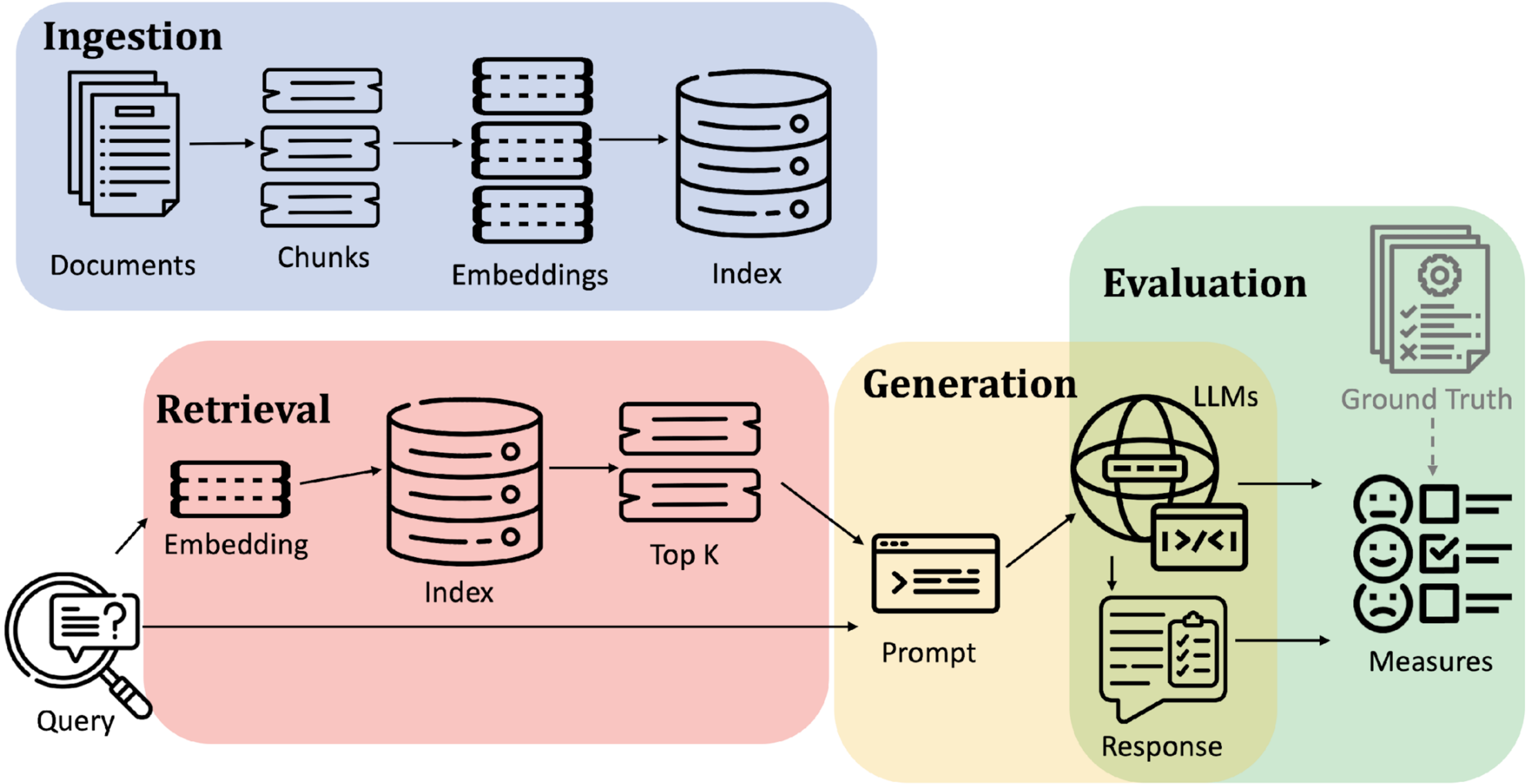
Evaluating Retrieval-Augmented Generation for Question Answering with Large Language Models

Ermelinda Oro, Francesco Maria Granata, 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

