Domain-Specific Fine-Tuning of GPT-2 for Emulating Academic Communication Styles

Keer Ni
UC Davis Student
knni@ucdavis.edu

Muhammad Reza
UC Davis Student
msreza@ucdavis.edu

Joshua Sun
UC Davis Student
zysun@ucdavis.edu

Abstract

Large language models (LLMs) have shown strong general-purpose language generation capabilities, but often lack the stylistic nuance required for academic communication. This project explores domain-specific fine-tuning as a means to adapt LLMs for more formal, scholarly use cases. We fine-tune GPT-2 (124M) using full-parameter training and Mistral-7B using QLoRA on a subset of academic abstracts from the arXiv dataset. Our preprocessing pipeline included text cleaning, abstract filtering, and prompt formatting. We evaluate the models using a mix of qualitative output analysis and human judgment, with a focus on academic tone, coherence, and factual consistency. While both models showed marginal improvements in structure and style, their performance remained limited—frequent hallucinations, generic phrasing, and shallow logic were common. Our results suggest that light fine-tuning alone is insufficient to fully align LLMs with academic expectations. This study highlights the challenges of adapting LLMs for scholarly use and lays the groundwork for further refinement through more targeted techniques and larger training resources.

1 Introduction

Large Language Models (LLMs) such as GPT-2 have shown impressive capabilities in generating fluent, general-purpose text. However, when applied to academic contexts, these models frequently fail to produce stylistically appropriate responses. The generated output may appear too casual for formal settings or overly rigid and artificial—what some might describe as "robotic." This lack of stylistic nuance limits the effectiveness of LLMs in scenarios where academic tone, structure, and clarity are crucial.

There is a growing need for models that can generate high-quality academic language. Potential applications include intelligent tutoring systems, research drafting tools, and writing assistants for educational and scholarly communication. A model better aligned with academic standards could enhance the productivity and communication of students, educators, and researchers alike.

While pretrained models like GPT-2 are effective generalists, they often fall short in specialized domains such as academia, where precise language, logical flow, and formal tone are expected. Our project addresses the challenge of adapting GPT-2

to emulate the speaking and writing style typical of academic professionals.

The core problem is to develop a domain-adapted model that improves stylistic fidelity to academic communication. We approach this by fine-tuning GPT-2 on a corpus of academic texts. This adaptation, however, brings forth several challenges: preserving the model's general language fluency, avoiding overfitting to narrow stylistic patterns, and developing objective methods to evaluate how well the model matches academic style.

The primary goal of this project is to fine-tune an existing LLM on academic writing data to produce more stylistically accurate and fluent academic text. Specifically, we aim to:

- Improve the model's ability to emulate formal scholarly tone and structure.
- Evaluate the quality of generated text using both automatic metrics and human judgment.
- Analyze the balance between domain adaptation and generalization, ensuring the model remains useful beyond the fine-tuning corpus.
- Contribute a replicable method for improving stylistic alignment in LLMs for educational and research-focused applications.

By addressing these goals, we hope to build a more reliable and effective academic writing assistant, advancing the utility of LLMs in scholarly contexts.

2 Methods

2.1 Problem Statement

While pretrained language models like GPT-2 are highly capable in general-purpose text generation, they often lack the stylistic precision needed for effective academic communication. Academic writing demands more than grammatical correctness—it requires formal tone, structured reasoning, and domain-specific terminology that general models are not explicitly trained to reproduce.

Our project investigates how domain-specific fine-tuning can improve an LLM's ability to emulate academic tone, structure, and fluency. Formally, given a pretrained language model M and an academic text corpus D, our objective is to learn a fine-tuned model M' such that the output distribution $P_{M'}(y \mid x)$ better matches the style and content typical of scholarly communication than the baseline distribution $P_M(y \mid x)$.

