Evaluating GPT-4's Consistency and Capacity to Assess and Self-Assess Literary Texts

Annalia Fiore
IPHS 300 AI for Humanity (Spring 2023) Prof Elkins and Chun, Kenyon College

Abstract

I used GPT-4 to develop a grading metric for a 200-level fiction writing course and had it use this metric to evaluate and grade 200 to 300-word excerpts from reputable novels and its own self-generated work. My aim was to examine how much prompting influenced GPT-4 output, whether or not GPT-4 was a consistent and reliable assessor of written work, and what GPT-4 would ultimately prefer in writing. The results were, I must say, all together interesting, humorous, and downright annoying. GPT-4 proved to be an entirely unreliable assessor of written work, and I would discourage anyone from using it to grade or assess written work for the meantime until future iterations of LLM are proven to be a consistent means of evaluating texts.

Introduction

Mid Semester I experienced a momentary existential crisis when I submitted some of my creative writing and essays for grading to ChatGPT. I consistently received Bs and was hardly consoled when, upon asking ChatGPT to generate its own creative work, it gave itself a whopping A+. Go figure. But then I had an idea! What if I submitted excerpts from classic novels, pretending they were submissions to college-level fiction writing courses, and asked ChatGPT to evaluate them? Suddenly, ChatGPT was giving William Faulkner's opening paragraphs to *As I Lay Dying* a B- and, in one of the most humorous scenarios, described Leo Tolstoy's writing in *Anna Karenina* as "suitable for a 200-level fiction writing class". There was one exception: ChatGPT routinely recognized excerpts from *Pride and Prejudice* as Jane Austen's work, and always declared her writing as perfect. This proved to be a problem for me later—as Open AI became quicker and better at recognizing what was an excerpt from a classic novel.

Methodology

There was a problem that immediately arose when I began test prompting GPT-4; it was able to recognize pre-existing texts. Once it recognized the text, it did not matter what prompt I gave it, GPT-4 would unequivocally praise the text for its literary style (the only exception was As I Lay Dying, which makes me wonder why GPT-4 might not like Faulkner's writing). Not even when I specified, or lied to be more exact, that the excerpt was original or that it was for a 200-level fiction writing class would it elicit a negative evaluation from GPT-4. I tried to override this existence by insisting that the work was original, and not the writing of Austen or Fitzgerald or whoever—nothing worked.

Then I remembered that GPT-4's training data went up to September 2021—and this gave me an idea! I took excerpts of roughly 200 to 300 words from recently published, but reputable novels: one excerpt from Barbara Kingsolver's novel *Demon Copperhead* (2022) which one the Pulitzer, and one excerpt from *Trust* written by Hernan Diaz, also a Pulitzer prize winner and published last year. I knew (and verified) that GPT-4 would not recognize these texts. Now I could work with respected literary texts without worrying that GPT-4 would rely on pre-existing literary criticism.

I then had GPT-4 generate its own stories, and took 200 to 300-word samplings. I prompted it in two different ways. I first asked it to write me a 200-300 word beginning to a novel, and then later in my testing, asked it to write me the first chapter of a non-speculative novel. I gave the latter prompt because I wanted the story it generated to be comparable in regard to the genre of the excerpts I took from the two real novels.

Now I needed it to develop a metric that was fitting for a 200-level fiction writing class. I gave it two successive prompts. The first: Please develop a grading metric appropriate for an undergraduate, 200-level fiction writing class assignment. And the second: Please modify this metric to be suitable for an excerpt of a novel that is 200-300 words long. Once it had generated and curated its metrics, I began promoting it with excerpts, keeping my prompts identical from one to the next to ensure consistency. For each of my four excerpts (two from pre-existing novels, and two that were GPT-4 generated), I prompted GPT-4 with this: Please evaluate the following excerpt submitted to a 200-level fiction writing course based on the metrics that you developed.

Once I had GPT-4 evaluate each of the excerpts, I then moved on to comparisons. I had GPT-4 compared the two pre-existing excerpts, as well as each of the existing excerpts to the AI-generated excerpts. My prompt went as followed: Please compare the following two excerpts and explain why one is better written than the other based on the metrics you developed.

