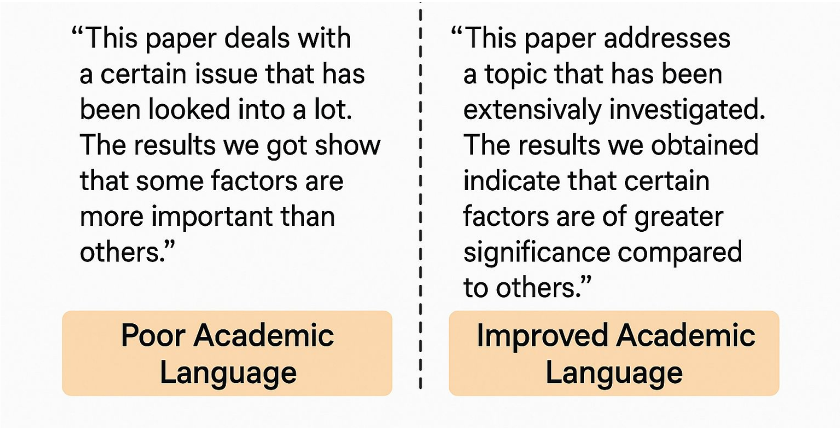
Domain-Specific Fine-Tuning of GPT-2 for Emulating Academic Communication Styles

Keer Ni, Muhammad Reza, Joshua Sun

UC DAVIS
COMPUTER SCIENCE

Motivations

- LLMs such as GPT-2 struggle with maintaining proper academic tone
- Output could either be too casual or too “robotic”
- There is a real-world need for strong academic language generation
    - Education, research assistance, AI tutors, etc.



“This paper deals with a certain issue that has been looked into a lot. The results we got show that some factors are more important than others.”

Poor Academic Language

“This paper addresses a topic that has been extensivaly investigated. The results we obtained indicate that certain factors are of greater significance compared to others.”

Improved Academic Language

Problem Definition

Problem Statement

Large language models like GPT-2 often lack the stylistic precision needed for academic communication. Our project explores how domain-specific fine-tuning can improve the ability of LLMs to emulate scholarly tone, structure, and fluency.

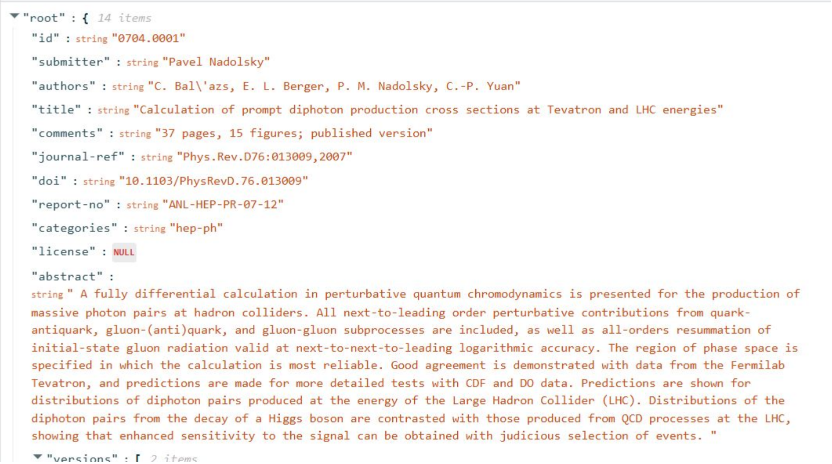
- We wish to build a model that emulates proper academic writing style
- Will achieve this by fine-tuning an already existing LLM on academic writing data
- Key challenge areas:
  - Maintaining general fluency
  - Preventing overfitting
  - Evaluating “academic style”

Project Goals

- Fine-tune some LLM on academic text
- Improve stylistic fidelity
- Offer a proper evaluation of the quality of the model
- Investigate domain generalization and adaptation efficiency

Dataset: Structure & Information

- “arXiv Dataset” by Cornell University et al. (found on Kaggle)
  - Article information/metadata from arXiv research papers in JSON format
  - Title, authors, article id, abstract, etc.
- ~2.7M entries/research papers
- “abstract” is the important variable



