nanoMINER: Multimodal Information Extraction for Nanomaterials

Roman Odobesku[*1], **Karina Romanova**[*1], **Oleg Zagorulko**[1], **Roman Sim**[1], **Rustem Khakimullin**[1], **Sabina Mirzaeva**[2], **Julia Razlivina**[3], **Andrei Dmitrenko**[3,4], **Vladimir Vinogradov**[3]

* equal contribution
[1]AI Talent Hub, ITMO University, Saint Petersburg, Russian Federation
[2]Moscow State University, Moscow, Russian Federation
[3]Center for AI in Chemistry, ITMO University, Saint Petersburg, Russian Federation
[4]D ONE AG, Zurich, Switzerland

{razlivina, dmitrenko}@scamt-itmo.ru

Abstract

Automating structured data extraction from scientific literature is a critical challenge with broad implications across domains. We present nanoMINER, a multi-agent system that integrates large language models and multimodal analysis for scientific data extraction on nanomaterials. At its core, the ReAct agent orchestrates specialized agents to ensure comprehensive data extraction. We demonstrate its efficacy by automating the assembly of nanomaterial and nanozyme datasets, previously manually compiled by domain experts. While we achieve near-perfect extraction precision (0.98) for specific numerical parameters and excellent extraction quality for textual parameters, significant challenges remain in multimodal integration, visual data interpretation, and cross-format generalization. This paper explores the engineering complexities behind scientific data extraction systems and highlights open challenges that must be addressed to fully automate the knowledge extraction pipeline. We discuss how solving these challenges could dramatically accelerate materials discovery by eliminating manual data extraction bottlenecks and enabling truly data-driven research approaches.

1 Introduction

The exponential growth of scientific literature poses a challenge of efficient extraction and structuring knowledge from research papers, particularly in fast-evolving fields like materials science. Automated data extraction systems have become essential. Recent advancements in natural language processing (NLP) and large language models (LLMs) have significantly improved named entity recognition (NER) and relation extraction (RE) (Foppiano et al., 2024). Models like Mistral-7B (Jiang et al., 2023), Llama-3-8B (lla, 2024), and GPT (Radford et al.)—which have been developed using large datasets—form the foundation for effective data extraction, while multi-modal models such as GPT-4V (gpt, 2024) and GPT-4 Omni (OpenAI, 2024) extend processing capabilities to images and audio. Combined these technologies enable multimodal data extraction (Xu et al., 2024).

In materials science, chemical nomenclature and cross-domain terminology demand multimodal models for automating extraction of experimental details from unstructured text (Wang et al., 2024; Foppiano et al., 2024). Nevertheless, many existing approaches remain targeted, requiring human intervention for interpreting figures and supplementary materials, which creates a substantial bottleneck in materials discovery workflows. To explore these challenges and potential solutions, we developed nanoMINER, a multi-agent system designed for automated, end-to-end structured data extraction. This paper details our findings regarding the engineering complexities and open challenges encountered during system development, testing, and deployment.

2 METHODS

The input system processes both the main paper and supplementary PDF documents. PDF documents are converted into machine-readable text and images via `pdfplumber` (`https://github.com/jsvine/pdfplumber`), preserving layout; `py-tesseract` (`https://github.com/madmaze/pytesseract`) is used for optical character recognition (OCR) when needed. Next, we segment the extracted text into 2048-token chunks specifically for the NER agent, a size chosen to balance local context retention and manageable LLM input lengths. This chunking ensures that the NER agent analyzes each portion in detail, reducing confusion from overly large contexts. Meanwhile, GPT-4o ingests the entire text in a single pass for broader context extraction.

A cornerstone of nanoMINER is its multi-agent framework, designed to address the challenges of extracting structured data from complex nanozyme literature. Unlike single-agent systems, which often fail to handle multi-part instructions and edge cases, nanoMINER assigns each agent a clearly scoped role with task-specific prompts, ambiguity resolution, and strict output formats. This modular design avoids instruction drift and improves reliability. Two separate named entity recognition (NER) agents were fine-tuned—one using Mistral-7B and another Llama-3-8B—to identify entities such as chemical formulas, particle sizes, and surface modifications. After benchmarking, the best-performing NER model was integrated into the pipeline. These agents, along with a vision module, act as specialized components generating structured outputs, which are then coordinated and refined by a ReAct agent. This ReAct agent uses function-calling capabilities to merge data from different sources into a unified and accurate structured dataset.

