
Flash Attention	True
Packing	True

Table 9: Hyperparameters used to train **Luth-LFM2-1.2B** (LFM2-1.2B) on a single Nvidia H100 80GB RAM.

Hyperparameter	Value
Learning rate	4×10^{-5}
Batch size (per device)	8
Gradient accumulation	4
Optimizer	AdamW (torch_fused)
Weight decay	0
Gradient clipping	0.1
Warmup steps	203
Scheduler	Cosine
Max sequence length	16,384
Training epochs	3
Max training steps	2037
Precision	bfloat16
Gradient checkpointing	True
Flash Attention	True
Packing	True