Analyzing Essay Prompts by Their Ability to Differentiate LLM and Human Responses

Ibrahim Musaddequr Rahman
iamr@umich.edu

1 Introduction

Large language models enable automated generation of natural language, which poses significant challenges for educators who rely on writing assignments as a learning tool. Due to model efficacy and the inherent difficulty of distinguishing LLM-generated essays from human-written ones, educators must now design assignments that remain resilient to LLM-based tools such as ChatGPT. Despite this pressing need, little analysis has determined which assignments are more or less susceptible to LLM completion. This project addresses that gap by aiming to quantify how "GPT-able" specific essay prompts are, thereby identifying which assignments enforce human authorship.

This goal is achieved through a two-step process. First, a classifier is trained to differentiate human and LLM writing as a proxy for human grading. Second, this classifier is run on a large set of assignment samples containing both human and LLM responses. A secondary model is then trained to estimate classifier efficacy based on the assignment itself, which could be applied to new prompts and enable educators to modify assignments to better differentiate LLM and human work.

In experiments, a small sentence transformer model trained to predict Qwen-2.5-7B's ability to imitate human writing on a creative writing dataset outperformed linear regression and random baselines, providing a meaningful ranking of prompts by their susceptibility to LLM completion. When this trained model was applied to analyze the dataset's prompts, it was found that generic assignments such as open-ended dialogue poorly differentiated LLM-generated text from human writing, while more philosophical assignments such as questions about death or memory better elicited distinctive human responses. All code and implementations for this project are available at https://github.com/iamr-gh/essay-gptability.

2 Related Work

Many studies about LLMs and education focus on user studies of how humans create prompts and consume outputs. For instance, Fleckenstein et al. asked teachers to differentiate student and LLM-generated essays for English as a Foreign Language tests (Fleckenstein et al., 2024). Similarly, Scarfe et al. examined LLM usage in a real educational setting by submitting unlabeled AI assignments within a university context. It was found that 94% of AI submissions were undetected by educators and that these submissions were scored higher on average than human submissions, illustrating the consequences when teachers do not consider LLM generation in assignment design (Scarfe et al., 2024).

Tang et al. analyze how well prompt engineering can improve reliability in automated essay scoring by comparing varying phrasing of objectives across different models (Tang et al., 2024). Lo et al. investigate the quality of LLM-generated feedback for essay improvement by comparing it with human feedback across control groups and analyzing resulting student essays (Lo et al., 2025). Jiang and Hyland compare LLM-generated argumentative essays to student writing and discover statistical patterns more apparent in LLM writing, particularly noting a lack of engagement features common in interactive discourse (Jiang and Hyland, 2025).

None of these studies, however, examine prompts that reliably differentiate humans from LLMs. While general performance compared to human answers is a related problem, it remains distinct because the focus in those studies is on outputs instead of the original prompt. Since most existing work concentrates on user studies, this leaves a clear gap in automated quantification of prompt efficacy. This gap is addressed by developing methods that evaluate prompts directly rather than their outputs.

3 Data Sources and Generation

3.1 Human-Written Text

The ASAP 2.0 dataset from the Learning Agency Lab was initially considered, which contains over 24,000 student essays from writing assessments (Burleigh, 2025). Each datum includes source texts, an assignment, student response, and grader score. For the final version, however, the Writing Prompts creative writing dataset from r/CreativeWriting was selected, which offers greater prompt diversity with a human response for each entry (Euclaise, 2024). To match ASAP 2.0's size and generation constraints, only the first 24,000 entries were used.

Table 1: Dataset Overview

Dataset	Essays	Prompts
ASAP 2.0	24,000	7
Writing Prompts	272,600	97,349
Writing Prompts (Trunc.)	24,000	18,940

3.2 LLM-Generated Text

For each assignment-human essay pair, an equivalent LLM-generated essay was created with the LLM having access to the same sources and assignment description. Initial prompts were generated using the `byLLM` Python library and manually extracted via the verbose debugging feature (Dantanarayana et al., 2025). For the full 24,000 essays, the vLLM inference engine was used on a Tesla V100 in the Great Lakes computing cluster (Kwon et al., 2023). Due to VRAM limitations, Qwen-2.5-7B was used with batch size 8.

4 Methodology

The approach involves two training stages to produce the final model. The primary goal is to predict how differentiable an LLM response will be from a human response given a prompt, and this measurement is accomplished through a trained classifier.

4.1 Classifying Human and LLM Text

For classification, the dataset is converted into labeled essays where human-written content is labeled as 0 and LLM-generated content as 1, ignoring all other assignment information. Cross-entropy loss is used for training and baselines are established that include random selection and logistic regression on a bag-of-words model. Due to overfitting concerns, a 50/50 train-test split is employed with 1 epoch at learning rate 1e-5. The logistic regression performed so well that it produced an overly discrete data distribution for subsequent sections, so it was deliberately undertrained to create a more continuous error distribution. Since the baselines performed strongly, more complex classifiers were not explored.

