Evaluating Generative LLMs: A Probabilistic Scoring Approach

Jeremy Kulcsar

jeremy.kulcsar@gmail.com

February 28, 2024

1 Introduction

This paper explores the evaluation of Large Language Models (LLMs) within the domain of Generative AI and Retrieval Augmented Generation. The objective is to develop a testing framework that facilitates regular performance assessment of LLMs based on various criteria, including correctness, bias, toxicity, and context relevance.

Drawing inspiration from traditional Software Development CI/CD practices, this testing framework aims to resemble unit testing. However, the inherent nature of LLMs and their answer generation process has revealed that a direct application of conventional scoring principles is inadequate. LLMs operate as probabilistic machines, making it impossible to guarantee consistent responses even when provided with identical inputs. Consequently, traditional scoring approaches only reflect the stochastic nature of LLM output, regardless of answer correctness.

To address this challenge, I propose a novel probabilistic scoring approach inspired by the methodologies employed by statisticians in their studies involving group interrogations. My approach deviates from evaluating a single LLM response and instead advocates for posing the same question multiple times. By examining the disparity among the generated answers, we can derive a score that reflects the variability in LLM responses, providing a more comprehensive evaluation of the performance.

Moreover, this framework emphasizes the importance of precise and unbiased prompt engineering to obtain accurate answers from LLMs. Effective question targeting, while avoiding bias, is a critical component of prompt engineering. By posing a question to an LLM multiple times (e.g., 100 times), we can assess the consistency of its responses. If the majority of responses are the same, it indicates a reliable answer. Conversely, if we observe a distribution of

10 different answers with a proportion of 10% each, it suggests inconsistency in the LLM's output.

In summary, the proposed framework presents a new approach for evaluating the performance of LLMs based on relevant criteria. By employing a probabilistic scoring approach and considering the disparity among generated answers, we can obtain a nuanced understanding of LLM performance. Furthermore, by incorporating constraints on prompt engineering, we ensure precise and unbiased question targeting, enabling more accurate evaluations. We anticipate that this framework will contribute to the advancement of LLM performance measurement in Generative AI research.

2 Background

In the project from which this framework is imagined and applied, a method called Retrieval Augmented Generation was used, where a user query is embedded to find closest pieces of information that underwent the same embedding within a knowledge base. However, the method shown will not be impacted by the context retrieval, but rather see it as an extension. This method can provide a more suited answer to a given user query, especially when it is about a closed domain, but it is also room for more errors and accidents, either by the increased complexity of the process, or by the multiple LLM queries which can have a probabilistic snowball effect of hallucinations.

The utilization of LLMs carries inherent risks that merit careful consideration. These risks include, but are not limited to, the following:

- Open-domain hallucination: LLMs may provide answers with unwarranted confidence that are factually incorrect or lack substantial foundation.
- Closed-domain hallucination: LLMs may generate responses that surpass the boundaries of the user query, the given data scope, and/or the retrieved context, leading to undesired outputs.
- Toxicity: LLMs can produce answers that contain biased, harmful or offensive content, potentially causing distress or harm to users.
- Data leakage: LLMs may inadvertently disclose sensitive or confidential information in their generated responses, posing risks to data privacy and security.
- Copyright risk: LLMs carry the risk of inadvertently including copyrighted content in their generated responses, potentially violating intellectual property rights.

Moreover, as an extension of the first score proposed here, we will also discuss the evaluation of the context retrieval process, necessitating an examination of the following factors:

- Context relevance: The retrieval of context must be assessed to ensure its alignment with the user query, avoiding the inclusion of irrelevant information that may hinder the accuracy and appropriateness of LLM-generated responses.

- Context adherence: LLMs should adhere to the retrieved context and generate responses that are consistent with the information provided, ensuring coherence and relevance in the generated content. This point will help in reducing closed-domain hallucinations.

These considerations underscore the need for vigilant evaluation and risk mitigation strategies when employing LLMs in practical applications. To address them comprehensively, a proposed scoring model will be presented, offering a systematic approach to address the identified points. Moreover, this scoring model aims to encompass additional factors, as long as they can be represented as relevant factors by the user.

3 The scoring model

The objective of the proposed model is to emulate the principles of a unit test. In this approach, a predetermined set of n key questions is generated in advance, each accompanied by its corresponding human evaluation. The main goal is to generate two numbers:

- a score that evaluates how well the LLM performs following one specific factor of relevance (e.g. correctness, bias, toxicity, etc.)

- a score that accurately reflects the degree of similarity between the output of the LLM and the human expected responses following the factor of relevance

To account for the probabilistic nature of the model, each key question is prompted m times. This repeated prompting enables the acquisition of multiple answers for the same initial question. By aggregating these diverse responses, a more probabilistic assessment of the LLM's performance can be obtained.

Note that throughout this paper, I talk about "the LLM's performance". This encompasses any LLM-based ensemble of models, RAG-enhancements, etc. As long as there is a user query at the beginning, and a response at the end.

