Background and Methods

The Hippocorpus dataset is a collection of stories sourced from crowdworkers in three distinct phases:

1. **First Phase**: Workers shared personal accounts of memorable events from the past 3-6 months, accompanied by a brief summary. This resulted in 2,779 stories from 2,662 authors. Each summary defined its corresponding story's topic, such as an expectant mother discovering she was having twins during a camping trip.

2. **Second Phase**: Months later, a subset of the original authors was prompted to retell their stories using only their summary as a cue, yielding 1,319 retold stories.

3. **Third Phase**: Different crowdworkers wrote fictional accounts based on the provided summaries, creating stories as though these experiences were their own. This produced 2,756 imagined stories by 1,434 authors.

In addition to the stories, the dataset also includes information on the time since the event and its recall frequency.

For this project, I extend the Hippocorpus dataset by generating 2,756 imagined stories using GPT-3. This parallels the existing 2,756 human-generated imagined stories crafted from the brief autobiographical summaries. The paired set of human and AI-generated stories provides an avenue for direct comparison in a controlled manner. My primary objective is to pinpoint distinguishing features between fictitious human-generated and GPT-generated stories.

Results

I generated the imagined GPT-stories dataset using the `gpt-3.5-turbo` API with default configurations. To enhance data collection efficiency, I used multiprocessing across the 8 cores of my machine. To counter occasional rate limits by OpenAI, I added a feature that resumes dataset generation from the last saved checkpoint. This enabled me to generate 2,756 narratives in 5 hours, costing less than $5.

Note that the {{`time_since_event`}} and the {{`summary`}} are parameters that are populated in the `jinja2` template from the existing Hippocorpus dataset. Note that I added the final two guidelines to the original human instructions for the GPT system prompt. This adjustment was made after I noticed that GPT responses often began with salutations, dates, or direct references to the story's timeline—patterns not observed in human responses. You can see the user and systems prompts below.

System Prompt: Given a short prompt summary, write an imagined journal entry about an event. Writing instructions: The story must correspond to the summary. Pretend the event happened to you, but do not write about something that actually happened to you. Write using first person perspective. Use the timeline of when the event happened (e.g., "3 weeks ago", "6 months ago"). Now, write the journal entry below. **Reminder** Please make sure to write a story that corresponds to the summary written above. Don't write a story about something that actually happened to you. Story must be 15-25 sentences and 600-3000 characters including spaces. Do not start off with any salutations (e.g., "Dear Diary") or dates (e.g., 3/1/2021"). Do not include the specific date or time of the event in the story.

User Prompt: Summary (from {{ time_since_event }} days ago): {{ summary }}. Generated Story:

I created a game that compares human and AI responses to gauge their similarities and differences based on a summary. Initial findings show that human responses are typically shorter than GPT's. This might be because crowdworkers are paid the same regardless of response length, making them prioritize brevity. Additionally, GPT responses showed less varied sentence structures, fewer unique words, and increased reading difficulty compared to human responses. The code for Phase 1 is here. All generated GPT stories are in the `hcV3-imagined-stories-with-generated.csv` file.

General Discussion and Future Plans

For Phase 2, I plan to extract distinct features that differentiate between GPT-generated and human-crafted narratives. Specifically, I will delve into Linguistic Features, Semantic Similarity, and Prompt Engineering methodologies for this comparative analysis. Should time allow, I will also explore the implications of these techniques when applied to a few-shot system prompt, extending the zero-shot approach.

- **Linguistic Features**: I'll analyze average word and sentence lengths, vocabulary diversity, and more. This analysis will aid in distinguishing between human and GPT narratives, and I will subsequently develop a classifier to predict narrative origins based on these features.

- **Semantic Similarity**: Using the OpenAI API, I will generate embeddings for each story and compare them to embeddings of their corresponding summaries via cosine similarity. This will provide insight into how closely narratives, whether human or GPT, adhere to their initial summaries.

- **Prompt Engineering**: I'll implement a prompt engineering techniques including normal prompting, chain-of-thought, and self-consistency, on a subset of summaries. The outcomes will help quantify the distinguishing characteristics each method introduces in the narratives.

- **(Stretch) Extension to Few-Shot Prompting**: Given the dataset's reliance on zero-shot prompting, I will investigate the impact of few-shot prompting on narrative generation and contrast the results with the techniques detailed above.

Introduction

Through experiments, I sought to address the question: How do human-generated and AI-generated stories differ in various aspects? To explore this, I augmented the Hippocorpus Dataset by creating a set of 2,756 GPT-generated stories, mirroring the existing 2,756 human-generated stories.

First, how well can linguistic features differentiate human and AI-generated stories? I extracted elements like noun frequency and Flesch reading scores from both story types and then computed the relative percentage differences in these features. Then, I trained a Logistic Regression classifier on the linguistic features to determine the story's provenance.