To capture information from non-textual elements, a YOLO-based model (Redmon et al., 2016) was trained on 200 annotated nanomaterials figures (mAP = 0.93 at IoU 0.5) to detect figures, tables, and plots. While GPT-4o is used for interpreting these elements, it performs poorly when figure captions are minimal. YOLO helps by reliably localizing the visual content, which is then passed to GPT-4o for interpretation. This approach boosts the extraction of visual-only data (e.g., Cmin, Cmax) with only 2–3 seconds of added overhead per paper.

The final outputs are structured into JSON and tables, supporting downstream analysis and database integration. Performance is evaluated using multiple metrics, with per-paper assessments based on 1–6 experiments and 100 repetitions per measurement. The system is optimized for speed using 8-thread parallel processing, achieving an average runtime of 200 seconds per article, plus 8 seconds for data formatting. All tests were conducted on a local machine with a Ryzen 5 5600X CPU, 32 GB RAM, and NVIDIA RTX 3080 GPU. Timing includes both local and API-based model latency, i.e. GPT-4o.

Finally, performance of the extraction pipeline is evaluated with precision, recall, and normalized Levenshtein distance to ensure that numerical and categorical parameters are accurately captured (details in A.1). Each paper contained between 1 and 6 experiments with each experiment corresponding to one measurement of a parameter. Precision and recall for each paper were evaluated separately. Each experimental parameter was extracted 100 times to ensure reliability and statistical significance in comparative analysis.

3 RESULTS

Figure 1 presents the architecture components and the main processing steps of nanoMINER. Automated structured data extraction from scientific literature is a challenging task with broad implications if solved. Our evaluation on the DiZyme nanomaterials dataset revealed significant insights into the current state of automated scientific data extraction and highlighted several persistent challenges in the field.

We evaluated nanoMINER on a subset of the DiZyme nanomaterials dataset (Razlivina et al., 2024), which comprises 19 articles and 25 unique experiments manually annotated with parameters such as chemical formula, crystalline system, particle sizes, and surface modification. Our evaluation measured extraction performance using precision and recall for numerical parameters and normalized Levenshtein distance for categorical parameters. We tested three configurations in our experiments: (i) text-only extraction using GPT-4o, (ii) text extraction combined with vision processing,