Several key challenges must be addressed to achieve this adaptation:

- Maintaining general language fluency: Ensuring that the fine-tuned model remains versatile and coherent in language use without becoming overly domain-specific.
- Preventing overfitting: Avoiding excessive memorization of the training corpus, which could lead to repetitive or unnatural output.
- Evaluating "academic style": Defining and measuring academic tone and structure objectively, which

is inherently qualitative and context-dependent.

By formally defining this adaptation task, we aim to provide a principled foundation for improving stylistic fidelity in LLMs used in academic settings.

2.2 Data Collection and Preprocessing

We use the arXiv Dataset curated by Cornell University and available on Kaggle, which contains metadata and article content for over 2.7 million research papers from the arXiv repository. The dataset is structured in JSON format, including metadata such as title, authors, article ID, and abstract. For the purposes of our fine-tuning task, we exclusively focus on the abstract field, which captures a compact, stylistically rich form of academic communication. We randomly sampled 500,000 abstracts from the full dataset. This number was chosen to provide a substantial training corpus while remaining computationally feasible for our available resources.

Mathematical expressions—often denoted in LaTeX by dollar signs (e.g., $...$)—were removed. These symbols frequently obscure natural language patterns and introduce noise that may confuse the language model. Length Standardization: Abstracts below a minimum character threshold or with missing/incomplete content were discarded to ensure linguistic substance and consistency across training samples. All text was tokenized using GPT-2's Byte-Pair Encoding (BPE) tokenizer to match the model's expected input representation.

It is important to note that no train-validation-test split was used. In contrast to traditional machine learning workflows where automatic metrics (e.g., accuracy, loss) guide model selection, our primary evaluation is human-centered. The model's performance is judged qualitatively based on fluency, style, and academic tone—making traditional validation less relevant. Furthermore, each abstract was treated independently and weighted equally during training. We did not perform any document-level comparisons or introduce weighting schemes based on domain, subfield, or citation count.

2.3 Methodology Overview

Our overall pipeline follows a standard domain adaptation workflow:

$$\text{Pretrained Model } M \rightarrow \text{Academic Dataset } D \rightarrow \text{Fine-Tuned Model } M'$$

We begin with a pretrained language model and fine-tune it on a large corpus of academic text (research paper abstracts) in order to adapt its output style toward academic communication norms. Two models were used in this study: GPT-2 and Mistral-7B, each employing different optimization strategies and fine-tuning methods.

For GPT-2, we used a full fine-tuning approach, where all model parameters are updated during training. This method is common for smaller models and allows for maximum control over stylistic adaptation. However, it comes at the cost of increased training time and memory requirements. For the larger Mistral-7B model, we applied Low-Rank Adaptation (LoRA), which updates only a subset of the model's parameters. LoRA drastically reduces both

time and space complexity, making it ideal for adapting larger models while maintaining general language capabilities.

GPT-2 was fine-tuned using the cross-entropy loss, which measures the likelihood of predicting the correct next token. Mistral-7B was trained using a causal language modeling (CLM) loss, which also operates on next-token prediction in a left-to-right fashion. That said, our focus was not on loss minimization. While these loss functions are standard for LLM training, human evaluation was prioritized far above automatic metrics. Our primary concern was not whether the model performed well numerically, but whether its output sounded convincingly academic.

A key part of our methodology involves domain transfer. The model is trained on a formal domain—academic abstracts from scientific papers—but evaluated on different types of academic tasks, such as generating or refining essay theses. This setup allows us to assess whether stylistic adaptation transfers across academic subfields and genres, rather than being limited to the training domain. We rely on human evaluators to judge model outputs in terms of stylistic formality, logical sentence construction, and proper coherence of text. This evaluation allows us to test not just technical adaptation, but how well the model performs under real-world academic communication tasks.

3 Experiments

3.1 Setup and Details

To evaluate the effectiveness of domain-specific fine-tuning for academic communication, we fine-tuned two different language models: GPT-2 (124M parameters) and Mistral-7B, using different tuning strategies and training configurations to reflect their computational requirements and capacities.