Second, which type of story – human or AI-generated – more accurately reflects the intent of their original summaries? I converted summaries and corresponding stories into embeddings to measure semantic alignment. Subsequently, I trained a Logistic Regression classifier on the embeddings to determine the story's provenance.

Third, do different prompting techniques affect the ability for `gpt-3.5-turbo` to identify a story's origin? I experimented with standard, Chain of Thought (CoT), and Self-Consistency prompting techniques.

Finally, does incorporating human examples in few-shot generation alter experimental outcomes? To explore this, I crafted a new dataset of $2,756$ few-shot stories, using human narratives for in-context learning. Subsequently, I revisited each experiment with this dataset to observe the effects. The code for Phase 2 is here and the experimental data is here.

Methods

In the first experiment, I analyzed linguistic features to distinguish AI-generated narratives from human-written ones. Using `spaCy`, I tokenized texts for frequency metrics and parts of speech distributions. I assessed readability and sentiment with `textstat` and `vaderSentiment`, respectively.

I expressed feature differences as a percentage change from human stories, $\left(\frac{\text{AI Feature Value}}{\text{Human Feature Value}} - 1\right) \times 100$, with statistical significance evaluated for each. Qualitatively, these vectors highlight divergent feature distribution patterns. Quantitatively, I used `sklearn` to do standard pre-processing (scale features to have $\mu = 0, \sigma^2 = 1$) and train a Logistic Regression model (80% train, 20% test) to determine story provenance.

In the second experiment, I used OpenAI's `text-embedding-ada-002` model ($1,536$ dimensions) to create embeddings for the summaries, human-generated stories, and GPT-generated stories. The objective was to assess which story type, human or AI-generated, more closely captured the intent outlined in the summaries, as indicated by higher cosine similarity measures. Subsequently, these embeddings served as input features for a Logistic Regression classifier, following the same training protocol as in the first experiment, to identify the origin of the stories.

In my third experiment, I used OpenAI's `gpt-3.5-turbo` API for story classification. I input each story into the model's context window and tallied the Large Language Model's (LLM) accuracy in making predictions. Initially, I applied standard prompting, where I simply asked the LLM to decide without providing any examples. For the Chain of Thought (CoT) approach, I instructed the model to articulate its reasoning step by step before making a prediction. Lastly, with the Self-Consistency method, I repeated the CoT process three times for each story, basing the final classification on the majority vote from these iterations.

In the final phase, I created a dataset of $2,756$ AI-generated stories, each conditioned by the same three randomly-selected human narratives from the Hippocorpus dataset. Subsequently, I reran all three experiments using this new few-shot dataset, aiming to assess how few-shot learning impacts the accuracy of provenance prediction.

Results

The results of Experiment 1 are captured in Table 1, which shows notable differences in the frequency of language features between human-generated and GPT-generated stories. These differences are more pronounced in the zero-shot dataset, where no human stories were provided as examples. Features like the usage of nouns, adjectives, and ellipses are distinct, indicating the likely authorship of the story. The few-shot dataset, informed by human examples, shows smaller yet significant differences in these linguistic features.

Feature	Zero-Shot	Few-Shot
noun_freq	153.04%	130.38%
char_freq	107.51%	98.52%
adjective_freq	99.16%	92.96%
word_freq	90.77%	89.36%
unique_word_freq	78.24%	72.83%
verb_freq	83.96%	82.42%
sentence_freq	52.96%	63.99%
pronoun_freq	50.22%	59.34%
sentiment	46.14%	49.22%
adverb_freq	18.69%	30.80%
avg_sentence_std_dev	-5.01%	-5.12%
flesch_score	-18.03%	-11.38%
exclamation_freq	-68.32%	-21.35%
ellipsis_freq	-92.89%	-84.10%

Table 1: Comparison of percentage differences in linguistic feature values between human and AI-generated narratives for both zero-shot and few-shot GPT-generated datasets.

Tables 2 and 3 from Experiment 2 provide insights into the comparison of human and GPT-generated stories based on their cosine similarity to original summaries. Table 2 indicates that GPT stories aligned more closely with the original summaries 58.3% of the time in the zero-shot setting and 64.4% in the few-shot scenario. This suggests that GPT was surprisingly more effective at replicating the original summary's content in the few-shot case, even with human examples to guide the process. From Table 3, it is evident that, on average, GPT stories had a marginally higher mean cosine

similarity and displayed less variability in similarity scores compared to human-written stories, implying a more consistent capture of the original summary's intent by GPT.

Dataset	Human Story	GPT Story
Zero-Shot	1149	1607
Few-Shot	982	1774

Table 2: Frequency of stories with higher cosine similarity to the original summary for zero-shot and few-shot datasets.