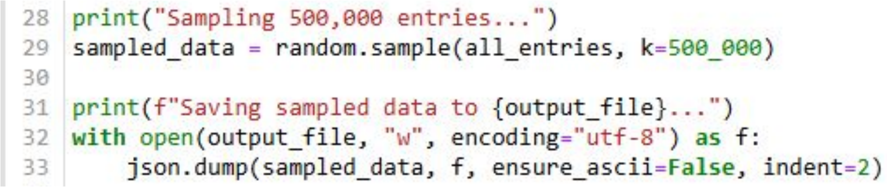
```
▼ "root" : { 14 items
    "id" : string "0704.0001"
    "submitter" : string "Pavel Nadolsky"
    "authors" : string "C. Bal\'azs, E. L. Berger, P. M. Nadolsky, C.-P. Yuan"
    "title" : string "Calculation of prompt diphoton production cross sections at Tevatron and LHC energies"
    "comments" : string "37 pages, 15 figures; published version"
    "journal-ref" : string "Phys.Rev.D76:013009,2007"
    "doi" : string "10.1103/PhysRevD.76.013009"
    "report-no" : string "ANL-HEP-PR-07-12"
    "categories" : string "hep-ph"
    "license" : NULL
    "abstract" :
    string " A fully differential calculation in perturbative quantum chromodynamics is presented for the production of
    massive photon pairs at hadron colliders. All next-to-leading order perturbative contributions from quark-
    antiquark, gluon-(anti)quark, and gluon-gluon subprocesses are included, as well as all-orders resummation of
    initial-state gluon radiation valid at next-to-next-to-leading logarithmic accuracy. The region of phase space is
    specified in which the calculation is most reliable. Good agreement is demonstrated with data from the Fermilab
    Tevatron, and predictions are made for more detailed tests with CDF and DO data. Predictions are shown for
    distributions of diphoton pairs produced at the energy of the Large Hadron Collider (LHC). Distributions of the
    diphoton pairs from the decay of a Higgs boson are contrasted with those produced from QCD processes at the LHC,
    showing that enhanced sensitivity to the signal can be obtained with judicious selection of events. "
  ▼ "versions" : [ 2 items
```

Dataset: Preprocessing

- Dataset downloaded into cache via Python’s kagglehub library

- Size of data reduced to 500,000 samples (via random sampling)

- Sampled data saved into new file in the project directory

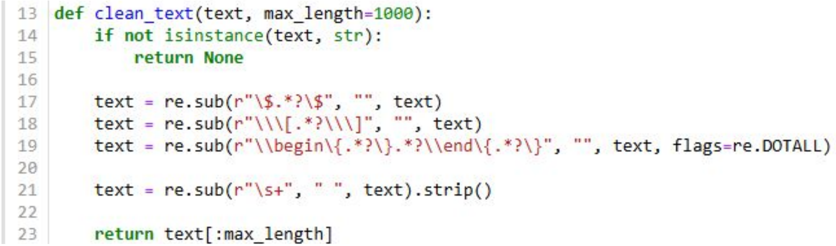


```
print("Sampling 500,000 entries...")
sampled_data = random.sample(all_entries, k=500_000)

print(f"Saving sampled data to {output_file}...")
with open(output_file, "w", encoding="utf-8") as f:
    json.dump(sampled_data, f, ensure_ascii=False, indent=2)
```

Dataset: Preprocessing

- Text filtering: Removing mathematical symbols from the data
  - In LaTeX, symbol notation is tedious and could hinder the model’s understanding
- Length Standardization: Setting minimum threshold for abstract length
  - Removing rows with incomplete or missing text data

