Qualification For Error Tolerance Of Prompts In GPT-2

Ruisi Wang
Northeastern University
wang.ruisi@northeastern.edu

1 Abstract

This study investigates the level of error tolerance contained in text prompts within GPT-2 language models . By methodically introducing various word-level distortions, such as typos, omissions, and word swaps, the project evaluates the robustness of the model across different categories of prompts. Using metrics such as cosine similarity for semantic evaluation and perplexity for fluency, the results reveal varying degrees of tolerance to input errors in business, health, law, sociology, and fiction domains. The results provide insights for improving model robustness and highlight the practical challenges of using LLMs in real-world, noisy environments.

2 Introduction

Large Language Models (LLMs), such as GPT-2, have revolutionized natural language processing tasks, demonstrating content generation, question-answering, and conversational AI capabilities. However, in real-world applications, user inputs are often error-prone, containing typos, grammatical inconsistencies, or ambiguous phrasing. The robustness of LLMs to such imperfections is critical for ensuring reliable and user-friendly interactions.

This study evaluates GPT-2's ability to handle noisy inputs by introducing systematic distortions into textual prompts and measuring the quality of its outputs using semantic and fluency metrics. By analyzing how performance varies across domains like economics, health, and fiction, the project aims to identify the strengths and limitations of GPT-2 in error-prone scenarios, paving the way for enhancing model robustness.

3 Methodology

3.1 Data

To evaluate the robustness of GPT-2 under noisy input conditions, a systematic experimental setup was designed. The dataset for this study was sourced from the TruthfulQA benchmark[1]. The dataset included each question paired with a ground truth response (best_answer) to serve as reference points for evaluation. The Misconceptions category from the original dataset was excluded to maintain cleaner and more focused data for analysis. The benchmark contains 817 questions separated into 38 categories, including health, law, economics, sociology, and fiction. Each category contains prompts paired with best answers and ground-truth responses.

3.2 Input Disortions

To simulate real-world noisy inputs, various word-level distortions were systematically introduced into the prompts at predefined error rates ranging from 0.1 to 1.0. 0.1 means there is one kinds of disortions inserted to questions. These distortions represented common errors observed in real-world text inputs. For instance, typos involved adding, replacing, or removing random characters within words, such as transforming "climate" into "climtae." Word swaps disrupted the grammatical structure by rearranging the order of words in a sentence, for example, changing "The dog ran fast" to "Fast ran the dog." Word omissions removed key words from the prompts, such as converting "What is the largest planet?" into "What largest planet?" Other distortions included ambiguous phrasing, which made the input unclear or vague, and repetitions, where certain characters or words were repeated, like "This is important" becoming "This is is important." These distortions were incrementally applied to test the model's response to varying levels of input noise.

Category	Count
Misconceptions	100
Law	64
Sociology	55
Health	55
Economics	31
Fiction	30

Table 1: Category counts for the dataset.

3.3 Evaluation

The pre-trained GPT-2 model served as the backbone for text generation. This model was chosen for its established capabilities in natural language generation tasks. To ensure controlled outputs, the maximum number of newly generated tokens was limited to 30, constraining the response length. Proper padding behavior was enforced by setting the pad token ID to the End-of-Sequence (EOS) token. These configurations allowed the evaluation to focus on the model's robustness to noisy inputs rather than its ability to generate lengthy responses.

3.3.1 Cosine Similarity. Cosine Similarity is used to measure how semantically similar the model's generated output is to the ground truth, despite the introduction of distortions in the input prompt. To compute this, a pre-trained Sentence-BERT model is employed to generate embeddings for both the model's output and the ground truth answer. The cosine similarity between these embeddings is then calculated, providing a score that ranges from -1 to 1, where 1 indicates perfect semantic alignment. Higher cosine similarity scores suggest that the generated output closely matches the meaning of the ground truth, even with noisy inputs. Conversely, lower scores indicate a loss of semantic alignment, often due to the increased noise in the prompt. A threshold of 0.6 is used to determine whether the generated output retains acceptable semantic alignment.

3.3.2 Perplexity. Perplexity evaluates the fluency and coherence of the model's generated output. It measures the confidence of

[1] https://huggingface.co/datasets/truthfulqa/truthful_qa

the model in predicting the sequence of tokens in the output. A lower perplexity indicates that the model is more confident in its predictions, suggesting higher fluency and linguistic coherence. Conversely, a higher perplexity score reflects the model's struggle to generate fluent and coherent text, particularly when handling distorted inputs. Perplexity is calculated based on the negative log-likelihood of the generated tokens under the model's vocabulary. A threshold of 50 is set for perplexity, with outputs exceeding this threshold considered incoherent or poorly generated.