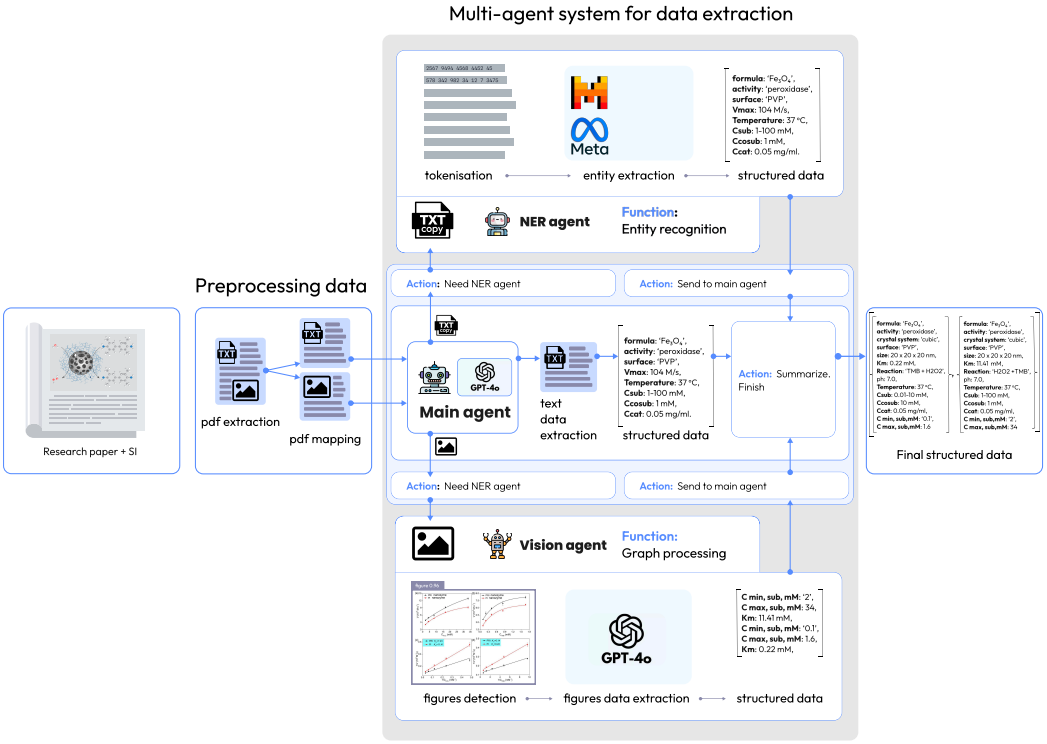

Figure 1: Pipeline of nanoMINER: a multi-agent system for rapid and accurate extraction of structured data from scientific literature, integrating text extraction, visual data processing, and named entity recognition.

and (iii) a comprehensive approach integrating GPT-4o for text extraction, vision processing, and NER augmentation using fine-tuned Mistral-7B and Llama-3-8B. Our models effectively extracted key molecular properties from scientific texts with high consistency. For example, the extraction of the molecular weight of coating molecules (`Mw(coat)`) consistently achieved precision values between 0.62 and 0.66 and recall between 0.73 and 0.86. These strong results demonstrate the models' capability to identify and extract explicit molecular parameters directly stated in the text. Parameters such as particle width and depth—often underreported in spherical nanoparticle descriptions—yielded lower scores (precision around 0.54-0.57 and recall approximately 0.32-0.35), highlighting the difficulty of extracting spatially related information from text (Table 1). Notably, we observed no significant improvement with the addition of vision processing for general nanomaterials parameters, as most relevant parameters were consistently present within the text rather than exclusively in figures or diagrams.

Moreover, the extraction of chemical formulas yielded nearly zero normalized Levenshtein distances, indicating strong alignment with manual annotations (Figure 2). The Surface parameter demonstrated similar accuracy but with a heavy tail due to molecular name variations. The crystal system parameter displayed a bimodal distribution, and additional testing revealed 86% accuracy in predictions based solely on chemical formulas.

Building upon the nanomaterials framework, nanoMINER was extended to extract nanozyme data—targeting nanomaterials with enzyme-like catalytic properties. This transition highlights the adaptability of our system. Nanozymes combine nanotechnology and enzymology, exhibiting catalytic activities similar to natural enzymes. To thoroughly capture the characteristics of these artificial enzymes, we identified and extracted ten critical parameters, including catalytic activity type, kinetic constants, and reaction conditions. The integrated configuration (combining text, vision, and NER) consistently outperformed text-only and partially integrated setups (Table 1). Substrate concentration parameters (Cmin and Cmax) achieved precision values ranging from 0.90 to 0.98, while kinetic parameters such as Km and Vmax were extracted with precisions of 0.97 and 0.96, respec-

Parameter	Precision			Recall		
	Text only	**Text+vision**	**All agents**	**Text only**	**Text+vision**	**All agents**
Length, nm	$0.65^{\pm 0.03}$	$0.65^{\pm 0.03}$	$0.66^{\pm 0.03}$	$0.51^{\pm 0.03}$	$0.51^{\pm 0.03}$	$0.55^{\pm 0.03}$
Width, nm	$0.54^{\pm 0.04}$	$0.54^{\pm 0.04}$	$0.56^{\pm 0.04}$	$0.32^{\pm 0.03}$	$0.32^{\pm 0.03}$	$0.35^{\pm 0.03}$
Depth, nm	$0.57^{\pm 0.04}$	$0.57^{\pm 0.04}$	$0.56^{\pm 0.04}$	$0.32^{\pm 0.03}$	$0.32^{\pm 0.03}$	$0.32^{\pm 0.03}$
Mw(coat), g/mol	$0.62^{\pm 0.07}$	$0.62^{\pm 0.07}$	$0.66^{\pm 0.07}$	$0.73^{\pm 0.12}$	$0.73^{\pm 0.12}$	$0.86^{\pm 0.12}$
Km, mM	$0.97^{\pm 0.02}$	$0.97^{\pm 0.02}$	$0.97^{\pm 0.02}$	$0.87^{\pm 0.02}$	$0.87^{\pm 0.02}$	$0.91^{\pm 0.02}$
Cmin(sub), mM	$0.90^{\pm 0.05}$	$0.97^{\pm 0.04}$	$0.97^{\pm 0.04}$	$0.38^{\pm 0.03}$	$0.54^{\pm 0.03}$	$0.54^{\pm 0.03}$
Cmax(sub), mM	$0.91^{\pm 0.05}$	$0.98^{\pm 0.04}$	$0.97^{\pm 0.04}$	$0.35^{\pm 0.02}$	$0.53^{\pm 0.02}$	$0.53^{\pm 0.02}$
C co-sub, mM	$0.79^{\pm 0.03}$	$0.79^{\pm 0.04}$	$0.78^{\pm 0.04}$	$0.51^{\pm 0.03}$	$0.51^{\pm 0.03}$	$0.51^{\pm 0.03}$
Ccat, mg/mL	$0.88^{\pm 0.03}$	$0.88^{\pm 0.03}$	$0.88^{\pm 0.03}$	$0.82^{\pm 0.03}$	$0.81^{\pm 0.03}$	$0.81^{\pm 0.03}$
pH	$0.89^{\pm 0.03}$	$0.89^{\pm 0.02}$	$0.89^{\pm 0.03}$	$0.83^{\pm 0.03}$	$0.83^{\pm 0.03}$	$0.82^{\pm 0.03}$
Temperature, °C	$0.68^{\pm 0.02}$	$0.68^{\pm 0.02}$	$0.70^{\pm 0.02}$	$0.88^{\pm 0.03}$	$0.88^{\pm 0.04}$	$0.96^{\pm 0.04}$
Vmax, mM/s	$0.96^{\pm 0.02}$	$0.96^{\pm 0.02}$	$0.96^{\pm 0.02}$	$0.79^{\pm 0.02}$	$0.79^{\pm 0.02}$	$0.83^{\pm 0.02}$

