Citation Benchmark

Ali Nazari, Illia Hashemirad, Shayan Salehi,
Mehdi Lotfian, Masih Najai, Amir Mohammad Fakhimi
<ali.nazari.8102@gmail.com><s.salehi1381@gmail.com>
<mehdilotfian.meloent@gmail.com><najafim2002@gmail.com>
<fakhimi.amirmohamad@gmail.com>
Sharif University of Technology, Tehran

August 10, 2024

Abstract

In this paper, we propose a comprehensive approach to enhance the evaluation of language models by focusing on the generation and utilization of reference citations. The evaluation of model outputs is crucial in determining their quality and relevance, yet current methodologies often rely heavily on human judgment, which can be time-consuming and unreliable. To address this issue, we have developed a standardized benchmark framework, including novel metrics that emphasize automated, correlation-based assessments of sentence fragments within documents. This framework includes optimized input processing techniques, structured input formats, and improved methods for extracting references from web pages. Leveraging diverse datasets such as ASQA, QAMPARI, ELI5, and Wikidata5m, our approach aims to create a robust and scalable evaluation system. By employing models such as Vicuna and LLaMA alongside established models like GPT-4, we study the effectiveness of these strategies in reducing human intervention and enhancing the accuracy of model evaluation.

1 Introduction

The rapid advancement of natural language processing (NLP) models has brought about significant improvements in the generation of human-like text, enabling applications across various domains, from automated customer service to content creation. However, evaluating these models remains a challenging task, particularly when assessing the relevance and accuracy of the generated content. Traditional evaluation methods often rely on human judgment, which can be subjective and resource-intensive. Furthermore, standardized benchmarks and metrics are needed to ensure the objective comparison of model performance.

This paper addresses these challenges by proposing a novel framework for evaluating language models, with a particular focus on the generation and use of reference citations. The importance of references in assessing the quality of model outputs cannot be overstated, as they provide an objective basis for evaluating the factual accuracy and relevance of generated text. Our approach aims to create a standardized benchmark that minimizes human intervention by introducing automated metrics that leverage correlation checks for sentence fragments within documents.

The proposed framework is built upon several key objectives:

- We aim to establish a standardized procedure for executing benchmarks, ensuring consistency and comparability across different models.

- We introduce new metrics designed to reduce human involvement in the evaluation process, particularly by implementing automated checks for the correlation of meaning within sentence fragments.

- We explore methods for optimizing input processing, such as reducing the token count by providing relevant snippets and summaries.

- We focus on developing structured input formats that improve the consistency and accuracy of model outputs, particularly in scenarios requiring multiple document references.

To validate our approach, we utilize a diverse set of datasets, including ASQA, QAMPARI, ELI5, and Wikidata5m, which cover a wide range of question types and information needs. These datasets provide a comprehensive foundation for testing our evaluation framework across different content domains and question formats. Moreover, we explore the use of various language models, including Vicuna, LLaMA, and established models like GPT-4, to assess the effectiveness of our proposed methods.

In the following sections, we will detail the methodologies used in the development of our benchmark framework, the implementation of new evaluation metrics, and the results of our experiments with different models and datasets. Through this paper, we aim to contribute to the field of NLP by providing a more robust, scalable, and objective approach to the evaluation of language models, ultimately improving their reliability and applicability in real-world scenarios.

2 Related Work

This section is for mentioning related works.

3 Implementation

The implementation section.

4 Results

The result section.

5 Conclusions

The conclusions section.

References

[1] L. Lamport. *LaTeX: a document preparation system: user's guide and reference manual*. Addison-wesley, 1994.