Generative AI Worked Examples – Report

Student: Kaushik Jayaprakash
Course: INFO 7390 – Art and Science of Data
Assignment: Generative AI Worked Examples (Fall 2025)

Dataset Choices

For this assignment, I selected **two text datasets with contrasting linguistic and stylistic characteristics** to compare how generative language models behave across different domains.

Dataset A – Wikitext-2 (Raw)

- Formal, encyclopedic writing
- Long paragraphs and structured exposition
- Suitable for evaluating how generative models learn factual, coherent, and topic-driven text
- Minimal preprocessing is required, making it ideal for a controlled fine-tuning task

Dataset B – Reddit TIFU (Short)

- Informal, conversational writing
- First-person narratives, slang, and emotional expressions
- High variability in tone and sentence structure
- Useful for testing how models adapt to social-media style text with personal storytelling elements

These two datasets complement each other by representing **formal vs. informal** language domains, allowing a clear comparison of model behavior when trained on different text styles.

Methodology

For both datasets, I fine-tuned **GPT-2 (small)** using causal language modeling:

1. **Tokenization** using GPT-2's Byte-Pair Encoding tokenizer
2. **Concatenation + chunking** into 128-token sequences
3. **Training** for one epoch with a small batch size to keep compute requirements manageable
4. **Evaluation** using validation loss and perplexity
5. **Generation samples** to qualitatively examine style adaptation

Both examples used the **same architecture and training process**, enabling a meaningful comparison driven only by dataset differences.

Key Findings

1. Perplexity Differences

- Wikitext-2 achieved a **lower perplexity** than Reddit TIFU
    - Due to being more consistent in structure and vocabulary
    - Easier for the model to predict next tokens in formal, patterned text
- Reddit TIFU's perplexity was **higher**
    - Caused by slang, emotional expression, abrupt topic shifts, and inconsistent phrasing

2. Style Adaptation

Fine-tuned GPT-2 adapted strongly to each dataset's linguistic style:

Wikitext-2 model produced:

- Expository, neutral, and topic-focused text
- Longer and more structured sentences
- A tone similar to Wikipedia or textbooks

Reddit TIFU model produced:

- Conversational, emotional, and narrative-driven text
- First-person phrasing ("I realized…", "Today I messed up…")
- Informal language, humor, and storytelling elements

3. Influence of Dataset Characteristics

The experiment shows that **dataset style dominates model behavior**, even with minimal fine-tuning:

- Formal datasets push the model toward coherent, informative writing
- Informal datasets lead to expressive, personal storytelling
- The same architecture can produce radically different outputs depending on the training data

Conclusion

By fine-tuning GPT-2 on two contrasting datasets—Wikitext-2 and Reddit TIFU—I demonstrated how generative language models internalize and reproduce the style, tone, and structure of the data they are trained on. Wikitext-2 enabled more predictable and structured generations, while Reddit TIFU resulted in diverse, conversational, and narrative outputs. These findings highlight the **sensitivity of generative models to dataset characteristics** and emphasize the importance of data selection in developing domain-specific AI systems.