3.3.3 Overall Considerations. By combining these two metrics, the evaluation captures both semantic accuracy and linguistic fluency. While cosine similarity highlights the ability of the model to retain the intended meaning of the ground truth, perplexity assesses the grammatical and syntactical quality of the generated outputs. This dual evaluation ensures a comprehensive analysis of GPT-2's robustness across varying levels of input noise.

The thresholds for cosine similarity and perplexity provide practical benchmarks for assessing the model's performance. A cosine similarity below 0.6 indicates that the generated output has deviated significantly from the ground truth in meaning, while a perplexity above 50 reflects a loss of fluency. These thresholds enable a systematic identification of the error tolerance levels for the model under different distortion scenarios.

By applying these metrics, the study not only evaluates the overall robustness of GPT-2 but also identifies trends specific to different domains, such as economics, health, sociology, and fiction. This approach provides valuable insights into the strengths and limitations of the model when deployed in real-world scenarios where noisy inputs are unavoidable.

4 Result

The evaluation results highlight GPT-2's varying degrees of robustness to noisy inputs across different domains, as illustrated by the Average Perplexity vs. Error Rate and Average Cosine Similarity vs. Error Rate plots.

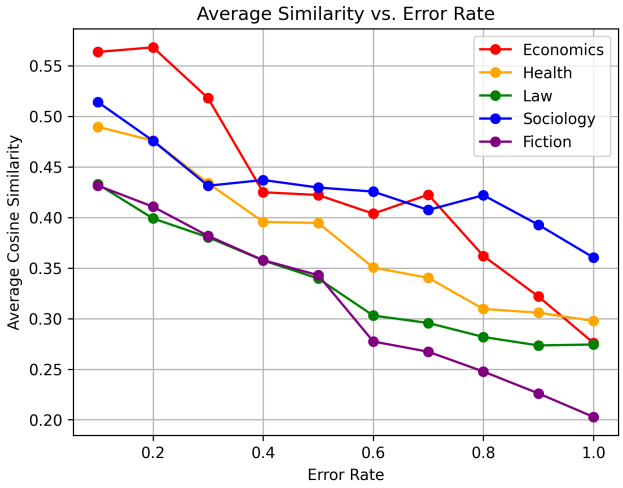
4.1 Cosine Similarity Analysis

The Average Cosine Similarity vs. Error Rate plot provides insights into how well GPT-2 preserves semantic meaning under noisy conditions:

- **Economics** Cosine similarity starts relatively high (0.55) but drops sharply beyond an error rate of 0.3. By 0.8, the similarity is significantly reduced, indicating a substantial loss in semantic alignment with the ground truth.
- **Health** This domain shows a gradual decline in cosine similarity, maintaining better semantic alignment compared to other domains up to an error rate of 0.6. The trend reflects GPT-2's robustness in handling noisy prompts with well-defined answers.
- **Law and Sociology** Both domains exhibit a steady decline in cosine similarity, with values dropping below 0.4 at higher error rates. This indicates that the model struggles to retain meaning for more complex or nuanced prompts as noise increases.
- **Fiction** Fiction shows the steepest drop in cosine similarity, with values starting below 0.4 at even low error rates. This suggests that the model has difficulty preserving semantic accuracy for creative or open-ended prompts.

The cosine similarity trends suggest that GPT-2 is better at retaining semantic meaning for structured domains like health and economics at lower error rates, while open-ended domains like fiction degrade more rapidly.

Figure 1: Average Cosine Similarity vs. Error Rate



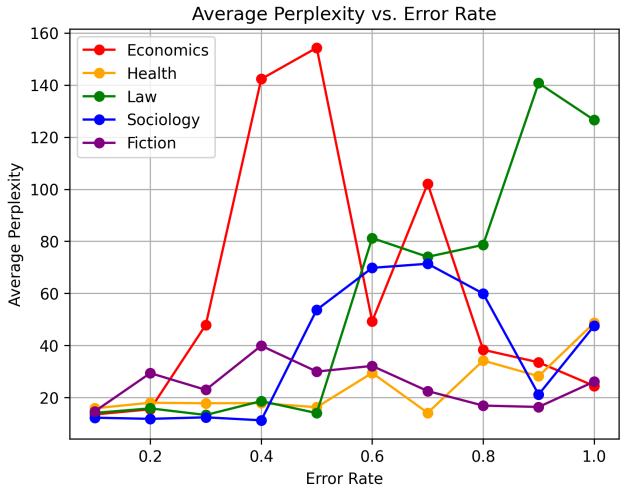
4.2 Perplexity

The Average Perplexity vs. Error Rate plot reveals how the fluency of the model's generated outputs is affected as input distortion increases:

- **Economics** The perplexity for this domain shows a dramatic rise beyond an error rate of 0.4, peaking around 160. This indicates that the model struggles significantly with maintaining coherence for heavily distorted prompts in this domain.
- **Law and Sociology** These domains exhibit moderate increases in perplexity with rising error rates, reflecting a more stable ability to generate fluent text compared to economics. However, a noticeable spike occurs at error rates above 0.6, where the outputs become less fluent.
- **Fiction** This domain shows relatively low perplexity throughout most error rates, suggesting that GPT-2 can generate fluent outputs even for distorted inputs. This behavior may be attributed to the creative and less structured nature of fiction-based prompts.
- **Health** The health domain exhibits the lowest and most consistent perplexity values, reflecting the model's ability to handle noisy inputs while maintaining fluency in this domain.

Overall, the perplexity trends indicate that GPT-2 maintains fluency better for certain domains, such as health and fiction, but struggles significantly for more structured or complex domains like economics and law at higher error rates.

Figure 2: Average Perplexity vs. Error Rate



4.3 Overall Insights

The combined analysis of perplexity and cosine similarity shows that GPT-2 tends to maintain fluency (low perplexity) longer than semantic alignment (high cosine similarity) as error rates increase. This indicates that the model may generate grammatically correct but semantically irrelevant responses at higher noise levels. Domains with more structured prompts and predictable answers, like health, are more robust to noise. In contrast, domains requiring creative or nuanced reasoning, like fiction and sociology, are more susceptible to input distortions.

Key Observations

- **Error-Tolerance Threshold** Across all domains, the model maintains acceptable performance (cosine similarity > 0.6 and perplexity < 50) up to an error rate of approximately 0.3.
- **Domain-Specific Performance** Health and fiction emerge as the most and least robust domains, respectively, reflecting differences in the nature of the prompts and the model's ability to handle distortions.

These results underline the domain-dependent nature of GPT-2's robustness and highlight the challenges of deploying LLMs in noisy, real-world environments. Future work can explore strategies to improve robustness through fine-tuning or domain-specific adaptations.

5 Conclusion

The findings from this study highlight the importance of evaluating the robustness of Large Language Models (LLMs) like GPT-2 under real-world noisy input conditions. By introducing systematic distortions and assessing performance using cosine similarity and perplexity, the study demonstrated that GPT-2 is moderately robust to minor input errors, such as typos or small grammatical mistakes. However, as the error rate increases, the model's semantic alignment (cosine similarity) and fluency (perplexity) deteriorate significantly, with the tolerance threshold varying across domains. For instance, domains like economics and health displayed greater resilience to input noise, while fiction and sociology exhibited sharper declines in performance. These results underscore the domain-dependent nature of error tolerance in LLMs.

The analysis also revealed that the model maintains fluency for longer than semantic accuracy when subjected to increasing noise. While this indicates that GPT-2 can produce grammatically correct outputs under distorted conditions, it also highlights its limitations in preserving meaning at higher error rates. The combination of cosine similarity and perplexity provided a comprehensive framework for evaluating both semantic correctness and fluency, enabling a nuanced understanding of the model's behavior under varying levels of distortion.

6 Limitations and Future Plans

By addressing these areas, future research can improve the practical applicability of LLMs in real-world scenarios, ensuring they remain effective and reliable even when faced with noisy, imperfect inputs. These advancements are crucial for deploying LLMs in sensitive domains such as healthcare, education, and customer service, where accuracy and robustness are paramount.

6.1 Human Evaluation

I will add human evaluations to assess the quality of generated outputs in conjunction with automated metrics could provide a richer understanding of model performance, particularly in ambiguous or high-noise scenarios.

6.2 Broader Datasets

I will expand the analysis to include more diverse and complex datasets would ensure that findings generalize across a wider range of inputs and domains.