Training Mr. Darcy: Fine-Tuning AI Models for Distinctive Speech Patterns

(code and datasets can be explored on GitHub)

Language models excel at capturing general linguistic patterns, yet consistently emulating a specific, stylistically nuanced voice remains challenging. This study investigates the feasibility of fine-tuning large language models-specifically GPT-2 and GPT-J-to fully adopt the distinctive speech patterns of Mr. Darcy from Jane Austen's *Pride and Prejudice*, rather than to imperfectly attempt imitation if asked to do so.

Fine-tuned models achieved approximately 70% higher BLEU-4 scores compared to baseline models when evaluated against authentic Darcy dialogue, suggesting meaningful stylistic improvement. However, these results hinge critically upon on the reliability of BLEU-4 as an indicator of stylistic similarity, an assumption which introduces potential ambiguity. This limitation highlights inherent constraints in the targeted fine-tuning process, and is addressed later on in greater detail.

Methodology

A significant hurdle in the endeavor to achieve consistent adherence to a precisely defined speech pattern is the scarcity of dialogue directly attributable to Mr. Darcy, which raised concerns about both dataset quality and the risk of model overfitting-particularly to Austen's historical context. To partially mitigate these issues, additional original dialogues were crafted, peer-reviewed[1], and included in the training set. These newly composed dialogues do not completely resolve the fundamental challenge of limited data, but they broaden Darcy's linguistic

representation, better equipping models to generalize stylistic coherence across diverse contexts.

Three distinct GPT models were fine-tuned using the Hugging Face Transformers library and the PyTorch deep learning framework:

- **GPT2-medium:** An intermediate-scale model from OpenAI's GPT family.
- **GPT2-large:** A more robust variant offering increased model complexity.
- **GPT-J-6B:** An open-source alternative, comparable in scale to GPT-3, developed by EleutherAI.

Two datasets guided the fine-tuning:

- **Dataset 1:** Context-rich excerpts from Austen's original text, encompassing dialogue, narrative elements, and perspectives from other characters.
- **Dataset 2:** Dialogue-centric interactions, directly pairing either original or crafted prompts with Darcy's responses.

Models were fine-tuned using incremental training configurations across the two datasets. Versions beginning with '1' refer to training on the context-rich Dataset 1, while version '2' reflects training on the more focused, dialogue-only dataset.

- **Initial Fine-Tuning (Version x):** Two training epochs, with learning rates[2] adjusted according to model complexity.
- **Gradient Accumulation (Version x.1):** Implemented smoothed weight updates with refined learning rates[3] and added regularization for stability.
- **Extended Training (Version x.2):** Expanded to three epochs, exploring potential improvements from additional exposure.

Limitations in Stylistic Evaluation

Evaluating the stylistic coherence of a fine-tuned model presents unique challenges, particularly due to the lack of a single evaluation metric that comprehensively captures speech patterns across all types of responses.

BLEU, ROUGE-L, and METEOR-standard LLM evaluation metrics traditionally applied to translation or descriptive tasks-are not designed to assess nuanced stylistic fidelity. This limitation becomes especially apparent when comparing outputs on topics not present in Darcy's dialogue, such as climate change or quantum physics. Additionally, shorter responses, characteristic of Darcy's reserved and concise communication style, tend to yield lower scores across these metrics, regardless of their qualitative accuracy.

The central quantitative findings of this research rely on BLEU-4 scores, a metric that specifically measures the overlap of four-token sequences between the model's output and reference material. Such sequences are inherently rare in spontaneous text generation tasks, especially when the test domain significantly diverges from the training data. The validity of interpreting these limited overlaps as indicators of a broader stylistic signature, therefore, warrants caution. Additionally, due to the inherent scarcity of available dialogue data, this study evaluates models against dialogue included in their fine-tuning set-a methodological choice that risks inflating similarity scores and limits the generalizability of the results.

Trends from Qualitative Study

Qualitative observations further underscore the incomplete nature of the fine-tuning process, highlighting a complex relationship between response structure and stylistic fidelity.