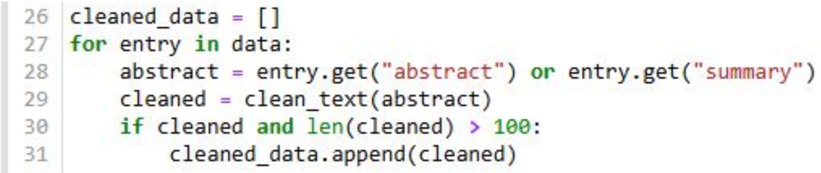


```python
def clean_text(text, max_length=1000):
    if not isinstance(text, str):
        return None

    text = re.sub(r"\$.*?\$", "", text)
    text = re.sub(r"\\\[.*?\\\]", "", text)
    text = re.sub(r"\\begin\{.*?\}.*?\\end\{.*?\}", "", text, flags=re.DOTALL)

    text = re.sub(r"\s+", " ", text).strip()

    return text[:max_length]
```



```python
cleaned_data = []
for entry in data:
    abstract = entry.get("abstract") or entry.get("summary")
    cleaned = clean_text(abstract)
    if cleaned and len(cleaned) > 100:
        cleaned_data.append(cleaned)
```

Dataset: Some Considerations

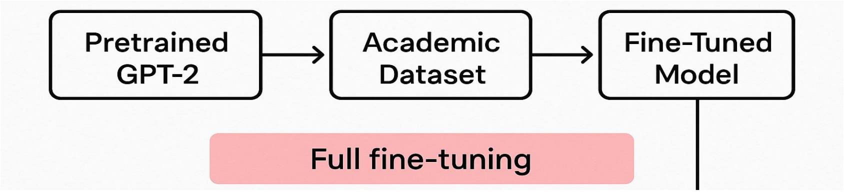
- All samples are used for training (no splits)
  - Human evaluation >> Loss

- The amount of samples (500,000) was chosen with discretion

- Article abstracts were not compared to one another
  - Each abstract was weighted equally in the model

Methodology Overview

- Pipeline: Pretrained model → Academic Dataset → Fine-Tuned Model
- GPT-2 relies on cross-entropy loss for fine-tuning
- Mistral-7B relies on CLM (Casual Language Modeling) loss
- But… Human evaluation >> loss



Full fine tune & LoRA

- **Full fine tune**: Fine-tunes all the parameters in the model
  - Often used for smaller models
  - Better results, but takes much longer and needs much more resources

- **LoRA (Low-Rank Adaptation)**: Fine-tunes only a subset of parameters
  - Greatly reduces time and space complexity
  - Ideal for fine-tuning especially large models

Domain Transfer

- Training domain example: Research paper excerpts
- Testing domain example: Essay thesis
- Goal: Assess cross-subfield stylistic adaptation

Experiment Setup

Model Configurations:

- **GPT-2 124M**: Full fine-tuning for **1.2 epochs**
- **Mistral 7B**: Parameter-efficient tuning using **QLoRA**, trained for **1 epoch**

Hardware:

- **GPU**: NVIDIA RTX **5070 Ti**

Training Time:

- **GPT-2**: ~2 hours
- **Mistral**: ~60 hours

Model	Epochs	Batch Size	LR	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

Experiment details

GPT-2: Abstract is converted into plain text, saved as a .txt file, where each abstract is separated by a empty line (\n). Then the plain text gets tokenized and feed into GPT-2 to make it predict the next token.

Mistral: Prompt tuning, where the prompt is always “Write an academic paragraph given the title.” The input is the title of and the output is the abstract. Then the whole prompt gets tokenized and mapped to train the model.