After training, the model is rerun on all original data. For each assignment-human-LLM essay tuple, the combined classification error e is calculated as:

$$\begin{aligned} e_{\text{human}} &= p_{\text{human}} \\ e_{\text{LLM}} &= 1 - p_{\text{LLM}} \\ e &= e_{\text{human}} + e_{\text{LLM}} \end{aligned}$$

where p_{human} and p_{LLM} are the classifier's output probabilities for correct predictions.

4.2 Predicting LLM Performance Per Prompt

After classifier annotation, error prediction is treated as a regression task, considering only the prompt and error value. Baselines include random choice and bag-of-words logistic regression, where random choice always predicts the mean training error and logistic regression is trained for 10 epochs at 1e-5 learning rate with an 80/20 split. Due to baseline underperformance and task difficulty, a transformer-based regression model was employed.

4.3 Transformer-Based Regression Model

A transformer architecture was designed that combines a frozen sentence encoder with a trainable regression head, which balances efficiency with representational capacity for resource-constrained deployment.

4.3.1 Architecture

The model comprises an `all-MiniLM-L12-v2` sentence transformer encoder and a three-layer regression head. Input texts are tokenized and processed by the 12-layer transformer to produce contextualized embeddings. Attention-masked mean pooling aggregates token representations into fixed-size 384-dimensional sentence embeddings, and the regression head transforms these embeddings through sequential linear layers (384→128→64→1) with ReLU activations and dropout ($p = 0.2$). This architecture contains approximately 50,000 trainable parameters.

4.3.2 Encoder Selection Rationale

`all-MiniLM-L12-v2` was selected based on three criteria. First, its compact size (33M parameters, 120MB) enables CPU inference. Second, pre-training on semantic similarity tasks via contrastive learning produces embeddings relevant for error prediction. Third, the fixed-size output eliminates task-specific pooling, which reduces memory and training time compared to standard BERT models. The 12-layer variant was chosen over the 6-layer alternative due to its superior semantic representation.

4.3.3 Parameter-Efficient Fine-Tuning

Low-Rank Adaptation (LoRA) was applied to reduce trainable parameters while maintaining adaptation capacity. LoRA injects trainable low-rank matrices into dense layers through the computation $\mathbf{W} = \mathbf{W}_0 + \mathbf{BA}$ where $\mathbf{W}_0$ remains frozen. Using rank $r = 16$ and scaling factor $\alpha = 32$, trainable encoder parameters were reduced from 33 million to 600,000 (98% reduction). This configuration balances expressiveness and efficiency, as lower ranks risk underfitting while higher ranks increase memory consumption with diminishing returns. LoRA also provides implicit regularization that mitigates overfitting on limited data.

4.3.4 Training Configuration

The transformer model was trained using mean squared error loss with AdamW optimizer (learning rate 2×10^{-4}, weight decay 0.01). A learning rate scheduler was employed with plateau detection (patience=2, factor=0.5), input sequences were truncated to 256 tokens, and batch size 8 was used on Apple M3 hardware with early stopping (patience=3 epochs) based on validation RMSE. The 80/20 train/validation split used stratified sampling to preserve label distribution.

5 Evaluation and Results

5.1 Classification

Two baselines were evaluated: random choice and bag-of-words logistic regression. Random choice has 0.5 expected accuracy given balanced classes, and due to strong baseline performance, a more complex transformer model was not used. Logistic regression achieved high performance even with a 50/50 split over a single epoch. As specified in Section 3.1, the final classifier was deliberately undertrained to generate a more continuous error distribution.

Table 2: Classification Results

Model	F_1	Accuracy
Random	0.500	0.500
Logistic Regression	0.962	0.963
Undertrained Logistic Regression	0.915	0.914

