PAPER 1

RAGAS:Automated Evaluation of Retrieval Augmented Generation

Shahul Es†, Jithin James†, Luis Espinosa-Anke∗♢, Steven Schockaert∗

Introduction:

This paper, provides us metrics which can be used to evaluate these different dimensions of RAG without having to rely on ground truth human annotations. We therefore focus on metrics that are fully self-contained and reference-free.

Scope and Objectives:

The literature review's scope centers on evaluating RAG systems' performance, steering clear of reliance on human-generated reference responses. This deliberate choice facilitates expedited evaluation cycles. The specific objectives encompass an exploration of how RAG systems harness LMs as knowledge repositories and a discerning identification of their inherent limitations.

Critical Analysis:

A critical analysis of RAG systems is presented, delving into both their strengths and weaknesses. RAG systems exhibit proficiency in leveraging LMs for extracting pertinent information from external documents, concurrently mitigating the risk of generating inaccurate or fictional content. However, challenges emerge in responding to inquiries about events occurring post-training data and in grappling with the memorization of rarely mentioned knowledge.

Identification of Trends and Gaps:

Within the landscape of RAG systems, prevalent trends emerge, such as the integration of standard LMs with retrieved documents to enhance strategies. The pivotal role of the retrieval component in furnishing contextual understanding for the LM is highlighted. Simultaneously, the review identifies gaps, including the complexity of evaluating RAG systems encompassing retrieval model performance and prompt formulation, along with a dearth of a comprehensive approach in existing literature devoid of human annotations.

Methodological Quality Assessment:

- **Faithfulness** refers to the idea that the an swer should be grounded in the given context. This is important to avoid hallucinations
- **Answer Relevance** refers to the idea that the generated answer should address the actual ques tion that was provided.
- **Context Relevance** refers to the idea that the retrieved context should be focused

Synthesis of Findings:

The synthesis of findings from the literature review underscores the significance of RAGAS as an enabler for the automated assessment of RAG systems. This methodology unravels insights into the delicate trade-offs between retrieval quality and LM performance, contributing substantially to comprehending the challenges and opportunities embedded in RAG research.

Conclusion:

The authors claim this method to be extremely efficient in evaluating RAG apps.This framework is easy to use and can provide deverlopers of RAG sys tems with valuable insights, even in the absence of any ground truth. We plan to use the same while evaluating our various LLM chatbots to compare and contrast between them.

References: 2309.15217.pdf (arxiv.org)

PAPER 2

HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models

AUTHORS - Junyi Li

Introduction:

To un derstand what types of content and to which ex tent LLMs are apt to hallucinate, we introduce the Hallucination Evaluation benchmark for Large Language Models (HaluEval), a large collection of generated and human-annotated hallucinated samples for evaluating the perfor mance of LLMs in recognizing hallucination.

Scope and Objectives:

The study ambitiously aims to evaluate the extent to which LLMs give rise to hallucinated content. This involves the establishment of a benchmark, HaluEval, for a nuanced assessment of hallucinations. The specific objectives include the creation of HaluEval and the provision of a curated collection comprising both hallucinated and normal samples for thorough analysis.

Critical Analysis:

In evaluating the research, notable strengths and weaknesses come to light. The paper introduces a distinctive two-stage framework for generating hallucinated samples, involving initial sampling followed by meticulous filtering. The incorporation of human-annotated examples bolsters the robustness of the validation process. However, the paper falls short in delving deeply into the underlying causes of hallucinations, primarily focusing on detection rather than mitigation.

Methodology:

Hallucination Evaluation benchmark for Large Language Models (HaluEval): a large collection of 35,000 hallucinated/normal samples for LLMs analysis and evaluation. HaluEval includes 5,000 general user queries with ChatGPT responses and 30,000 task-specific examplesthe deficient performance of LLMs in recognizing hallucinations can be improved by pro viding explicit knowledge and adding intermediate reasoning steps

Results:

The study sheds light on the prevalence of hallucinated responses in ChatGPT, a widely utilized LLM, revealing an occurrence rate of approximately 19.5%. These hallucinations often involve the generation of fabricated information pertinent to specific topics, adding granularity to the understanding of LLM behavior.

Identification of Trends and Gaps:

Trends surface as LLMs display a tendency to generate unverifiable content, especially when retrieving information from external sources. Contextual variations impact the frequency of hallucinations, indicating a nuanced landscape. However, the paper falls short in plumbing the depths of the root causes of hallucinations, necessitating further research to bridge these critical gaps.