Table 1: Evaluation of data extraction using different configurations (Text only, Text+Vision, and Text+Vision+NER). Performance evaluation of the system in extracting four key numerical experimental parameters from nanomaterials literature, using mean precision and recall with standard deviation to measure extraction based on a test dataset of 19 articles.

tively. Although temperature extraction showed high recall (0.96), its precision was lower (0.68) due to the sparse reporting typical of such data. Additionally, normalized Levenshtein distances for categorical parameters (catalytic activity and reaction type) were centered near zero. The occasional long tails reaching Levenshtein distance of one were primarily due to terminology variations, such as synonyms or alternate phrasing that the system could not fully reconcile (Figure 2).

We conducted a detailed error analysis to understand the limited impact of the vision agent on parameters beyond Cmin and Cmax. In 19 test articles, figures and tables relevant to these parameters were rare (1 figure and 1 table per paper), and the text extraction agent already captured most numerical data (e.g., Km, Vmax). Thus, the vision agent's main value lies in extracting Cmin and Cmax when they appear only in figures. However, its performance drops with partially labeled or ambiguous plots, limiting its broader utility. Improving figure interpretation and label clarity remains an open challenge for expanding its effectiveness. While the vision agent can provide critical data when it appears solely on figures, further improvements in figure interpretation and label disambiguation are necessary to boost its performance on parameters beyond Cmin and Cmax.

Furthermore, using the developed tool, we expanded the existing nanozyme database to 100 samples from 42 articles and automatically identified seven articles lacking experimental measurements, which were filtered out during processing. The system's ability to detect papers without quantitative data eliminates the need to screen unsuitable articles manually. The total processing of 49 articles took 2 hours and 48 minutes. Thus, nanoMINER allows us to rapidly expand the existing state-of-the-art database through automated data extraction and validation, opening new research avenues in design and discovery of new materials.

4 Conclusion, Discussion, and Future Work

NanoMINER represents the first-of-its-kind multi-agent solution for automated nanomaterials data extraction. The system accurately and rapidly extracts structured data from complex nanomaterials and nanozyme literature. By integrating NLP, computer vision, and NER techniques, we have achieved significant extraction quality—up to 0.98 precision for nanozyme-specific kinetic parameters and near-perfect extraction of chemical formulas and other properties, as evidenced by near-zero normalized Levenshtein distances. Furthermore, the dramatic reduction in processing time (approximately 3.5 minutes per article versus roughly 90 minutes manually) addresses a critical bottleneck in materials research. While nanoMINER effectively automates this complex and time-consuming process in many cases, we acknowledge certain limitations and opportunities for improvement in our approach.