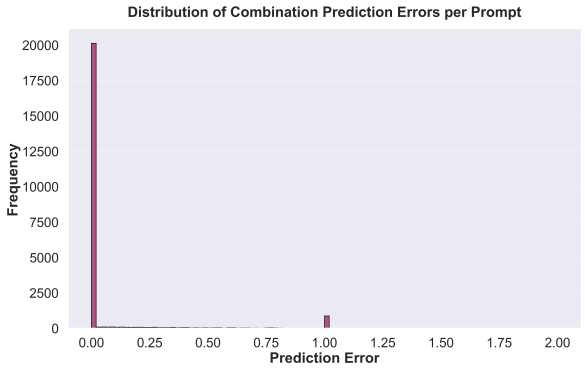

Figure 1: Error Distribution of Fully Trained Classifier

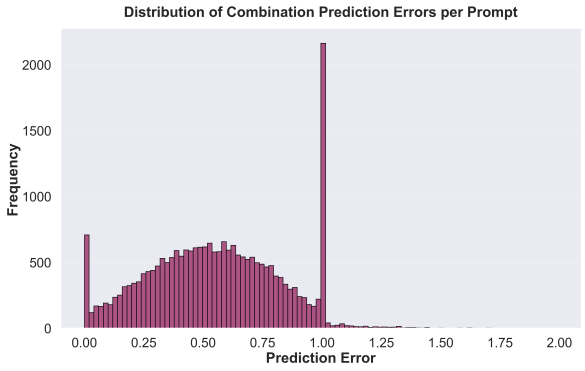

Figure 2: Error Distribution of Undertrained Classifier

5.2 Regression

Regression baselines include random choice and bag-of-words logistic regression. Random choice always predicts the mean training error, while logistic regression is trained for 10 epochs. Performance is reported as mean squared error for each model in Figure 3.

5.3 Interpretation of Regression Model

Using the best-performing transformer model, prompts from the initial dataset were analyzed. Tables 4 and 5 show the top 5 and bottom 5 prompts, where ranking indicates predicted LLM ability to replicate human writing.

Table 3: Regression Results

Model	MSE
Random	0.0825
Logistic Regression	0.1830
Fine-Tuned Sentence Transformer	0.0743

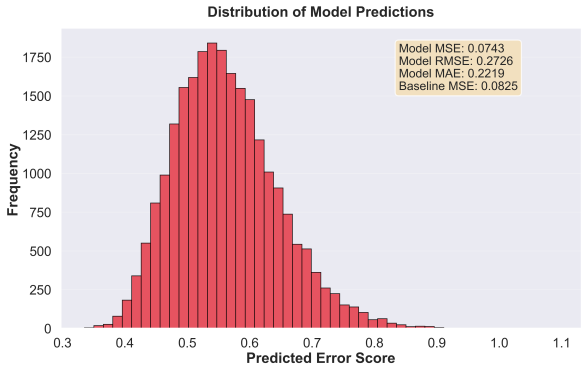

Figure 3: Prediction Distribution of Transformer Model

6 Discussion

6.1 Classification & Regression Performance

The efficacy of the simple classifier stems from the quality of the data used. Because a smaller language model was used for generation, the essays are of lower quality and more distinguishable from human work than would be expected with higher quality models. The undertrained classifier generates a richer dataset for the regression model, as shown in Figures 1 and 2. The complexity of the prompt dataset causes the simplistic bag-of-words model to underperform relative to random choice, as shown in Table 3. The sentence transformer outperforms baselines and offers a meaningful differentiation of prompts, as shown in Figure 3.

6.2 Analysis of Prompts

High-scoring prompts indicate where LLMs imitate humans effectively, while low-scoring prompts indicate where human output is more distinctive. LLMs perform better on generic dialogue tasks that are highly open-ended, as seen by the lack of names and specificity in Table 4. Conversely, humans show greater uniqueness on philosophical tasks, such as the questions about the afterlife in Table 5. The lower scoring prompts are frequently longer than the higher scoring prompts, further emphasizing that specificity provides more opportunity for humans to distinguish themselves.

Table 4: Top 5 Prompts (Most GPT-able)

Rank	Prompt Preview
1	[CW] Write a scene between 2 or more characters using only dialogue . . .
2	[WP] He sighed. . . .
3	[FF] “It was the last time I spoke to him.” . . .
4	[WP] She was almost beautiful. . . .
5	[CW] Write the end of a relationship in dialogue only. . . .

Table 5: Bottom 5 Prompts (Least GPT-able)

Rank	Prompt Preview
1	[WP] On the planet you live on, memories can only be fully remembered once. But because of this, when you replay the memory, its as if you are right back in it. . . The touches, the smells, the. . .
2	[WP] You die. As you regain consciousness, you realize that you are in heaven. However, your expectations of a utopia are quickly ruined when you see that all posts have been abandoned, and a w. . .
3	[WP] One day during group meditation you reach that elusive and ephemeral state of enlightenment - but instead of the promised bliss, your new understanding of things shakes the foundations of your. . .
4	[WP] The afterlife is not based on how you live, but how you die. The more horrific and painful the death, the better the heaven. Conversely, the happier you are when you die, the deeper into. . .
5	You are a Marine in Vietnam, the year is 1967. While out on a patrol in the Central Highlands, your platoon stumbles upon an old abandoned Catholic church that seems to have been built during Frenc. . .

7 Conclusion

A novel two-stage framework was introduced for quantifying prompt susceptibility to LLM completion. The approach trains a classifier to differentiate human and LLM writing, then uses a secondary regression model to predict classification error directly from prompt text. Experiments demonstrated that a sentence transformer approach outperforms simpler baselines, yielding meaningful rankings that reveal generic dialogue prompts as highly vulnerable while philosophical, specific prompts better elicit distinctive human responses.