Results

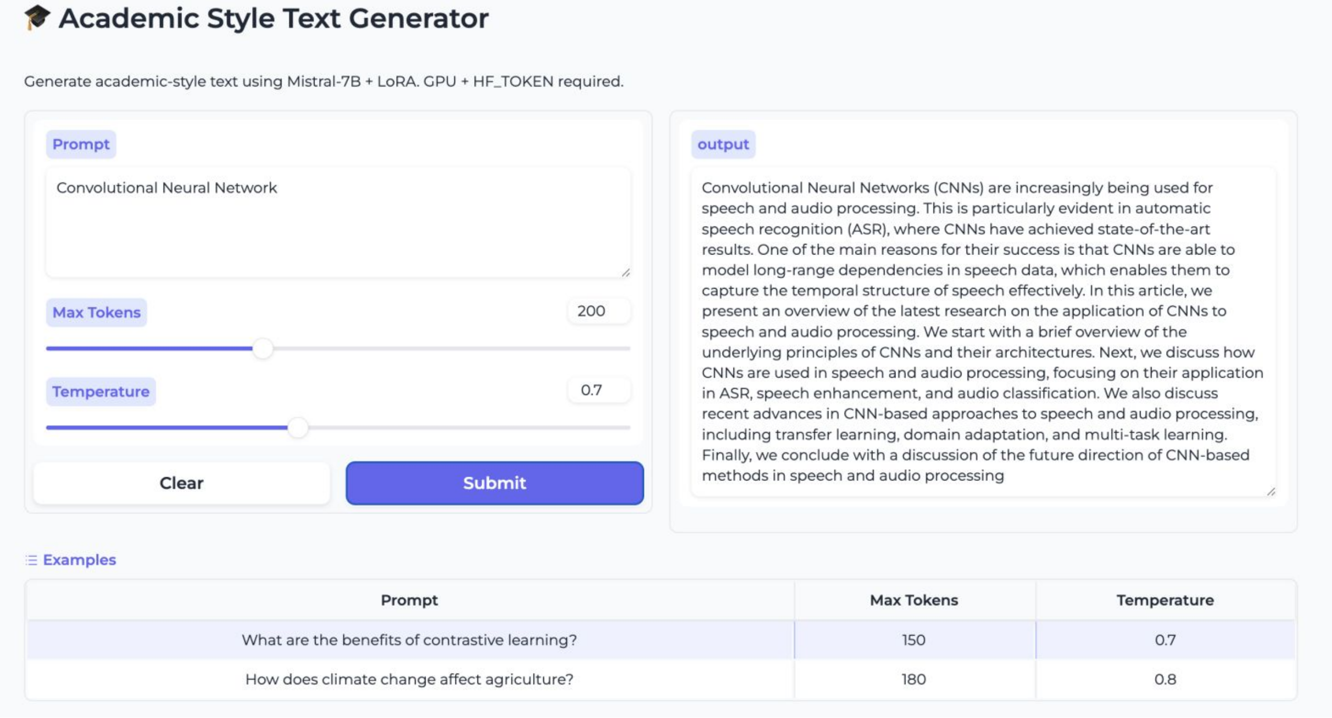
Prompt: LoRA

Output (GPT-2): 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

Output (Mistral): 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. LoRA also offers additional benefits such as faster training times, reduced memory requirements, and better generalization to unseen data. We demonstrate the effectiveness of LoRA on a variety of in-context learning tasks, including

Live Demonstration

GPT-2 and Mistral models



🎓 Academic Style Text Generator

Generate academic-style text using Mistral-7B + LoRA. GPU + HF_TOKEN required.

Prompt: Convolutional Neural Network

Max Tokens: 200

Temperature: 0.7

output:

Convolutional Neural Networks (CNNs) are increasingly being used for speech and audio processing. This is particularly evident in automatic speech recognition (ASR), where CNNs have achieved state-of-the-art results. One of the main reasons for their success is that CNNs are able to model long-range dependencies in speech data, which enables them to capture the temporal structure of speech effectively. In this article, we present an overview of the latest research on the application of CNNs to speech and audio processing. We start with a brief overview of the underlying principles of CNNs and their architectures. Next, we discuss how CNNs are used in speech and audio processing, focusing on their application in ASR, speech enhancement, and audio classification. We also discuss recent advances in CNN-based approaches to speech and audio processing, including transfer learning, domain adaptation, and multi-task learning. Finally, we conclude with a discussion of the future direction of CNN-based methods in speech and audio processing

Examples

Prompt	Max Tokens	Temperature
What are the benefits of contrastive learning?	150	0.7
How does climate change affect agriculture?	180	0.8