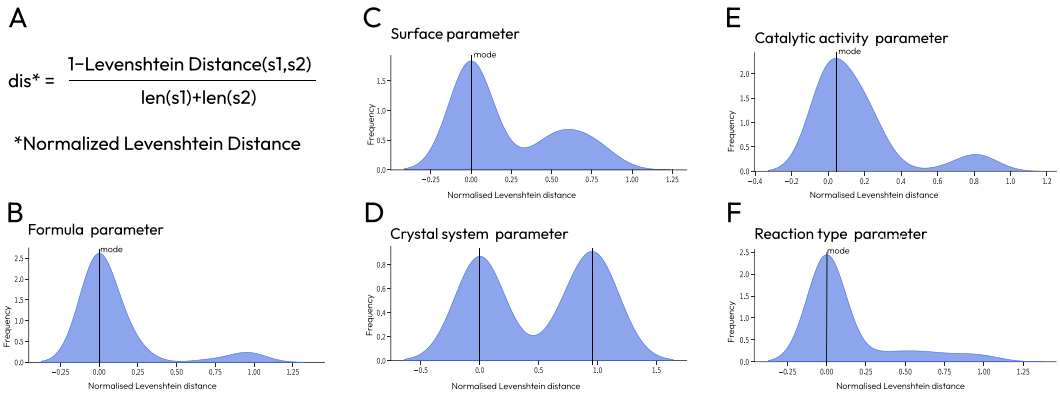

Figure 2: The performance of categorical nanozyme parameter extraction using normalized Levenshtein distance (A) is depicted for chemical formula (B), crystal system (C), surface molecule (D), catalytic activity (E) and reaction type (F) parameters.

The acceleration in data collection provided by nanoMINER has transformative implications for discovering new materials. The rapid transition from unstructured scientific texts to curated datasets will facilitate hypothesis generation and testing. Comprehensive materials databases will make it possible to implement AI-first approaches to materials design. Despite these advancements, several challenges remain. First, our analysis shows that visual data extraction does not significantly improve extraction accuracy for general parameters beyond Cmin and Cmax. A deeper breakdown of when figure processing is critical—such as in cases where information is exclusively present in charts or supplementary information—could help refine our approach. The system's generalizability to highly varied journal styles, noisy scans, or low-quality OCR outputs requires further evaluation. Future work should assess robustness under real-world conditions, including inconsistencies in formatting and document quality.

Another key area for improvement is validation and feedback mechanisms. Currently, nanoMINER focuses on producing structured outputs, but incorporating real-time verification—such as cross-checking extracted values with known material constraints—could further enhance reliability. Adaptive feedback loops, where models iteratively refine their outputs based on confidence scores or external validation sources, represent a promising direction for future development.

Addressing these challenges can further optimize nanoMINER for scalability, accuracy, and adaptability, ensuring its effectiveness across a wider range of scientific literature and application domains. More generally, future studies should explore integration of dynamic feedback loops within multi-agent frameworks, enabling real-time adjustments during data extraction. Furthermore, incorporating anomaly detection and on-the-fly data correction will pave the way for fully autonomous and adaptive extraction pipelines. These developments will not only streamline workflows in materials science but also extend to other domains such as biomedicine, environmental sciences, and more.

In conclusion, our work with nanoMINER demonstrates both the significant progress and substantial remaining challenges in automated scientific data extraction. By focusing research efforts on the identified open problems, particularly in scientific figure understanding, multimodal integration, validation, and format generalization, the community can develop tools that dramatically accelerate materials discovery through truly comprehensive and reliable data extraction.

ACKNOWLEDGMENTS

The authors thank the Priority 2030 Federal Academic Leadership Program for financial support.

REFERENCES

Evaluating gpt-4v (gpt-4 with vision) on detection of radiologic findings on chest radiographs. Radiology, 2024.

Introducing meta llama 3: The most capable openly available llm to date, 2024.

L. Foppiano, G. Lambard, T. Amagasa, and M. Ishii. Mining experimental data from materials science literature with large language models: an evaluation study. *Science and Technology of Advanced Materials: Methods*, 4:2356506, 2024.

A. Q. Jiang et al. Mistral 7b, 2023.

OpenAI. Gpt-4o system card, 2024.

A. Radford et al. Language models are unsupervised multitask learners.

J. Razlivina, A. Dmitrenko, and V. Vinogradov. Ai-powered knowledge base enables transparent prediction of nanozyme multiple catalytic activity. *J. Phys. Chem. Lett.*, 15:5804–5813, 2024.