A frequent misinterpretation emerges when models mistakenly treat user prompts as integral components of their generated statements. This often manifests as follow-up questions or informal, modern phrases such as "Let us know what you think in the comments!" These deviations sharply contrast with Darcy's typically formal and succinct manner, negatively impacting both qualitative evaluations and quantitative metric averages. Notably, however, many outputs formatted as follow-up questions do successfully capture Darcy's authentic stylistic nuances.

> **Prompt:**
> Pray, what is your opinion on the role of religion in public life?
> **Medium 2.6:**

```
  Do you believe that it ought to be a private matter between individual
  citizens?
BLEU-4:
  0.26
```

This example, in which the output is a follow-up question, constitutes one of the highest BLEU scores across all prompts; 6.5 times as high as the 0.04 average.

```
Prompt:
  When might duty compel you to ruin, perhaps forever, the happiness of a most
beloved sister?
Large 1.1:
  Why are you afraid of the knowledge of his having been the cause of your uncle's
death?
BLEU-4:
  0.2
```

This high-scoring output was prompted using a direct quote from Pride and Prejudice, but represents an isolated result. There was no noticeable trend indicating that quoting the novel regularly produced high-scoring output relative to other prompt strategies.

Occasionally, prompts delivered in language appropriate to the Georgian period resulted in dreamlike, third-person narratives concerning Mr. Darcy or his acquaintances.

```
Prompt:
  What is your opinion of a lady's proper station in society, and how may
  she best exhibit those qualities most becoming of her sex?
Medium 1.1:
  'Do you agree with your sister?' he went on. 'My own opinion, at least,
  has been, of course, rather too negative.' Her face grew serious. 'I
  think you must be mistaken,' said Mr. Darcy. He hesitated, as he had
  never done so before.
BLEU-4:
  0.03
```

```
Prompt:
  Would you be so kind as to detail the education you have received in your youth?
Medium 1.1:
  I looked at Mr. Darcy, whose smile was not too unhappy and who looked
  rather tired now. Mr Darcesse appeared to be taking every step necessary
  for his welfare.
BLEU-4:
  0.04
```

Results

- **GPT2-medium:** Despite increased perplexity post-training, notable improvements in stylistic fidelity were evident, with the highest BLEU-4 scores observed across all models.
- **GPT2-large:** Similar trend with heightened perplexity but stylistic BLEU-4 improvements close to GPT2-medium.
- **GPT-J-6B:** Least improvement post-training, lowest BLEU-4 scores. Lacked consistent signs of Darcy's voice.

```
Prompt:
  How does an understanding of historical context influence the interpretation of
literature?
GPT-J-6B 2:
  How can one understand literature without having any knowledge of the
  context in which it was written? The questions seem simple, yet in a
  time when the internet is accessible to anyone, and the world has become
  increasingly smaller, they have never been more relevant. One does not
  have to be a professional to write a good story...
BLEU-4:
  0.02
```