Methodological Quality Assessment:

The paper's methodology blends automatic generation and human annotation, presenting a comprehensive evaluation approach. The dependence on ChatGPT's capacity, however, introduces a noteworthy caveat, highlighting the nuanced relationship between methodology and the quality of hallucinated samples. This intricate balance underscores the complexity of assessing LLM behavior.

Conclusions

Benchmark helps researchers address LLM hallucination issues. It provides annotated samples for analyzing content types that cause hallucinations. Researchers can also evaluate LLMs' ability to recognize hallucinated content. Additionally, our benchmark, paired with human annotation, assesses whether LLM outputs contain hallucinations

References: pdf (openreview.net)

PAPER 3

Itihasa: A large-scale corpus for Sanskrit to English translation

Rahul Aralikatte1 Miryam de Lhoneux1 Anoop Kunchukuttan2 Anders Søgaard1

Introduction:

The research introduces “Itihāsa,” a large-scale dataset containing 93,000 pairs of Sanskrit shlokas and their English translations from two Indian epics, The Rāmāyana and The Mahābhārata1. The purpose is to aid the translation process and democratize the knowledge within these texts.

Scope and Objectives:

The scope encompasses the creation of a comprehensive Sanskrit-English translation corpus. Objectives include facilitating better translation systems for Sanskrit and potentially creating a parallel corpus for other Indian languages.

Critical Analysis:

The paper’s strength lies in its large dataset size and the methodological rigor in dataset preparation. However, the complexity of Sanskrit and the nuances of translation pose challenges, as indicated by the performance of standard translation models on this corpus.

Identification of Trends and Gaps:

A trend is the focus on improving translation systems for Sanskrit, a morphologically rich language. The gap identified is the lack of large-scale, high-quality translation datasets, which “Itihāsa” aims to fill.

Methodological Quality Assessment:

The methodological quality is high, with meticulous data preparation involving both automated OCR extraction and manual inspection. The accuracy of the extracted text is commendable, though some OCR errors persist.

Synthesis of Findings:

The findings underscore the complexity of Sanskrit translation and the need for advanced models capable of handling the language’s intricacies. The “Itihāsa” dataset is a step towards addressing these challenges.

Conclusion:

The literature review reveals that “Itihāsa” is a valuable resource for advancing Sanskrit translation research. Its significance lies in its potential to improve translation accuracy and contribute to the preservation and accessibility of ancient texts.

8. References: 2106.03269.pdf (arxiv.org)

PAPER 4

Advanced RAG Techniques: an Illustrated Overview

IVAN ILIN

1. Introduction:

We're diving into this topic called Advanced Retrieval Augmented Generation (RAG) Techniques. These techniques mix up finding information with creating new things, all to make our language models cooler.

2. Scope and Objectives:

We're zooming in on how RAG works in LlamaIndex and LangChain. We want to understand their ins and outs, and how they can be useful in real life. Our plan is straightforward: figure out what RAG is all about, get the hang of it, and see how it's making a difference in how computers understand language.

3. Critical Analysis:

As we look into these papers, we're discovering the good and not-so-good sides. RAG techniques are awesome at finding what you're looking for, but sometimes they get a bit tricky with complicated stuff. These studies are like detectives on a mission. They've got a solid plan, exciting results, and ideas that make us curious, especially for smaller language models.

4. Identification of Trends and Gaps:

Here's the buzz - everyone's excited about mixing RAG with vector databases. It's like leveling up your search skills. But, we need to keep it simple and not make things too complicated. The papers spill the beans on things we still need to figure out. Maybe we need easier ways for more people to join in. And, of course, we need to balance speed with keeping things simple.

5. Methodological Quality Assessment:

These studies are doing it right. They're not just telling us stuff; they're showing us with clear explanations, good experiments, and real-life examples. Everything seems legit, but hey, the world of open-source projects is moving fast. We’ve got to keep an eye on things changing.

6. Synthesis of Findings:

We've found some real gems! RAG is like magic, making computers better at understanding and using language. It's not just theory; these findings are like connecting the dots between what we know and what we can actually do.

7. Conclusion:

In a nutshell, RAG techniques are like the guiding stars shaping the future of how computers understand language. The mix of finding and creating is a game-changer.

This literature review isn’t just words on paper; it's like a compass pointing your technical paper to exciting new territories.

8. References: Advanced RAG Techniques: an Illustrated Overview | by IVAN ILIN | Towards AI