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016. doi: 10.1109/CVPR.2016.91.

X. Wang, S. L. Huey, R. Sheng, S. Mehta, and F. Wang. Scidasynth: Interactive structured knowledge extraction and synthesis from scientific literature with large language model, 2024.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357, 2024.

A Appendix

A.1 Performance Metrics

To rigorously evaluate our extraction system, we employ several complementary metrics that assess different aspects of performance. All evaluations were conducted on our manually annotated test dataset, comprising 19 papers with 1-6 experiments per paper. Each experiment corresponds to one measurement for a parameter, and measurements were repeated 100 times to ensure statistical robustness.

We define our evaluation criteria at the *experiment* level rather than at the parameter level:

- **True Positive (TP):** An experiment where all extracted parameters correctly match the gold-standard annotation.
- **False Positive (FP):** An experiment that either does not exist in the gold-standard or contains one or more incorrect parameters.
- **False Negative (FN):** A valid experiment from the gold-standard that the system fails to extract.
- **True Negative (TN):** Correct identification that no experiment is present when none exists in the gold-standard.

A.1.1 Precision

For each paper i, precision quantifies the proportion of correctly extracted experiments relative to all extracted experiments:

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{1}$$

The overall precision is calculated as the average across all papers:

$$\text{Precision} = \frac{1}{N}\sum_{i=1}^{N} \text{Precision}_i = \frac{1}{N}\sum_{i=1}^{N} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{2}$$

A.1.2 RECALL

For each paper i, recall measures the proportion of correctly extracted experiments relative to all experiments that should have been extracted:

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{3}$$

The overall recall is calculated as the average across all papers:

$$\text{Recall} = \frac{1}{N}\sum_{i=1}^{N}\text{Recall}_i = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{4}$$

A.1.3 NORMALIZED LEVENSHTEIN DISTANCE

The Levenshtein distance $L(s_1, s_2)$ measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform string s_1 into string s_2. To facilitate interpretation across strings of different lengths, we utilized the normalized Levenshtein distance:

$$\text{NLD}(s_1, s_2) = 1 - \frac{L(s_1, s_2)}{\max(|s_1|, |s_2|)} \tag{5}$$

Where $|s_1|$ and $|s_2|$ are the lengths of strings s_1 and s_2, respectively. This normalization produces values in the range $[0, 1]$, where:

- NLD $= 0$: Perfect match (identical strings)
- NLD $= 1$: Maximum dissimilarity (completely different strings)

This formulation is particularly useful when comparing strings of significantly different lengths, as it provides a more balanced similarity assessment.

A.2 NER AGENT TRAINING AND HYPERPARAMETER TUNING

We used pre-trained language models (PLMs) for further fine-tuning in our system. We carefully tuned the accumulation step, warmup steps, optimizer, and learning rate to achieve the best results with a Tree-structured Parzen Estimator (TPE). A Bayesian optimization technique aims to find the optimal set of hyperparameters that maximize (or minimize) an objective function. The accumulation step determines the number of gradients combined before updating the model weights, which can help stabilize training and reduce memory consumption. Warmup steps gradually increase the learning rate during the initial training phase, preventing the model from diverging due to large gradients. Llama-3-8B was finetuned using a cross-entropy loss function, and the training process was carefully monitored to ensure convergence and optimal performance. We employed a learning rate scheduler with a linear warmup period followed by cosine annealing, which has been shown to improve generalization. As shown in Figure 3, the warmup period was set to 334 steps, with an initial learning rate of 1.4e-6, gradually increasing to a maximum of 3e-6. The learning rate was then annealed according to the cosine annealing schedule over the remaining training steps. We used the AdamW 41 optimizer with a weight decay of 0.01 and a batch size of 1. The accumulation step was set to 16, allowing for more efficient use of GPU memory while maintaining a larger adequate batch size. Figure 3 illustrates the learning rate schedule and the corresponding training loss throughout training. As shown in Figure 3B, the model achieved convergence after approximately 1,500 steps, with the validation loss stabilizing satisfactorily. Mistral-7B was fine-tuned similarly but with different hyperparameters.(Figure 4) Specifically, it was also trained using cross-entropy loss. Still, it employed approximately 368 warmup steps. The initial learning rate was set to 5e-7, and the training process utilized a batch size of 4 with an accumulation step of 4. Additionally, we employed the Adafactor optimizer 42 for the fine-tuning process. These adjustments in the training configuration allowed for a more tailored approach to fine-tuning the Mistral-7B model, potentially impacting its performance and adaptation to the specific task at hand.

