



### **Evaluating Retrieval-Augmented Generation for Question Answering with Large Language Models**

Ermelinda Oro, <u>Francesco Maria Granata</u>, Antonio Lanza, Amir Bachir, Luca De Grandis and Massimo Ruffolo

## Introduction

#### • Emergence of RAG Systems:

- Integrate external information retrieval with natural language generation.
- Enhance capabilities of language models for more informative and contextually relevant responses.

#### • Evaluation challenges:

- Difficulty in evaluating performance without ground truth data.
- Impedes accurate assessment of system utility and applicability.

#### Research objectives:

- Investigate reliability and validity of existing evaluation methodologies.
- Examine correlation between various metrics and human evaluations.
- Highlight strengths, limitations, and areas for improvement in evaluation metrics.

#### • Key contributions:

- Comprehensive evaluation framework with state-of-the-art components.
- Comparison of diverse evaluation metrics.
- Rigorous experiments across multiple datasets, including NarrativeQA and FinAM-it.
- Analysis of metric strengths and limitations through correlation analysis.

## Framework for RAG and Evaluation



# **Evaluation strategies**

### Classical Retrieval Stage Metrics:

- Recall@K, Precision@K, mAP
- Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (nDCG)

### Answer Generation Stage Metrics:

- Syntactic metrics: BLEU, ROUGE, Precision, Recall, F1 Score, Exact Match.
- Semantic metrics: BERT Score, BEM Score.

### • LLM-derived Metrics:

- RAG triad: Answer Relevance, Context Relevance, Groundedness.
- Answer Correctness.

### Manual evaluation:

- Conducted by three independent human annotators.
- Evaluation based on relevance, accuracy, and coherence.
- 5-point Likert scale: Very Poor, Poor, Neither, Good, Very Good.
- Resolve discrepancies and ensure unbiased evaluations.

### Datasets

Dataset	Questions	Language	Content Type
NarrativeQA	50	English	Books
NarrativeQA	50	English	Movies
FinAM-it	50	Italian	Financial documents





LBPAM	DOCUMENTO CONTENENTE LE INFORMAZIONI CHIAVE TOCQUEVILLE EURO EQUITY ISR I U
Il presente documento forniace le informationi chiove re legge, hanno lo scopo di alutarni a capire le caratteristici d'investimento.	skilos s queste profette di investmento. Non si tratis di un documente presentativale. La leformationi, prescribe par e, i rischi, i costi e i guedigii a le pedite potenciali di queste prodetto e di alatami a bre un raffronto con alci prodetti
	PRODOTTO
TOCQUEVILLE BURO EQUITY ER, Astone I Comparis della SICAV LEPAM HUNDS Codes ISN: 198010942385 Ideatore LEP AM (1: "Società di gestione" o "LEP AM")	L'Autorité des marchés financiers (AMP) é responsible della régitante d'LEP AM in relations d'presente documento contravente la informazioni chive. LEP AME propo La Bargor Pocale, é autoritata in Francia con il neurono GP. 2000001 e disciplinat dell'Autorité des marchés financiers (AMP).
Sito web: www.bpare.com - Chamaro ii +33 (0) 1 57 24 informazioni	21 00 per ultoriori
	COST ONISTO PRODUCTOR
Tipe: Organismo di investimonto callettivo in valori mo dell'europona.	bilari - Società di investimento a capitale variable, di diritto francese constituito in Frances. Classificazione: Adoni di pasal
Observation de conservant de Comparte à duralités of Contraut de conserver, attaineures na portechigie compar- à quale de la meneral attaines della asse sure e de l'analismentere construction di interviententes socialisme Il Comparte à posities attaines de la meneral de posi- difia socia di posities attaines. Facili d'accourt della socia di posities este conserva, l'attai dessana d'antaine cachades, e sense da conservatio de aprica (const discusso La antal de tout servane de aprica (const discusso La antal de tout servane de aprica (const discusso da da tout servane) de aprica (const discusso de aprica de tout servane de aprica (const discusso de aprica de tout servane) de aprica (const discusso de aprica de tout servane) de aprica (const discusso de aprica de aprica de aprica de aprica (const discusso de aprica de aprica de aprica de aprica (const discusso de aprica de aprica de aprica de aprica (const discusso de aprica de aprica de aprica de aprica (const discusso de aprica de aprica de aprica de aprica (const discusso)) de aprica de aprica de aprica (const discusso) de aprica de aprica de aprica de aprica (const discusso) de aprica de aprica de aprica de aprica (const discusso) de aprica de aprica de aprica de aprica (const discusso) de aprica	And the multi-field assessment during a periodic of productions as negligible supervises of a wei, and performance supervises on a distribution of the supervised structure of the supervised structu
Anterester and a long of all statistics of a long of all statistics of all statistic	
Il Comparto è capasto por almono il 60% al morcati asi investimente (bino al 10% del partimento netto; se nee adottata dalla società di gestiane che gestiane gli OC en priviligiarano il a oblicato e di OC con un apprecio e l'investimante minimo del 73% in tito il di società e in que	zen del pese dell'erreazen. L'opposizione azionanti è ozzanza scamita invezzionati direzzio comitta CNCM+ o fond di si trazta del CC monti, pesso con estimate departiti di appreco in negati di LRP AN e 1924 e quelle consi institucati. Invezzio CC ana si abitattatto negativatattati di RN una quella GLRP AN e 1924 e 1924 e SL compatible con la ficculta di LRP AN, nandei transa fusibazi di arcumenti fisanziari derivati, nel respetto di un en orteret di CC Useni al PNA.
In factore della conditioni di mercano e al fini dalli di credita e alchi stromeni dei mercano monetterio, deran discontrato di stromeni di mercano monetterio, deven discontrato di stromeni di territorio, mercano di strome di stromento, alchi la Grangvaro posi di gualdoni della tenererio, il Comparto centino e di strato per penegare il suo elettitto di gri fomazzio elletteri ano spodei spererio il 2024 del partico il Comparto posi territori e di stroperio di territori il Comparto posi tengene di supporto a strobalto di esti- ti comparto posi fargiare di supporto anto-bitole di con	using do relation, a sufficient adds participants and particle super-site (Comparing to Americe III) and a super-site of Comparing to Americe III and America IIII and America IIIIII and America IIIII and America IIIII and America IIIII and America IIIIII and America IIIII and America IIIIII and America IIIIIIIIIII and America IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