GPT-2 (124M) was fine-tuned using full-parameter training for approximately 1.2 epochs. This approach updates all weights in the model and is suitable for smaller architectures where training time and memory are manageable. Mistral-7B was fine-tuned using Quantized LoRA (QLoRA), a parameter-efficient method that updates only low-rank matrices within transformer layers. This configuration allowed us to adapt a significantly larger model while keeping computational demands within reach. Training was conducted for 1 epoch.

For GPT-2, the training data consisted of research paper abstracts saved as a plain text .txt file, where each abstract was separated by a blank line (\n\n). The text was tokenized using GPT-2's native Byte-Pair Encoding (BPE) tokenizer. The model was trained in a standard left-to-right fashion, predicting the next token in the sequence.

For Mistral-7B, we used a prompt-based format designed to emulate a realistic academic writing task. Each training example followed the structure:

Prompt: "Write an academic paragraph given the title:" + [title]

Target: [abstract]

The combined prompt and target were tokenized and passed through the model for causal language modeling. This format

encouraged the model to generate outputs that match academic tone and structure conditioned on a research title.

Training was performed on a single NVIDIA RTX 5070 Ti GPU. Due to the significant difference in model sizes and tuning strategies, training times varied considerably.

Model	Epochs	Batch Size	Learning Rate	Loss Weights	GPU	Time
GPT-2 (124M)	1.2	16	Default	0.1 / 0.3 / 0.5	RTX 5070 Ti	~2 hours
Mistral-7B (QLoRA)	1	8	2e-4	0.1 / 0.3 / 0.5	RTX 5070 Ti	~60 hours

Fig. 1: GPT-2 vs Mistral-7B Model Training

Loss weights refer to the mixing coefficients used in loss calculation (if applicable) for balancing between objectives (e.g., matching fluency vs. stylistic targets).

To facilitate demonstration, both models were deployed as APIs via Hugging Face, allowing for real-time testing and human feedback.

3.2 Results and Analysis

To assess the effectiveness of our fine-tuning strategies, we qualitatively examined outputs from both the GPT-2 (124M) and Mistral-7B (QLoRA) models.

Prompt: LoRA

GPT-2 Output: “LoRA) method to solve the linear problem of constructing a convex hull for an arbitrary non-degenerate group , in this paper we propose and prove that, under certain conditions on , such as the dimension of its manifold with respect (or), the -norm is bounded above by . We also provide a characterization of the -norm in terms both of the linearity condition and uniqueness of the subgroup”

Mistral-7B Output: “We present LoRA (Loosely Regularized Adapters), a novel approach to fine-tune large language models (LLMs) for in-context learning tasks. Unlike conventional fine-tuning methods that modify the model's weights, LoRA only adds a few hundred additional trainable parameters to the model, reducing the risk of overfitting and maintaining the model's core functionality. Our experiments show that LoRA achieves comparable performance to conventional fine-tuning, while requiring significantly fewer trainable parameters…”

Overall, the generated outputs from both models exhibited significant limitations. The GPT-2 output often adopted a formal tone but lacked syntactic correctness, clarity, and coherence. It frequently produced fragmented or mathematically nonsensical sentences, likely due to limited capacity and a lack of structure in the input format. The Mistral-7B output, while more fluent and on-topic, remained overly generic. It often reproduced templated descriptions without meaningful elaboration, indicating shallow understanding. The model tended to prioritize surface-level academic phrasing over factual precision or argumentative structure.

Through repeated evaluations, we observed the following recurring error types across both models:

- Syntactic Issues: Ungrammatical or malformed sentences, particularly from GPT-2.

- Logical Inconsistencies: Contradictory or vague claims that break the internal coherence of the text.
- Hallucinations: Frequent fabrication of facts, methods, and results, especially when responding to less common or underspecified prompts.