3.1 Measures and variables

Let's first begin by defining the measures and variables that will be used for the scoring model:

- The user query Q, which is also called the key question

- The number of key questions n

- The factor of relevance F
- The LLM's generated answer A, which is a function of the key question Q and the position $j \in \{1, \cdots, m\}$ in the number of times the question was asked
- The checker LLM's evaluation E, which is a function of the main LLM's generated answer A (thus Q) and the factor of relevance F
- A probability measure $\mathbb{P}$ which will be associated with the LLM's output

3.2 Problem formulation and objective

Let's say that we have n key questions $\{Q_i\}_{i \in \{1, \cdots, n\}}$. Each key question Q_i will be repeated m times, which will lead to m answers $\{A_j(Q_i)\}_{j \in \{1, \cdots, m\}}$.

Each element of the set $\{Q_i,\ A_j(Q_i) \mid i \in \{1, \cdots, n\},\ j \in \{1, \cdots, m\}\}$ will be evaluated following a factor of relevance F by a checker LLM. The checker LLM will be asked to evaluate the main LLM's response and return either 0 or 1. 0 means that the performance regarding the factor of relevance is bad (e.g. answer correctness is low), and 1 means that it is good (e.g. answer correctness is high).

In summary, for each key question Q_i, we will get:

- m answers $\{A_j(Q_i)\}_{j \in \{1, \cdots, m\}}$
- m checker LLM evaluations $\{E_j(F,\ Q_i) \in \{0, 1\}\}_{j \in \{1, \cdots, m\}}$

Thus, the first piece we can calculate is a probabilistic score of the LLM over a factor of relevance F for a specific key question Q_i. Mathematically, it is the *conditional* probability of having a positive checker LLM evaluation knowing the key question:

$$\mathbb{P}(E(F,\ Q_i) = 1 \mid Q_i)$$

The second piece will be the *unconditional* probability, describing the generalised probabilistic score of the LLM over a factor of relevance F regardless of the key question:

$$\mathbb{P}(E(F,\ Q_i) = 1)$$

These probabilistic scores can be used for the following:

- The unconditional score gives a first score that truly evaluates an LLM's performance, taking into account its probabilistic nature. This can act on its own as an "unsupervised" observed score, giving an idea of the LLM's performance under a factor of relevance in general.

- If the human review can provide a binary output for each question (e.g. write 1 if a question is harmful otherwise 0), then we can get a tangible evaluation by simply making the difference between the human score and the found probability. In that case, the conditional probability should be used.

3.3 Approach

The first step is to calculate $\mathbb{P}(E(F,\ Q_i) = 1 \mid Q_i)$ for a given key question. This one is quite straightforward given how the problem was formulated:

$$\mathbb{P}(E(F,\ Q_i) = 1 \mid Q_i) = \frac{\sum_{j=1}^{m} E_j(F,\ Q_i)}{m}$$

We just need to count how many times the checker LLM returned 1 for the specific question Q_i, and sum over m, namely how many trials have been done.

However, it is not certain we could get $\mathbb{P}(E(F,\ Q_i) = 1)$ by simply extending this approach and summing positives for every question. $E(F,\ Q_i)$ is indeed different for every question Q_i, so it is legitimate to assume the computation isn't that straightforward. In any case, we will rigorously find the answer.

We can use Bayes's Theorem to express the conditional probability as a function of the unconditional one:

$$\mathbb{P}(E(F,\ Q_i) = 1 \mid Q_i) = \frac{\mathbb{P}(Q_i \mid E(F,\ Q_i) = 1) * \mathbb{P}(E(F,\ Q_i) = 1)}{\mathbb{P}(Q_i)}$$

Which gives:

$$\mathbb{P}(E(F,\ Q_i) = 1) = \frac{\mathbb{P}(E(F,\ Q_i) = 1 \mid Q_i) * \mathbb{P}(Q_i)}{\mathbb{P}(Q_i \mid E(F,\ Q_i) = 1)}$$

3.3.1 $\mathbb{P}(Q_i)$

The term $\mathbb{P}(Q_i)$ is simply the proportion of Q_i among all questions $Q_i,\ i \in \{1, \cdots, n\}$:

$$\mathbb{P}(Q_i) = \frac{1}{n}$$

3.3.2 $\mathbb{P}(E(F,\ Q_i) = 1 \mid Q_i)$

As shown in the previous paragraph, this term can be expressed as the following:

$$\mathbb{P}(E(F,\ Q_i) = 1 \mid Q_i) = \frac{\sum_{j=1}^{m} E_j(F,\ Q_i)}{m}$$

3.3.3 $\mathbb{P}(Q_i \mid E(F,\ Q_i) = 1)$

We can intuitively find the expression by rephrasing the meaning of the conditional probability: given n key questions Q_i that were reprompted m times each, resulting in a set of m answer $E(F,\ Q_i)$ each, we want to know what is the probability of picking one of the n questions knowing that $E(F,\ Q_i) = 1$.

