Benchmarking RAG

SOTA2 - Group C

Anton Kesy, Étienne Muser, Katharina Schindler, Lukas Fehrenbacher, Nico Ruschmann

Offenburg University of Applied Sciences

WS 2024/2025

Herausforderungen bei der Evaluierung von RAG

- geeignete Metriken finden um die Qualität der Antworten zu bewerten
  - traditionelle Metriken wie Exact Match oder F1 sind oft unzureichend
  - semantische Ähnlichkeit verwenden -> robustere und automatisierte Bewertungsmethoden
- Inkonsistente experimentelle Setups
  - große Vielfalt an experimentellen Setups
  - Mangel an Standardisierung erschwert Vergleich und Verständnis
- Wahl geeigneter Datensätze
  - bestehende Datensätze sind nicht geeignet, da LLMs bereits über umfangreiches Wissen verfügen
  - neue Datensätze speziell für RAG-Systeme entwickeln
- Einfluss der Retrievalqualität verstehen
  - Retrievalqualität hat erheblichen Einfluss auf Leistung von RAG-Systemen
  - Verwendung von State-of-the-Art-Retrievalsystemen und Rerankern kann Qualität verbessern
  - Beziehung zwischen Retrievalqualität und LLM-Größe untersuchen ist wichtig

Herausforderungen bei der Evaluierung von RAG

- Notwendigkeit einer ganzheitlichen Bewertung
  - Qualität des Retrievers und des Generators berücksichtigen
  - Einfluss von Retrievalqualität auf die Antwortqualität
- Herausforderungen der domänenübergreifenden Generalisierung
  - viele Studien konzentrieren sich auf domänenspezifische Datensätze
  - Generalisierung über verschiedene Domänen hinweg bewerten
- Grenzen von LLM-basierten Bewertungen
  - Verzerrungen und Ungenauigkeiten in LLM-generierten Labels
  - Verwendung mehrerer LLMs oder Kombination von LLM- und menschlichen Bewertungen

BERGEN

- RAG Benchmarking Library
- Open Source
- Unterstützung von vielen Konfigurationen
- Sehr einfach erweiterbar (YAML File)

BERGEN - Konfigurationen

- Retreivers
- Rerankers
- LLMs
- Datasets
- Metrics
- Fine-Tuning

BERGEN - Metrics

- Surface-based
  - Match
  - Exact Match
  - Presicion
  - Recall
  - F1
- LLM-based
  - BEM
  - GPT-4
  - LLMeval

Anwendungsgebiete von *BERGEN*

- Vergleich verschiedener RAG-Systeme:
  - ermöglicht die standardisierte Bewertung von RAG-Systemen
    - verschiedene Komponenten (Retriever, Reranker, LLM)
- Analyse von RAG-Komponenten
  - unterstützt die Untersuchung des Einflusses verschiedener RAG-Komponenten auf die Leistung
    - Analyse von Komponenten (R,R,L)
- Mehrsprachige RAG-Experimente
  - unterstützt mehrsprachige Datensätze

Vorteile von *BERGEN*

- Reproduzierbarkeit
  - standardisierte Umgebung
- Flexibilität und Erweiterbarkeit
  - basiert auf dem Hugging Face Hub
    - einfach um neue Modelle, Datensätze und Metriken
- Umfassende Unterstützung
  - große Auswahl an Modellen, Datensätzen, Metriken und Trainingskonfigurationen
- Vereinfachung von Experimenten

Nachteile von *BERGEN*

- Begrenzte Rechenleistung
  - Feinabstimmung großer LLMs kann rechenintensiv
- Fokus auf Wikipedia
  - Experimente auf Wikipedia-basierte Sammlung
  - Leistung auf anderen Datensätzen noch unklar
- Hauptsächlich QA-fokussiert
  - Der Schwerpunkt von BERGEN liegt derzeit auf QA-RAG
  - Die Anwendbarkeit auf andere RAG-Anwendungen muss noch weiter erforscht werden

Tolles Paper

- Super kurz
- Klar strukturiert
- Gute Beispiele

RAGBench: Datensatz & TRACe Framework

Bewertung von Retrieval-Augmented Generation (RAG)-Systemen

RAGBench Datensatz

- Neuer, großer Datensatz: 100.000 Beispiele
- Branchenspezifisch: 5 Domänen (Biomedizin, Allg. Wissen, Recht, Kundensupport, Finanzen)
- Reale Daten: z.B. Benutzerhandbücher
- Variiert in: Kontextlänge, Anzahl Dokumente, Generierungsmodell

TRACe Evaluierungsframework

- 4 Metriken zur RAG-Bewertung
  - **Kontextrelevanz**: wie relevant ist der abgerufene Kontext für die Eingabefrage
  - **Kontextnutzung**: wie viel des abgerufenen Kontexts wird vom Generierungsmodell verwendet
  - **Adhärenz** (Faithfulness): wie gut ist die Antwort des Modells an den bereitgestellten Kontext gebunden und ob Halluzinationen vermieden werden
  - **Vollständigkeit**: wie vollständig berücksichtigt die Antwort alle relevanten Informationen im Kontext

Anwendungsbereiche von RAGBench

Industrie:

- **Chatbots und digitale Assistenten**: Domänenspezifische Antworten für Gesundheitsberatung, Finanzanalysen oder **juristische Informationen**.
- **Kundenservice**: Automatisierung von FAQs und technischem Support.

Wissenschaft:

- **Medizinische Forschung**: Beantwortung komplexer Fragestellungen aus der Biomedizin.
- **Technischer Support**: Analyse und Lösung technischer Probleme.

Vorteile und Nachteile von RAGBench

Vorteile:

- **Standardisierung**: Einheitliches Framework für die Bewertung von RAG-Systemen mit klaren Metriken (Relevance, Utilization, Adherence, Completeness).
- **Vielfältige Anwendungen**: Abdeckung von 5 Domänen wie Medizin, Finanzen und Recht.
- **Granulare Analysen**: Identifikation von spezifischen Schwächen wie Halluzinationen.
- **Effizienz**: Feinabgestimmte Modelle wie DeBERTa bieten kostengünstige Alternativen zu großen LLMs.

Nachteile:

- **Annotationen**: Abhängigkeit von GPT-4-Labels, die Verzerrungen enthalten können.
- **Hoher Rechenaufwand**: Besonders bei langen Kontexten, z. B. juristische Texte.
- **Relevanz-Bewertung**: Schwierigkeit, semantische Ähnlichkeit von funktionaler Relevanz zu trennen.
- **Limitierte Generalisierbarkeit**: Spezialisierte Anwendungsfälle möglicherweise nicht abgedeckt.

Vergleich: BERGEN und RAGBench

Der Vergleich ist dem Leser als Übung überlassen