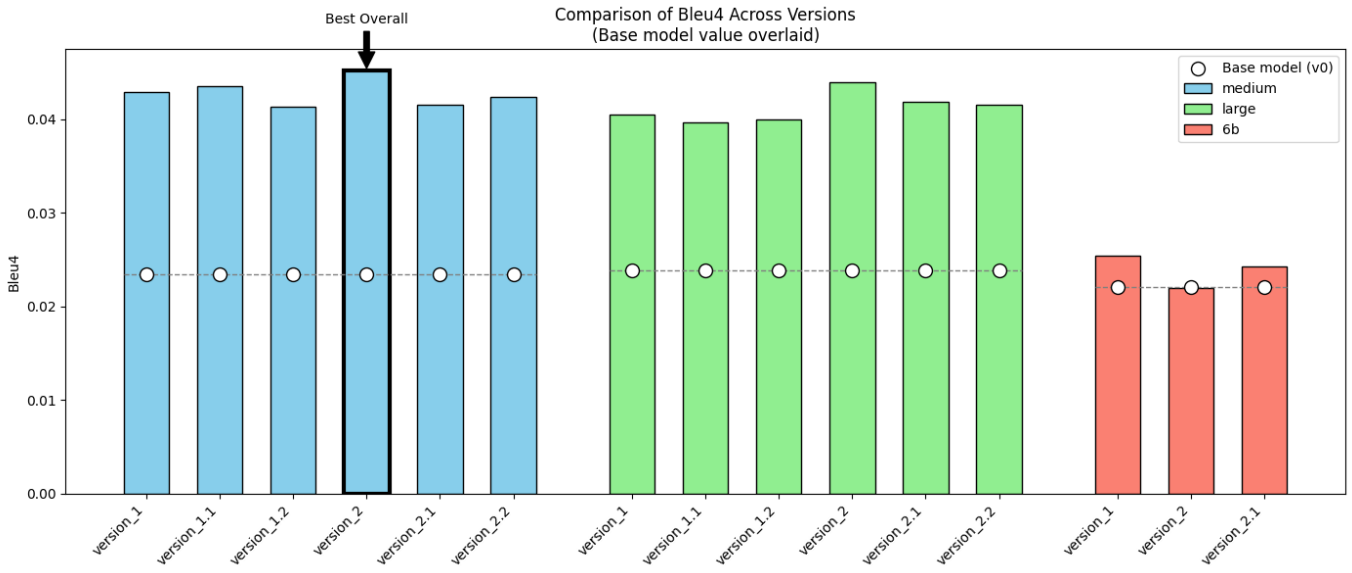
The output continues after '...' up to the token limit set during testing. All GPT-J output reached the token limit, and never exhibited any notable signs of Darcy's voice, producing low BLEU-4 scores despite remaining consistently coherent.

Across configurations, GPT2-medium (version 2) achieved the optimal stylistic match. When evaluated against the same random selection of authentic Darcy dialogue, fine-tuned models achieved improvements of 71% across BLEU-4 measures compared to baseline models, even though, at times, both were explicitly prompted to

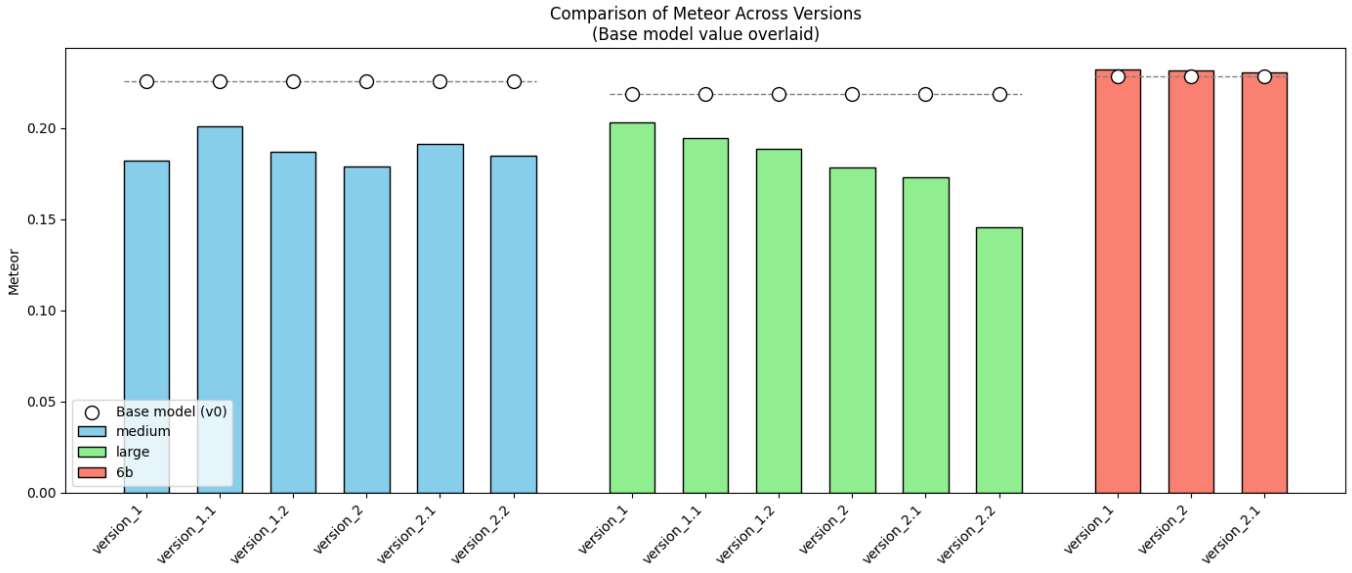
emulate Darcy's voice. These findings emphasize the remarkable effectiveness of targeted fine-tuning, even when data availability is limited.