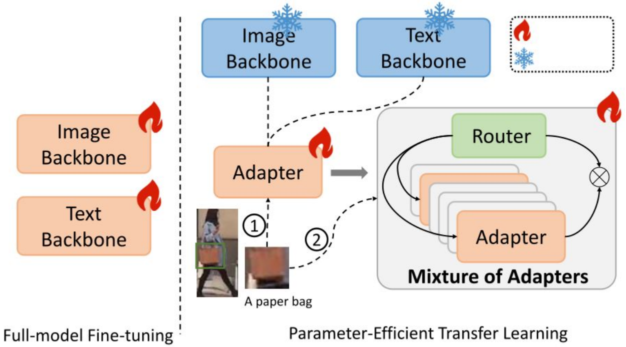
Error Analysis

- Categories of common errors:
  1. Syntactic issues
  2. Logic/consistency
  3. Hallucination (very often!)
- How we plan to address them in future iterations: More samples (use full 2.7M), use better base model (e.g. Deepseek), adjust prompt for each input so it learns facts instead of style-only

Related Work

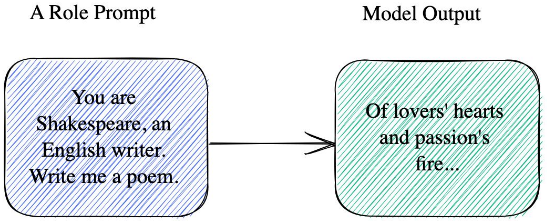
1. Mixture-of-Domain-Adapters: Diao et al. (ACL 2023)

- Introduce modular adapters for injecting domain-specific knowledge
- Enable multi-domain adaptation without full model retraining
- Inspire our approach to preserving general fluency across academic domains



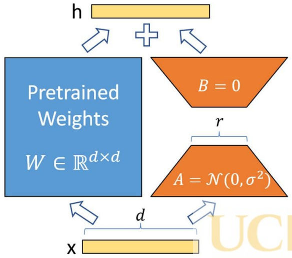
2. Role Prompting: Wang et al. (NAACL 2024)

- Use task-specific prompts to guide domain adaptation
- Prevent catastrophic forgetting while improving style consistency
- Aligns with our goal of balancing domain specificity and generalization


A Role Prompt
Model Output
You are Shakespeare, an English writer. Write me a poem.
Of lovers' hearts and passion's fire...

3. Fine-Tuning in Low-Rank Subspaces: Zhang et al. (ACL 2023)

- Show that fine-tuning is effective in low-dimensional parameter subspaces
- Justify our use of **QLoRA** for Mistral-7B as a parameter-efficient method
- Emphasize tuning efficiency over full model updates


h
Pretrained Weights
$W \in \mathbb{R}^{d\times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

UC DAVIS COMPUTER SCIENCE

Broader Impact

Educational Applications

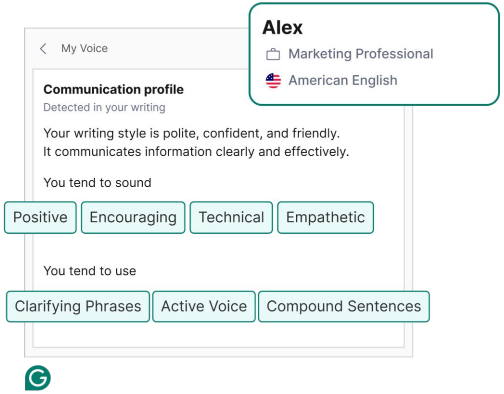
1. AI writing assistants for students and researchers
2. Tools for lecture summarization and academic tutoring
3. Improves accessibility for non-native speakers

Risks & Misuse

1. Potential for ghostwriting or plagiarism
2. Generation of fake but plausible academic content
3. Risk of stylistic bias toward overrepresented fields

Future Directions

1. Add citation grounding and source verification
2. Build safeguards: style checks, reference validation
3. Promote responsible and transparent deployment