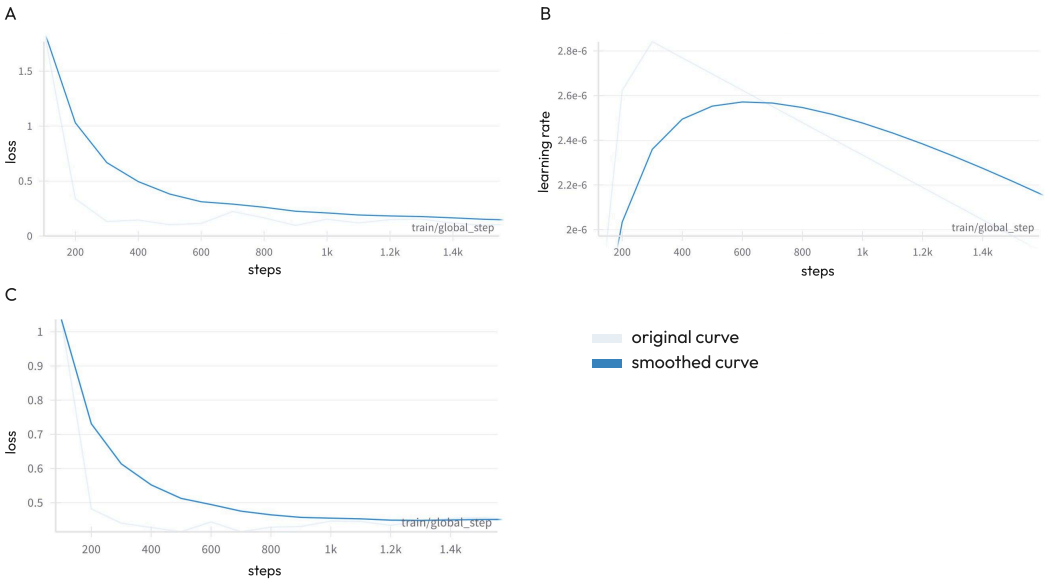

Figure 3: Llama-3-8B tuning process with original and smoothed curves: A) training loss function, B) learning rate, C) validation loss function.

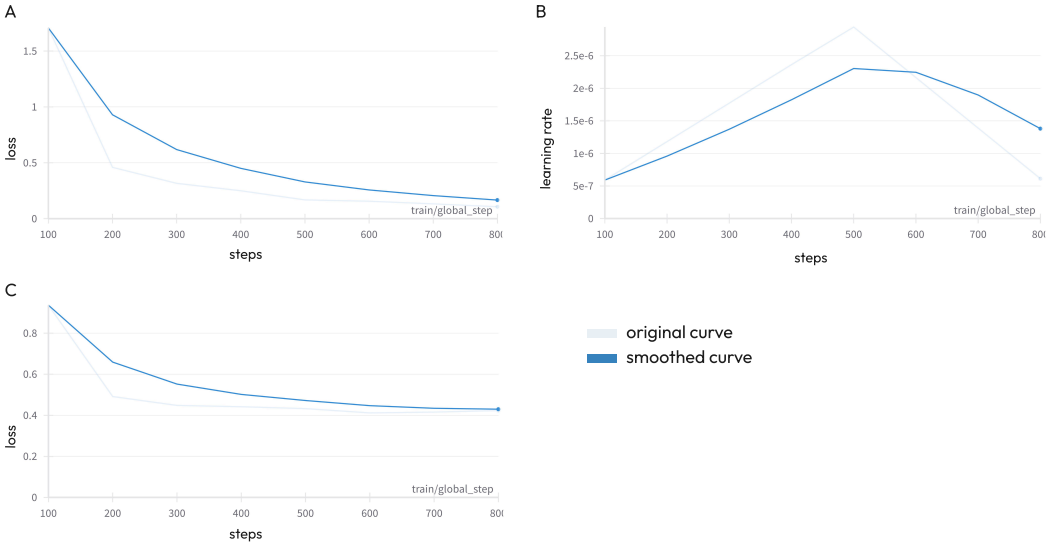

Figure 4: Mistral-7B tuning process with original and smoothed curves: A) training loss function, B) learning rate, C) validation loss function.