Columbia University Capstone Project Initial Due Diligence Report

September 21, 2023

Fidelity INVESTMENTS

BUTIS

Table of Contents

Project Description

Fidelity Investments proposes a project to develop an evaluation framework for **Large Language Models (LLMs)** in the context of the financial industry. The aim is to quantify key facets like correctness, sensitivity, and reasoning in LLM-generated content, focusing on financial and regulatory jargon.

Our Understanding

Final Deliverable:

- Create an LLM-produced content evaluation **methodology** and an **automated way** of measuring
  - Define **one or more** metrics
  - Paper about the evaluation framework
  - Model or Process
- (Stretch Goal) A network graph that shows the **relationship** between the different entities

System Input : An output generated by LLM

System Output :

- Evaluation of the response by defined metrics
- Score of aspects such as correctness, reasoning, variation

Industry Solution - Use Case

Problem | Create an LLM evaluation framework to quantify the following from the context of the financial industry

Use Case: Bloomberg and Deutsche Bank both utilize the LLM evaluation framework to **refine** and train their proprietary Large Language Models. This adoption ensures the accuracy and reliability of their AI-driven solutions in the **financial sector**

Bloomberg

Evaluate output performance on financial and general-purpose tasks using diverse metrics

- Employed **Financial Tasks metrics** like FPB, FiQA SA, Headline, NER, and ConvFinQA, reporting F1 scores and exact match accuracy
- Implemented internal **Sentiment Analysis** and reported F1 weighted by label support
- Explored NER performance on internal datasets, focusing on entity-level F1 score evaluation

Deutsche Bank

Evaluate the effectiveness and potential of Large Language Models in the financial domain

- Prioritized validation accuracy and real-world use cases like document summarization
- Valued interpretability, explain ability, and comparison against traditional models
- Ensured robustness against adversarial inputs and **dynamic financial scenarios**

BloombergGPT | Large Language Models in Finance | **Outcome**

https://www.arxiv-vanity.com/papers/2303.17564/#S5
https://www.nvidia.com/en-us/on-demand/session/gtcspring23-s51160/

Popular Method Overview - Traditional Metrics

Perplexity Score

- Evaluates how well the model **predicts** a sequence of words
- Lower perplexity indicates better language generation

BLEU Score

- Measures the **similarity** between generated text and reference text based on n-grams
- higher score is better

ROUGE Score

Evaluates text **summarization** by comparing overlapping n-grams between generated and reference text

METEOR Score

A machine **translation evaluation metric** that assesses the quality of machine-generated translations by measuring the similarity and fluency of the generated text compared to human reference translations

Short comes

Traditional metrics provide some **measure of similarity** between generated and reference texts, they often fail to capture the deeper, semantic quality of generated content. They can sometimes **reward ungrammatical or nonsensical outputs** that happen to share n-grams with reference texts. For tasks like generative AI, more advanced or task-specific evaluation metrics might be required to truly gauge performance.

Popular Method Overview – Nontraditional Metrics

BERT Score	Evaluates the quality of machine-generated text, such as translations or **text summaries**, by measuring the similarity between the generated text and reference text at both the token and sentence levels
Mover Score	Measures the **dissimilarity** between the generated text and reference text by computing the minimal cost of transforming one into the other while considering the **Earth Mover's Distance** (EMD) metric
Entailment Score	Quantifies the degree of entailment based on the **alignment** and agreement between the words, phrases, and structures in the hypothesis and premise
BLEURT Score	Relies on a pre-trained model to compute a score that measures the **semantic similarity** between the generated text and reference text
Question Answering Score	Quantifies how effectively a system can **answer questions** posed in natural language by comparing its responses to reference answers

https://arxiv.org/abs/1904.09675
https://aclanthology.org/D19-1053/
https://www.sciencedirect.com/science/article/pii/S2949719123000250
https://aclanthology.org/2020.acl-main.704.pdf
https://arxiv.org/abs/2104.08202

Popular Method Overview – LLM-assisted Metrics

Project Description	The **complexity** of natural language makes it hard for metrics with exact formula to measure the quality of the outputs of LLMs Existing LLMs have shown their potential ability to obtain the **deeper statistical information** from the natural language It is reasonable to apply LLMs to evaluate the performance of another LLM