The effectiveness of this approach is heavily dependent on the quality and diversity of input data. Classifier performance directly influences the regression model's utility, and the undertraining strategy was specifically designed to produce the continuous error distribution necessary for secondary prediction. The dataset required unique construction, balancing prompt diversity against generation

constraints to curate 24,000 creative writing examples after the ASAP 2.0 dataset's limited prompt variety was deemed insufficient.

8 Challenges and Failed Methodologies

8.1 Data Generation

Various generation attempts were unsuccessful due to issues with time requirements, output quality, or cost.

Table 6: Generation Timing

Method	Time
Local torch (Qwen-2.5-7B), serial	200 hrs
OpenRouter (Grok-4.1-fast), serial	101.3 hrs
OpenRouter (K2-0711), serial	62.0 hrs
Local Ollama (Qwen-2.5-7B), batch (4)	41.0 hrs
OpenRouter (GPT-OSS-20B), serial	37.0 hrs
OpenRouter (GPT-OSS-20B), parallel (10)	API limits
Great Lakes vLLM (Qwen-2.5-7B), batch (8)	3.5 hrs

The initial dataset contained 24,000 human-prompt essay pairs (Burleigh, 2025), and generating equivalent LLM essays proved nontrivial. Prompt length and included source texts created complexity that complicated generation. Attempts to use commercial infrastructure economically led to free models on OpenRouter, but these are rate-limited at 1,000 requests per day for paid accounts, with free accounts having even lower limits. vLLM on Great Lakes was eventually used, though this required a compromise on model quality.

8.2 Dataset Diversity

Initially, the ASAP 2.0 dataset was used (Burleigh, 2025), which contains 24,000 human essays responding to complex prompts that each include assignments and evidence. Unfortunately, only 7 unique prompts exist despite the volume, and this lack of diversity leads to similar LLM-generated essays while providing insufficient information for the secondary predictor to learn trends.

To resolve this, a Reddit creative writing dataset was used, which offered greater prompt diversity with a unique human response per prompt (Euclaise, 2024). However, classification remained easy when using smaller LLMs via vLLM because of low output quality. With a 50/50 split, simple bag-of-words logistic regression proved highly effective when fully trained, failing to produce the continuity needed for the secondary predictor. This was overcome by deliberately undertraining the baseline classifier.

9 Future Work

Due to training data limitations, future work should apply this approach to more diverse, higher-quality samples. This research optimized for speed and cost, using small LLMs to generate large volumes of lower-quality output. Future work could use larger LLMs to generate smaller volumes of higher-quality output, which may better represent current academic settings. Such an approach would provide greater diversity in data through a more difficult classification problem.

Ideally, a broader dataset of human work across contexts would improve generalizability. Such datasets exist, and the Purdue Corpus & Repository of Writing is one example (Staples and Dilger, 2018). These are frequently protected from public scraping due to LLM training concerns, but a similar open dataset could be created by aggregating other resources. The creative writing dataset used in this paper was created this way.

References

Logan Burleigh. 2025. Asap 2.0 dataset. Kaggle dataset. Accessed 2025-09-25.

Jayanaka L. Dantanarayana, Yiping Kang, Kugesan Sivasothynathan, Christopher Clarke, Baichuan Li, Savini Kashmira, Krisztian Flautner, Lingjia Tang, and Jason Mars. 2025. Mtp: A meaning-typed language abstraction for ai-integrated programming. *Proc. ACM Program. Lang.*, 9(OOPSLA2).

Euclaise. 2024. Writing prompts.

Johanna Fleckenstein, Jennifer Meyer, Thorben Jansen, Stefan D. Keller, Olaf Köller, and Jens Möller. 2024. Do teachers spot ai? evaluating the detectability of ai-generated texts among student essays. *Computers and Education: Artificial Intelligence*, 6:100209.

Feng (Kevin) Jiang and Ken Hyland. 2025. Does chatgpt write like a student? engagement markers in argumentative essays. *Written Communication*, 42(3):463–492.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Noble Lo, Alan Wong, and Sumie Chan. 2025. The impact of generative ai on essay revisions and student engagement. *Computers and Education Open*, 9:100249.

Peter Scarfe, Kelly Watcham, Alasdair Clarke, and Etienne Roesch. 2024. A real-world test of artificial intelligence infiltration of a university examinations system: A "turing test" case study. *PLOS ONE*, 19(6):1–21.

Shelley Staples and Bradley Dilger. 2018. Corpus and repository of writing (crow): Learner corpus articulated with repository. Website. Available at https://crow.corporaproject.org.

Xiaoyi Tang, Hongwei Chen, Daoyu Lin, and Kexin Li. 2024. Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, 10(14):e34262.