These shortcomings highlight the difficulty of aligning model outputs with the nuanced expectations of academic writing through limited fine-tuning.

To address these limitations, we propose the following improvements in future iterations:

- Scale Up the Dataset: Rather than sampling 500,000 abstracts, we plan to utilize the full 2.7 million-entry arXiv dataset to better capture domain-specific patterns.
- Use Stronger Base Models: Leveraging more capable base models, such as Deepseek or Mixtral, may improve factual grounding and stylistic richness.
- Refine Prompt Design: Instead of generic prompts, we aim to customize prompts to encourage factual content generation—not just mimicry that follows academic tone.

To summarize, while our fine-tuning procedures showed slight improvements in stylistic formality and structure, neither model consistently produced high-quality academic output. Error patterns suggest a need for deeper training, more expressive prompts, and stronger base models to achieve reliable academic tone and factual coherence

4 Related Work

Our work builds on recent advances in domain adaptation and parameter-efficient fine-tuning for large language models.

Diao et al. (2023) propose Mixture-of-Domain-Adapters, a modular framework for injecting domain-specific knowledge without full model retraining. Their approach enables multi-domain adaptation while maintaining general fluency—an idea that informs our strategy for adapting to academic language without sacrificing the model's broader linguistic capabilities.

Wang et al. (2024) introduce Role Prompting, which uses task-specific prompts to steer model behavior in different domains. This method helps prevent catastrophic forgetting while promoting stylistic consistency, closely aligning with our objective of balancing domain specificity with generalization across academic genres.

Zhang et al. (2023) demonstrate that fine-tuning in low-rank subspaces can be highly effective, supporting the use of parameter-efficient methods such as LoRA. Their findings motivate our use of QLoRA for adapting large-scale models like Mistral-7B, allowing us to reduce resource demands while achieving meaningful stylistic adaptation.

Together, these works highlight the importance of modularity, prompt conditioning, and efficient optimization—all of which directly inform our methodology for domain-specific fine-tuning in academic communication.

5 Conclusion and Future Work

This project explored the domain-specific fine-tuning of large language models to improve their ability to generate academic-style text. Our experiments with GPT-2 (using full fine-tuning) and Mistral-7B (using QLoRA) demonstrated modest improvements in stylistic fidelity and tone. GPT-2 outputs reflected more formal academic phrasing, while Mistral-7B maintained coherence with fewer trainable parameters, showcasing the efficiency of parameter-efficient adaptation.

These models have promising applications in educational contexts, such as AI writing assistants for students and researchers, tools for lecture summarization, and academic tutoring aids. Importantly, they can also help improve accessibility for non-native English speakers navigating scholarly environments.

However, the potential for misuse must be acknowledged. Risks include plagiarism, ghostwriting, and the generation of plausible but factually incorrect or biased content. Additionally, our current models lack citation grounding and do not support verifiable sourcing, which limits their reliability for research purposes.

Future work will involve integrating citation tracking, retrieval-augmented generation, and reference validation mechanisms to ensure factual grounding. Expanding support to multilingual and underrepresented academic domains will also be crucial. Overall, this project highlights both the opportunities and limitations of adapting LLMs for academic use, underscoring the importance of responsible development and transparent deployment.

References

Shizhe Diao, Tianyang Xu, Ruijia Xu, Jiawei Wang, and Tong Zhang. 2023. Mixture-of-Domain-Adapters: Decoupling and Injecting Domain Knowledge to Pre-trained Language Models' Memories. *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5113–5129, Association for Computational Linguistics.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better Zero-Shot Reasoning with Role-Play Prompting. *In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Association for Computational Linguistics.

Zhong Zhang, Bang Liu, and Junming Shao. 2023. Fine-tuning Happens in Tiny Subspaces: Exploring Intrinsic Task-specific Subspaces of Pre-trained Language Models. *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1713, Association for Computational Linguistics.