## Metrics

#### • Goal:

- Evaluating the quality of generated answers across the entire pipeline.
- BEM score:
  - Uses a BERT model trained for answer equivalence task.

#### Answer Correctness (RAGAS):

• Employs LLM to extract factual statements and calculates F1 score for factual correctness.

#### • Answer Relevance (RAGAS):

• Computes mean cosine similarities between the original question and artificial questions generated by an LLM based on the predicted answer.

#### • Answer Relevance (TruLens):

• Prompts an LLM to evaluate answer relevance with respect to the input prompt.

#### Spearman Rank Correlation Coefficient:

- Non-parametric measure of statistical dependence between rankings of two variables
- Used to measure the interrelationships and relative effectiveness among various evaluation metrics.



### Prompt

You are a chatbot having a conversation with a human. Given the following extracted parts of a long document and a question, create a final answer. If you don't know the answer, just say that you don't know, don't try to make up an answer. Context: {CONTEXT} Chat history: {CHAT\_HISTORY} Human: {HUMAN\_INPUT} Chatbot :

## Results

• Correlations with human judgement

Metrics	BEM	AR TruLens gpt-3.5- turbo	AR RAGAS gpt-3.5- turbo	AC RAGAS gpt-3.5- turbo	AR TruLens gpt-4-turbo	AR RAGAS gpt-4-turbo	AC RAGAS gpt-4-turbo
Narrative QA (Books)	0.735	0.436	0.234	0.718	0.42	0.15	0.67
Narrative QA (Movies)	0.704	0.565	0.483	0.792	0.213	0.411	0.781
FinAM-it	0.208	0.178	0.153	0.053	0.280	0.230	0.531

### Results



Comparison of Metrics Across Different Datasets and Models

Metrics

# Results

- NarrativeQA Dataset (Books and Movies):
  - Ground truth-based metrics align well with human perception of answer quality.
  - Reference-free metrics (e.g., AR RAGAS) show poor correlation (0.234 for books, 0.483 for movies).

### • FinAM-it Dataset:

- Lower correlations across all metrics.
- Complexity and diversity of financial content pose greater evaluation challenges.

### General Findings:

- All metrics struggle to robustly approximate human evaluation.
- Indicates the need for improvement in evaluation methods, particularly reference-free metrics.

# Conclusions and Future Work

- Ground truth based metric like BEM and AC RAGAS are significantly more robust than the ground truth free metrics.
- Significant challenges in achieving high correlation with human judgments.
- Room for improvement, especially with complex, domain-specific datasets like FinAM-it.
- Improve accuracy and reliability of existing metrics.
- Explore new methodologies to capture qualitative aspects of generated answers.
- Leverage advanced language models for additional context and domain knowledge.
- Develop ensemble or multi-task evaluation approaches.
- Mitigate biases and subjectivity in human annotations.