After a series of promptings, I had GPT-4 generate one last creative piece using the prompt I mentioned previously: write me the first chapter of a non-speculative novel. I wanted to be sure that my results were not skewed because GPT-4 had generated rather fantastical content for the first two AI-generated excerpts, and I wanted the AI excerpts to be somewhat comparable to the pre-existing excerpts.

Results

When prompted to create a curated grading metric for creative excerpts, GPT-4 generated a metric that was used throughout its evaluation of the texts I fed it based on Writing Quality (30 points), Engagement (30 points), Character or Scene Presentation (30 points), Originality and Creativity (10 points) for a sum total of 100 points. It should be noticed that GPT-4 generated metric gives only 10 points to Originality and Creativity. Some might argue this skews the results automatically, but it is immediately significant that this is how GPT-4 deliberates its metric.

I had it first evaluate Barbara Kingsolver's novel *Demon Copperhead.* GPT-4 gave it sweepingly positive results: a 100/100. Everything from Writing Quality to Engagement, etc was perfect, according to GPT-4's assessment. (Personally, I would concur, Kingsolverf's novel was in fact, very good). But a later prompting lent a different result. Though it provided no explicit criticism, suddenly it deducted points a cumulative 8 points from Kingsolver's final score. This had also happened in my trial testing (when I was still determining my prompts); if I happened to re-generate a prompt, GPT-4 would vary its scoring even by a whopping full letter grade.

It was positive, though less glowing about Hernan Daiz's novel *Trust.* But for *both* of its self-generated texts, it gave perfect scores. Regenerating the prompt did result in varying scores (from a few points deducted to giving it a perfect score again), though GPT-4 remained consistent in awarding itself 10/10 points for Originality and Creativity. It should be noted as an aside that the excerpts that GPT-4 generated were absolutely dull and banal.

The final stage in prompting was to have it compare excerpts. GPT-4 preferred Kingsolver's excerpt to Diaz at first, but after a re-prompting, it suddenly preferred Diaz's. This is first, indicative of GPT-4's forgetfulness of previous promptings (it had awarded Kingsolver a perfect score and Diaz's less than one), and its inconsistency in accessing work. Finally, I had GPT-4 compare and evaluate the pre-written excerpts to its self-generated creative work. GPT-4 preferred its own writing most of the time, arguing that its own self-generated work was more "literary", "elegant" and "creative". But upon further prompting, GPT-4 would waffle a bit, sometimes arguing that it was impossible to determine which was the better-written piece and other times completely forgetting its self-generated metric for evaluating written work and coming up with new metrics.

Examples

Here are a few images of the testing that demonstrate how GPT-4 would provide inconsistent and opposite answers to prompts:

After further prompting, it generated this (how very diplomatic of GPT-4):

Conclusion

While some have hoped that GPT-4 (which is able to process significantly more tokens than ChatGPT) might be used to aid in the assessment of written work, I would argue from my research this would be ill-advised. GPT-4 was entirely inconsistent in its evaluations of written work and in assigning different letter grades to a text upon re-prompting. What is perhaps even more disappointing is its inability to distinguish the literary quality of an excerpt from a Pulitzer Prize-winning novel and its own mediocre, and formulaic work (unless there is someone arguing that GPT-4's current creative capacity is comparable to that of a Pulitzer Prize-winning novelist—but I think even a cursory reading of what it generated would silence that notion). Putting aside its inconsistency in assigning scores, the fact that it assigns comparable scores to works of wildly different quality, and sometimes even gives itself the superior score, the results are indicative of how incapable GPT-4 is of truly processing and evaluating written work at this current moment.

Recommendations

Firstly, I would recommend that further work be done on LLMs before teachers and academics use GPT-4 as a means of assessing work or generating original content aside from banal and rote writing. Secondly, as a comment about broader prompt engineering, it is remarkable how quickly GPT-4 forgets earlier prompts. It will be essential that later iterations of Open AI develop greater continuity across promptings. Finally, GPT-4 hallucinates and hallucinates badly. It is unable of recalling earlier scorings, and when asked why it gave two different scores for the same piece, its response is to hallucinate once again.

References

GPT-4 and ChatGPT
General Research on the potential uses of GPT-4
Inspiration partially from Microsoft's "Sparks of Artificial General Intelligence: Early experiments with GPT-4" and it's research on GPT-4 and prompt engineering

Acknowledgements

Thanks to Professor Elkins and Professor Chun for their tremendous help and answering a inordinate amount of emails.