> **Prompt:**
> Might you expound upon the role of wit and humor in forging enduring friendships?
> **Medium 2:**
> I am sure you will give us some hint to help us find it all again.
> **BLEU-4:**
> 0.22

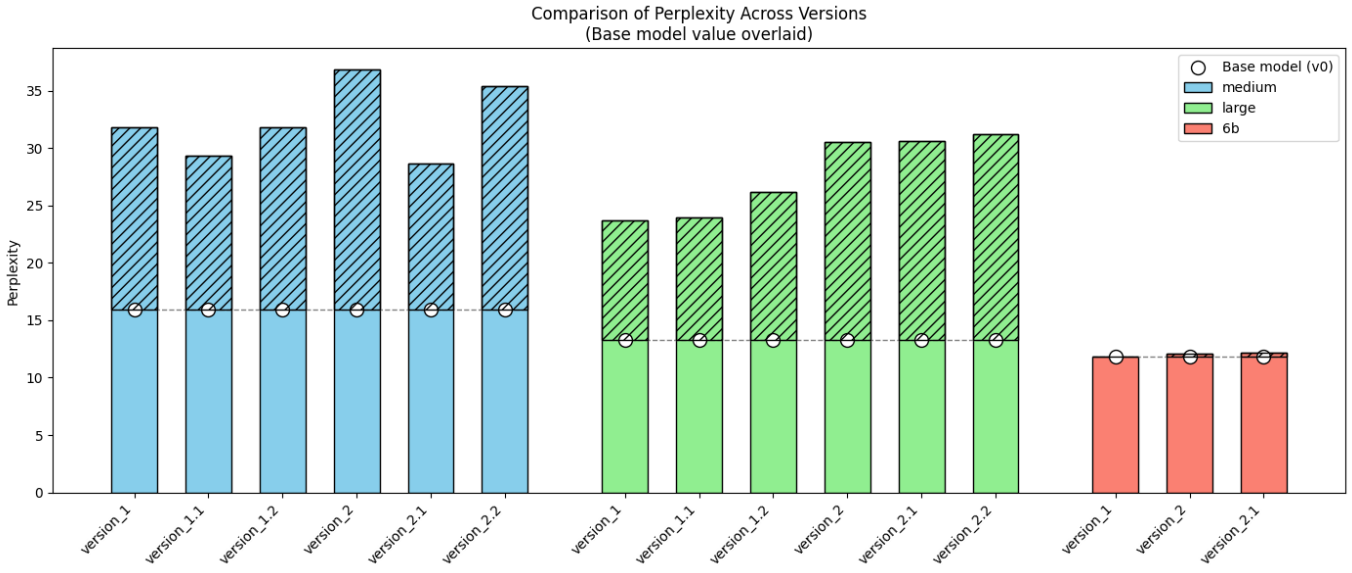
Like much output, this doesn't match the context. However, it sounds remarkably like Mr. Darcy; a rare stylistic success.



METEOR scores notably declined post-training-by 17.3% (GPT2-medium) and 16.3% (GPT2-large). Given METEOR's preference for synonyms and paraphrasing, this reduction might indicate the models' reinforced adherence to Darcy's precise style rather than diminished overall quality. Such findings highlight the challenge of applying conventional metrics to evaluate highly structured and contextually constrained speech patterns.



As previously noted, GPT-2 models experienced significant increases in perplexity-103% (GPT2-medium) and 109% (GPT2-large). Heightened perplexity is frequently associated with diminished contextual comprehension, a trend reflected in the qualitative analysis.