In other words, we want to know what is the proportion of prompt trials from Q_i that returned 1:

$$\{\text{trials from } Q_i \ / \ E_j(F,\ Q_i) = 1,\ i \text{ fixed},\ j \in \{1, \cdots, m\}\}$$

Among all prompts trials that returned 1:

$$\{\text{trials from } Q_i \ / \ E_j(F,\ Q_i) = 1,\ i \in \{1, \cdots, n\},\ j \in \{1, \cdots, m\}\}$$

For a given key question Q_i, we can get the number of prompt trials where $E_j(F,\ Q_i) = 1$ by simply summing:

$$\sum_{j=1}^{m} E_j(F,\ Q_i)$$

The total number of prompt trials where $E_j(F,\ Q_i) = 1$ over all questions can be given by:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} E_j(F,\ Q_i)$$

Thus, we can express the conditional probability $\mathbb{P}(Q_i \mid E(F,\ Q_i) = 1)$ as the following:

$$\mathbb{P}(Q_i \mid E(F,\ Q_i) = 1) = \frac{\sum_{j=1}^{m} E_j(F,\ Q_i)}{\sum_{i=1}^{n} \sum_{j=1}^{m} E_j(F,\ Q_i)}$$

3.3.4 Putting it together

We can now change the terms from the equation we got using Bayes's Theorem and find the following:

$$\mathbb{P}(E(F,\ Q_i) = 1) = \frac{\sum_{j=1}^{m} E_j(F,\ Q_i)}{m} * \frac{1}{n} * \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} E_j(F,\ Q_i)}{\sum_{j=1}^{m} E_j(F,\ Q_i)}$$

Which simplifies to:

$$\mathbb{P}(E(F,\ Q_i) = 1) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} E_j(F,\ Q_i)}{nm}$$

3.3.5 Observations

The probability $\mathbb{P}(E(F,\ Q_i) = 1)$ gives a probabilistic score on the LLM's performance. It is independent of the initial question Q_i, despite depending on $E(F,\ Q_i)$ which depends on Q_i.

The implication is counter intuitive: $\mathbb{P}(E(F,\ Q_i) = 1) = \mathbb{P}(E(F,\ Q_j) = 1)$ for any Q_i and Q_j, regardless of what $\{E_k(F,\ Q_i)\}_k$ and $\{E_k(F,\ Q_j)\}_k$ look like.

On the other hand, it might appear as simply counting how many times we see 1 across all results, and divide by the total number of elements. This section can therefore be seen as the mathematical confirmation that this approach is valid, and no dependence to the initial question should be considered.

What we can get from this is, $\mathbb{P}(E(F,\ Q_i) = 1)$ is a representative and global score across all the questions, focusing solely on the performance of the LLM.

3.4 Probabilistic scores

3.4.1 Unconditional probability

If we name S_u the probabilistic score resulting from the evaluation of an LLM over a factor of relevance F, we can write:

$$S_u(F) = \mathbb{P}(E(F,\ Q_i) = 1)$$

Where:

- Q_i is any element from the set of n key questions, $Q_i,\ i \in \{1, \cdots, n\}$
- F is the factor of relevance
- E is the response from the checker LLM on whether the answer from Q_i respects F
- $\mathbb{P}$ is the probability measure associated with the main LLM, used on E

Then $S_u(F)$ represents the general performance of the main LLM under the factor of relevance F.

3.4.2 Conditional probability

If we name S_c the probabilistic score resulting from the evaluation of an LLM over a factor of relevance F knowing the question Q_i, we can write:

$$S_c(F, Q_i) = 1 - |\mathbb{P}(E(F,\ Q_i) = 1 \mid Q_i) - H(F, Q_i)|$$

Where:

- Q_i is the question for which the probability is calculated

- F is the factor of relevance
- E is the response from the checker LLM on whether the answer from Q_i respects F
- H is the human estimation of how the answer from Q_i will respect F (e.g. expected toxicity knowing Q_i)
- $\mathbb{P}$ is the probability measure associated with the main LLM, used on E

Then $S_c(F, Q_i)$ represents a per-question measure of how the model's score is close to what the human expects, knowing the question.

3.4.3 Total probabilistic score

After having computed both scores, we can establish the total probabilistic score S_t as the following:

$$S_t(F) = \frac{S_u(F) + \frac{\sum_{i=1}^{n} S_c(F,Q_i)}{n}}{2}$$

Which can be rewritten as:

$$S_t(F) = \frac{nS_u(F) + \sum_{i=1}^{n} S_c(F, Q_i)}{2n}$$

4 Example: Evaluating the toxicity of a model

For this example, we will evaluate the toxicity of a model. Particularly, we will evaluate whether the model is likely to generate harmful content when it returns an answer to a given user query.

4.1 Methodology

The methodology will respect the following steps:

1. Prepare a set of user queries, where the toxicity is known in advance (e.g. harmful content)
2. Let one or many humans evaluate the toxicity of each of them with a score between 0 and 1, 0 meaning no presence of harmful content and 1 meaning very evident presence of harmful content (note: the more people do this part, the better the human evaluation)
3. Run them through the model a specific number of times each
4. Evaluate the toxicity with a specialised checker LLM that was built for checking harmful content and returning 0 if it sees none, otherwise 1
5. Build the conditional, unconditional and total probabilistic scores
6. Use them to evaluate the toxicity of the model regarding harmful content

4.2 Application