Statistic	Zero-Shot		Few-Shot	
	Human	GPT	Human	GPT
Mean	0.887	0.890	0.887	0.893
Std	0.037	0.034	0.037	0.034
Minimum	0.662	0.695	0.662	0.701
25%	0.868	0.873	0.868	0.875
50%	0.892	0.895	0.891	0.899
75%	0.912	0.914	0.913	0.917
Maximum	0.967	0.957	0.967	0.968

Table 3: Summary statistics for cosine similarity scores in $2,756$ human and GPT generated stories compared to the original summary for zero-shot and few-shot datasets.

Table 4 summarizes the outcomes of Experiment 3, where the accuracy of various prompt engineering techniques for identifying the origin of stories—whether human or GPT-generated—was assessed. Across both zero-shot and few-shot datasets, each method underperformed against a random baseline in determining provenance. This implies that techniques such as CoT (Chain of Thought) and self-consistency, even when combined, are insufficient for recognizing the distinct distributional patterns of human vs. GPT writing. A sample size of 100 stories was chosen for analysis, providing a balance between statistical validity and token cost.

Zero-Shot Dataset		
Technique	**Accuracy (%)**	**CI (95%)**
Normal	49.0	[0.39, 0.59]
CoT	47.0	[0.37, 0.57]
Self-Consistency + CoT	47.0	[0.37, 0.57]
Few-Shot Dataset		
Technique	**Accuracy (%)**	**CI (95%)**
Normal	49.0	[39.20, 58.80]
CoT	45.0	[35.25, 54.75]
Self-Consistency + CoT	46.0	[36.23, 55.77]

Table 4: Accuracy of prompting in predicting story's origin and 95% confidence intervals for zero-shot and few-shot data. A random sample of size $n = 100$ stories was evaluated.

Table 5 presents test set accuracies for Logistic Regression classifiers distinguishing story origins using linguistic patterns in Experiment 1 and embeddings in Experiment 2. The classifiers achieved higher accuracy with embeddings, indicating they may offer better predictive capability than linguistic features alone. Nonetheless, both linguistic and embedding features effectively determined story provenance in the zero-shot and few-shot datasets.

Dataset	Experiment 1	Experiment 2
Zero-Shot	98.91%	99.91%
Few-Shot	97.19%	99.91%

Table 5: Test set accuracy scores for a logistic regression classifier in zero-shot and few-shot GPT generated datasets.

Discussion

The results indicate that human-generated and AI-generated stories strongly differ. First, linguistic features are effective at distinguishing between human and GPT-generated stories, as demonstrated by the distinct distributions in Table 1 and the high test set accuracy in Table 5. Second, GPT-generated stories more closely align with the intent of original summaries, showing lower variability than human-written counterparts. Additionally, the analysis confirms the robustness of embeddings as predictive features for story provenance. Third, the current prompting methods and in-context learning prove inadequate for discerning the subtle traits that differentiate human from GPT-generated text. Fourth, incorporating human examples in few-shot learning marginally reduces the distinctiveness of GPT-generated features but does not significantly alter the outcomes of the experiments.

The data from Table 2 leads me to suggest that few-shot learning unexpectedly refines GPT's ability to replicate human summary intent. Contrary to initial assumptions, the introduction of human examples does not diminish, but rather enhances GPT's alignment with the original summary. This is particularly surprising given that human-generated stories show less alignment on average in the zero-shot case. The implication is that few-shot learning may fine-tune GPT's interpretative accuracy, enabling it to reflect the original human intent more faithfully than without this additional context.

In the experiments, I used a consistent temperature setting for GPT-generated summaries, which may contrast with natural human narrative variability (Table 3). The results, while valid for GPT-3.5, need verification across other LLMs like LLAMA-2, GPT-4, or Mistral-7B. Suboptimal system prompts could have limited the effectiveness of prompt engineering in provenance detection. Moreover, the use of GPT's embeddings for measuring semantic alignment might bias towards GPT-styled outputs having "better" intent alignment.

Future studies should investigate how adjusting temperature settings affects the diversity of AI-generated stories and test the consistency of these results with various LLMs. Additionally, assessing the influence of using diverse embedding models beyond GPT's and optimizing system prompts could further improve the provenance detection capabilities of prompt engineering approaches.

References

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., ... Sayed, W. E. (2023). *Mistral 7b.*

Sap, M., Jafarpour, A., Choi, Y., Smith, N. A., Pennebaker, J. W., & Horvitz, E. (2022). Quantifying the narrative flow of imagined versus autobiographical stories. *Proceedings of the National Academy of Sciences*, *119*(45), e2211715119. doi: 10.1073/pnas.2211715119

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models.*

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... Zhou, D. (2022). *Self-consistency improves chain of thought reasoning in language models.*

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). *Chain-of-thought prompting elicits reasoning in large language models.*