Alex
Marketing Professional
American English

My Voice

Communication profile
Detected in your writing

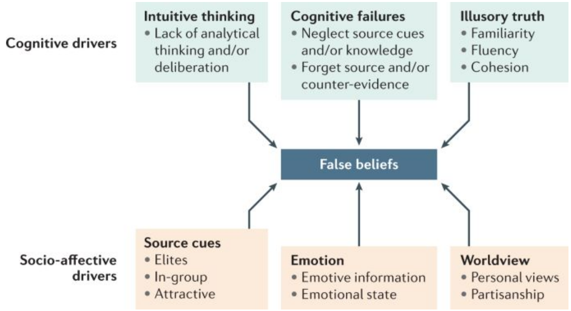
Your writing style is polite, confident, and friendly.
It communicates information clearly and effectively.

You tend to sound

Positive | Encouraging | Technical | Empathetic

You tend to use

Clarifying Phrases | Active Voice | Compound Sentences



Cognitive drivers

Intuitive thinking
- Lack of analytical thinking and/or deliberation

Cognitive failures
- Neglect source cues and/or knowledge
- Forget source and/or counter-evidence

Illusory truth
- Familiarity
- Fluency
- Cohesion

False beliefs

Socio-affective drivers

Source cues
- Elites
- In-group
- Attractive

Emotion
- Emotive information
- Emotional state

Worldview
- Personal views
- Partisanship

Limitations

Citation Grounding

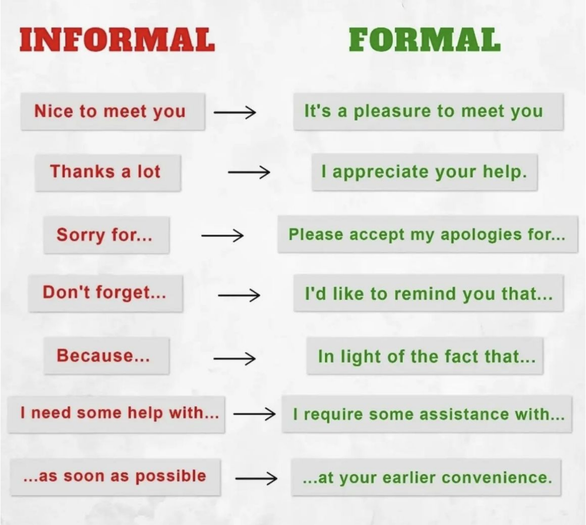
1. Current model does **not track or generate citations**
2. Outputs may lack verifiable sources

Style Generalization

1. **Style classifier** may underperform on **non-English** or informal academic texts

Planned Improvements

1. Add **retrieval-augmented generation** for factual grounding
2. Incorporate **citation modules** for reference accuracy



INFORMAL		FORMAL
Nice to meet you	→	It's a pleasure to meet you
Thanks a lot	→	I appreciate your help.
Sorry for...	→	Please accept my apologies for...
Don't forget...	→	I'd like to remind you that...
Because...	→	In light of the fact that...
I need some help with...	→	I require some assistance with...
...as soon as possible	→	...at your earlier convenience.

Conclusion

Key Outcomes

- Fine-tuned GPT-2 improved academic tone across multiple prompts
- QLoRA on Mistral-7B maintained style fidelity with fewer trainable parameters

Looking Ahead

- Incorporate **citation tracking** and **retrieval-augmented generation** for factual grounding
- Expand to **multilingual academic domains** and more diverse subfields
- Fine-tune and evaluate with **human-in-the-loop feedback**

Future Direction for Work (Example)

- TED Talk Transcript data (separate realm)
  - Vary in length
  - Much longer than research paper abstracts in general
  - Some TED Talks rely on humor/wit to engage audiences (not necessarily academic language)
  - Potential RQ could revolve around text for speeches/talks instead of academic language

Contribution of the Team Members

- Joshua: Proposal, Data preprocessing, Model training, Help with slides
- Keer: Proposal, API construction, Slides, Paper
- Muhammad: Data collection, Data preprocessing, Slides, Paper

Thank you for listening!

Q&A