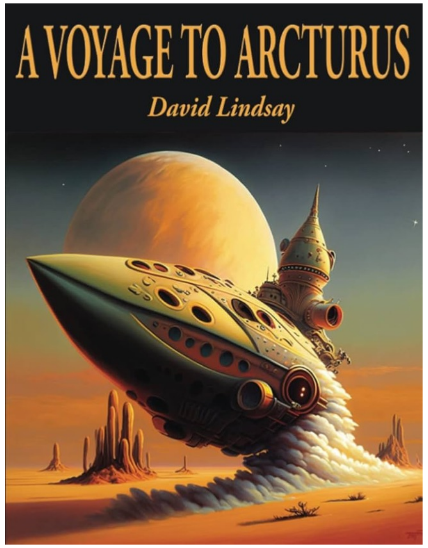


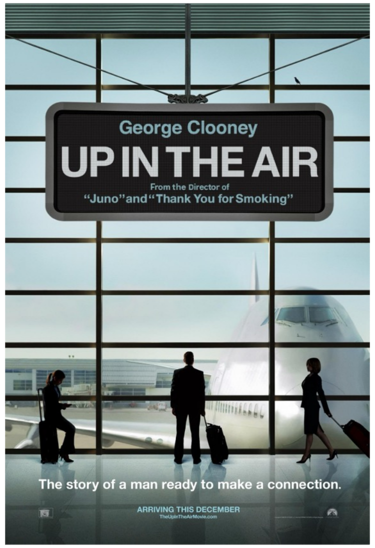
Evaluation strategies

- **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

OBIETTIVO

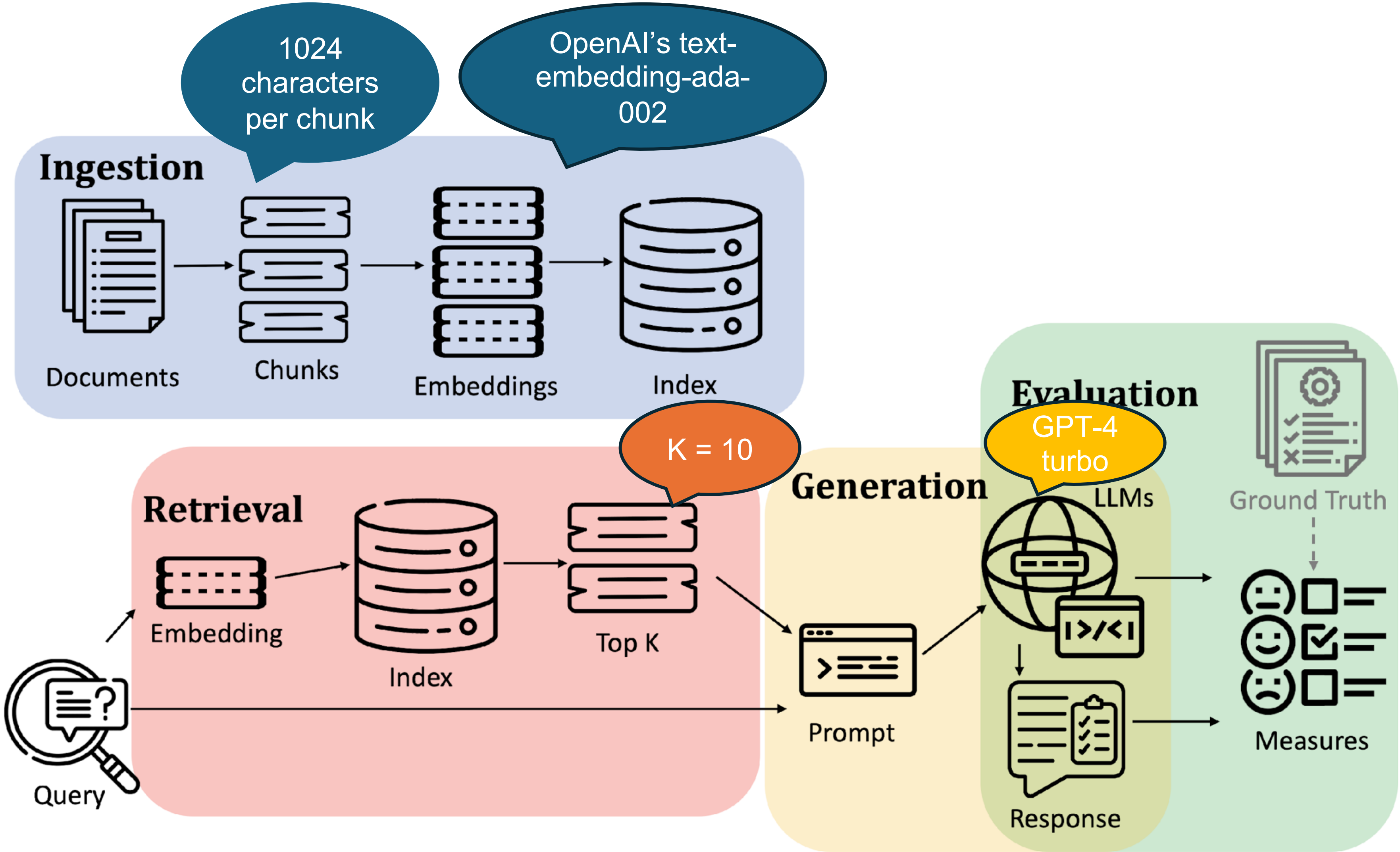
PRODOTTO

COS'È QUESTO PRODOTTO?