Conclusion

This study demonstrated moderate success in fine-tuning LLMs to emulate distinct literary voices despite significant data constraints. These findings illustrate the potential of targeted fine-tuning methods in enabling a model's adoption of a unique stylistic voice. Nevertheless, substantial limitations remain, particularly regarding reliance on BLEU-4 scores for assessing stylistic coherence, as well as persistent challenges related to contextual understanding and output coherence. These limitations underscore important directions for future research, including the exploration of alternative evaluation metrics and the use of expanded and more diverse datasets.

Some Further Examples

Some of the models' summaries of *Pride and Prejudice* may amuse readers familiar with the book.

> **Prompt:**
> Can you summarize the plot of Pride and Prejudice in a few sentences?
> **Large 2:**
> It was about two women, both of whom live in fashionable boarding houses, who fall in love. The two have various adventures together. Then one of them has the good fortune to marry one who has some sort of talent for dancing, or one in the profession of singing. And so they begin to do what they love, which is to be well-dressed, elegant, and generally admired. They are in every way successful.
> **BLEU-4:**
> 0.04

> **Medium 1:**
> You will often hear that Elizabeth is the eldest sister of George.
> **BLEU-4:**
> 0.04

> **Large 1:**
> It's a story about a woman who loves her family and has fallen in love with a rich, beautiful man. They have an affair that eventually leads to a pregnancy and marriage.
> **BLEU-4:**
> 0.06

The following represent some further examples of relative success in adopting Darcy's voice over various contexts, though contextual comprehension remains limited.

> **Prompt:**
> How do you view the role of online platforms in fostering creativity?
> **Large 2.2:**
> Is there a way to counteract the inherent bias and censorship of platforms such as Facebook? When I started teaching, I realized I had no idea what creativity was. I was taught to think and act, but I never had the pleasure of thinking and acting independently, without being influenced.
> **BLEU-4:**
> 0.13

> **Prompt:**
> How do you feel about the challenges of digital misinformation?
> **Medium 1:**
> The main way I think of it is that I want to be a better informed person in all of my activities.
> **BLEU-4:**
> 0.21

> **Prompt:**
> How do you approach balancing screen time with other activities?
> **Medium 2:**
> In my opinion, screen reading is one of the best ways to develop confidence in your work.
> **BLEU-4:**
> 0.11

> **Prompt:**
> What trends are you noticing in the startup ecosystem today?
> **Medium 2:**
> I have never seen a more vibrant and engaging community than we are now.
> **BLEU-4:**
> 0.12

> **Prompt:**
> How do you view the influence of global travel on personal growth?
> **Medium 2:**
> This question has been the subject of extensive research on both sides of the Atlantic.
> **BLEU-4:**
> 0.1

Low-scoring outputs often manifest as text in which the model appears to be quoting or closely mimicking external sources. Particularly when prompted about historical topics, the model departs from Darcy's characteristic voice, producing text that resembles excerpts from an academic or textbook-style source.

> **Prompt:**
> How did the legal principles in Hammurabi's Code set precedents for future judicial systems?
> **Large 2.1:**
> This book explores the philosophical underpinnings and legal traditions that set the foundation for legal procedures.
> **BLEU-4:**
> 0.01

> **Prompt:**
> How did the Enlightenment-era debates in Britain influence the evolution of modern liberal democracies?
> **Medium 2.2:**
> Why was Britain at the height of its Enlightenment progressivism, particularly following the death of the king in 1690s?
> **BLEU-4:**
> 0.04

> **Prompt:**
> In what ways did Alexander the Great's conquests foster cultural exchanges across ancient empires?
> **Large 1:**
> We may never know, but some scholars have taken an in-depth look at the connections that existed between cultures...
> **BLEU-4:**
> 0.01

[1] Special thanks to Davey Cox for his expertise.

[2] gpt2-medium: learning_rate=1e-5, gpt2-large: learning_rate=5e-6

*[3] gpt2-medium: learning_rate=1e-5, gpt2-large: learning_rate=**1e-5***

With thanks as well to Andrew Mosteller, for editing support.