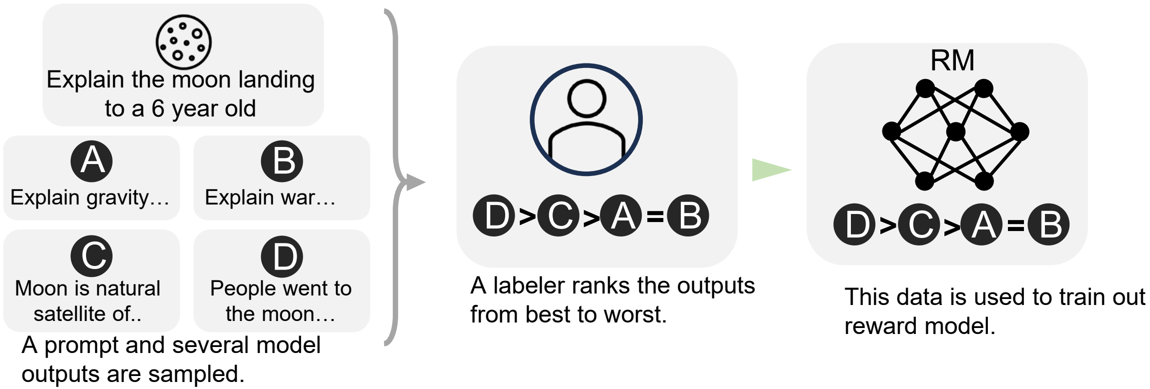
Example

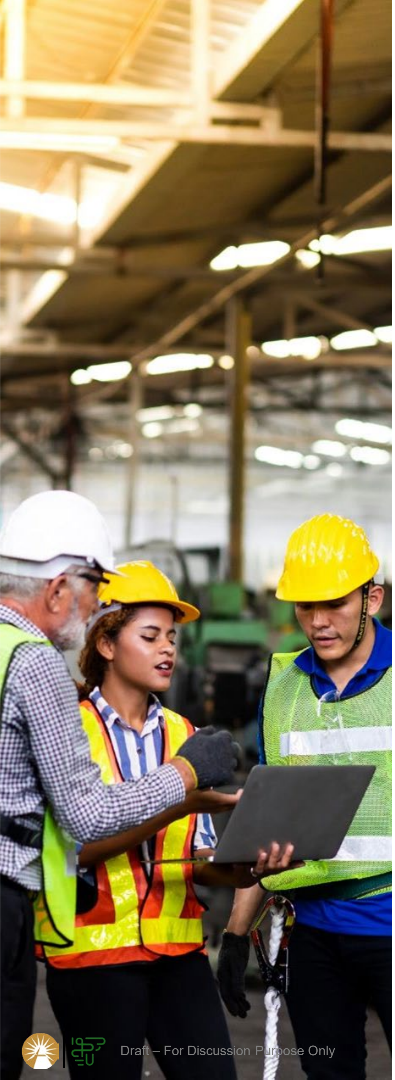
- Wider and Deeper LLM Networks are Fairer LLM Evaluators
- ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate
- PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization
- G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment
- GPTScore: Evaluate as You Desire
- SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models
- Is ChatGPT a Good NLG Evaluator? A Preliminary Study

Popular Method Overview – LLM-assisted Metrics

InstructGPT

- Use a 6B LLM as the **reward model**
- A prompt and **several model** outputs are sampled
- A labeler ranks the outputs from the best to worst
- Use the feedback from labelers to train the reward model.





General Questions

- Is there a specific conference or journal targeted for the paper's publication?
- Given that the datasets in the project description are unlabeled text data, do we need to create our own Q&A pairs for metric evaluation?
- Is an explainable evaluation required?
- Clarification for the input, summary of the document only, or more?
- Is it open to use existing LLMs?
- Should we deploy the model?
- Will computing resources be provided, as training a model for metrics may require significant computational power?

LLM Related Questions

- Could you provide examples and approaches of LLM applications in the financial industry?
- What is Post LLM in the project description?

Project Description:

The project involves the creation of an evaluation framework to quantitatively measure the quality of abstractive summarization models. The team will need to:

1. Generate summaries using different LLMs or prompts.
2. Use several evaluation metrics (e.g., GLUE, ROUGE, METEOR, etc.) to pick the best model or prompt based on the above definition of a "good" summary.
3. Recommend metrics, methods, or post LLM models to use as part of an evaluation framework.

Project Related

- The project specifies that Reasoning prompts should relate to financial and regulatory jargon. Does this also apply to the Correctness and Sensitivity evaluations?
- Besides finance and regulatory jargon, should prompts strictly target finance-related info or include other aspects too?
- Any suggestion for us to make it a successful project?
- What is the expected meeting frequency for us?

Thank you

Contact Us:

Taichen: tz2555@columbia.edu