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.

Settings



1024 characters per chunk
OpenAI's text-embedding-ada-002
Ingestion
Documents
Chunks
Embeddings
Index
K = 10
Retrieval
Embedding
Index
Top K
Query
Generation
Prompt
GPT-4 turbo
LLMs
Response
Evaluation
Ground Truth
Measures

Prompt

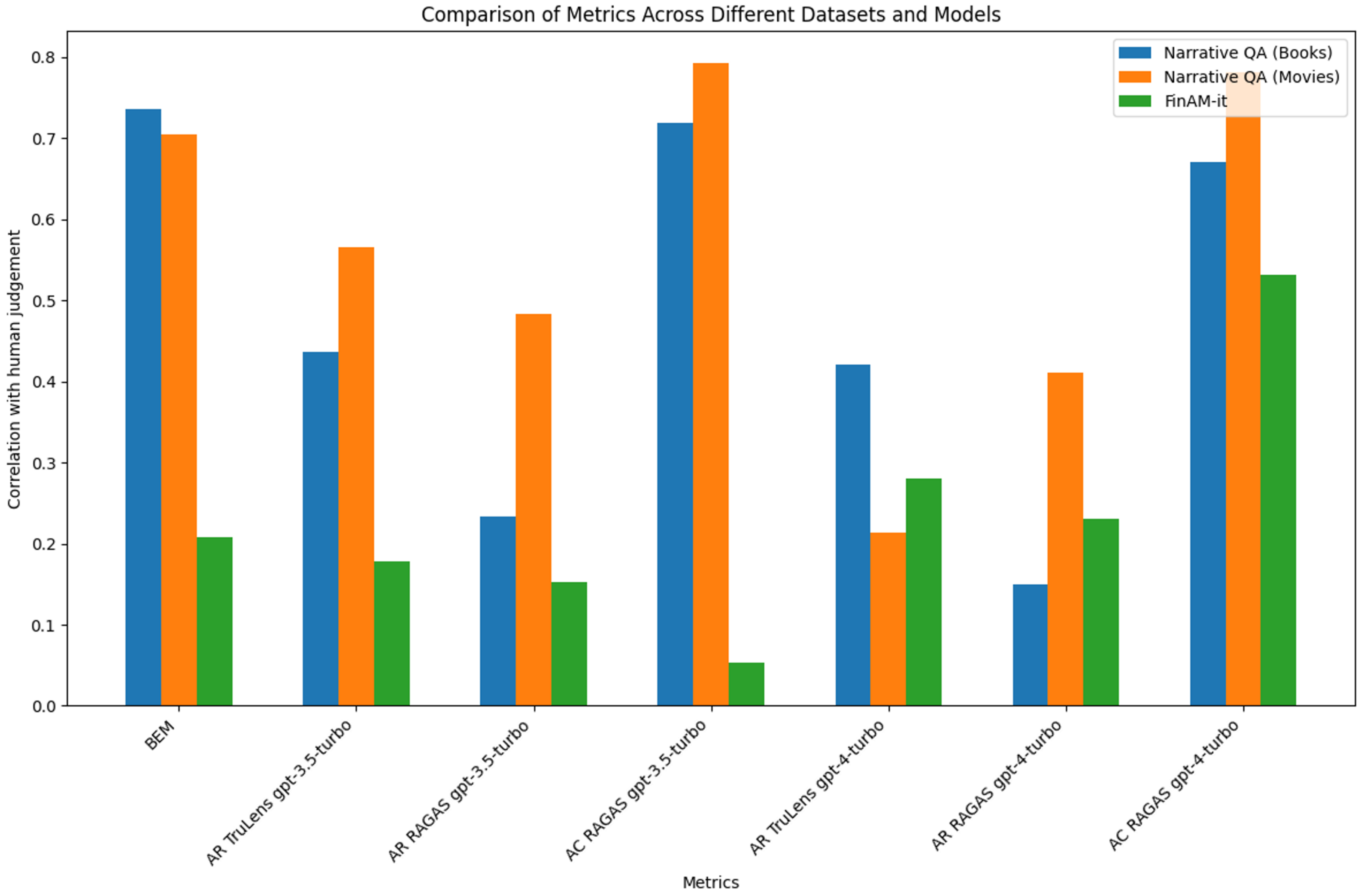
```
You are a chatbot having a
    conversation with a human.
Given the following extracted parts
    of a long document and a question,
     create a final answer.
If you don't know the answer, just
    say that you don't know, don't try
     to make up an answer.
Context: {CONTEXT}
Chat history: {CHAT_HISTORY}
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
Narrative QA (Books)
Narrative QA (Movies)
FinAM-it
Correlation with human judgement
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
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
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.