Extract, Match, and Score: An Evaluation Paradigm for Long Question-context-answer Triplets in Financial Analysis

Bo Hu Han Yuan Vlad Pandelea Wuqiong Luo Yingzhu Zhao Zheng Ma
Singapore Decision Science Center of Excellence, Global Decision Science, American Express
{Bo.Hu, Han.Yuan1, Vlad.A.Pandelea, Wuqiong.Luo, Yingzhu.Zhao, Zheng.Ma2}@aexp.com

Introduction

- The rapid advancement of large language models (LLMs) has sparked widespread adoption across diverse applications, making robust evaluation frameworks crucial for assessing their performance.
- While conventional evaluation metrics remain applicable for shorter texts, their efficacy diminishes when evaluating the quality of long-form answers.
- This limitation is particularly critical in financial analysis involving extended questions, extensive context, and long-form answers.

Contributions

- Benchmark the state-of-the-art LLMs on long-form analysis of earning call transcripts from 10 largest constituents in S&P 500 index.
- Identify the inefficiency of conventional evaluation methods, such as ROUGE, in differentiating generation quality in scenarios of long triplets.
- Propose a generalizable evaluation paradigm named EMS (Extract, Match, and Score) to provide fine-grained evaluations and demonstrated its superiority over RAGChecker in long triplets.

Method

- Extracts salient points from both the reference and candidate answers

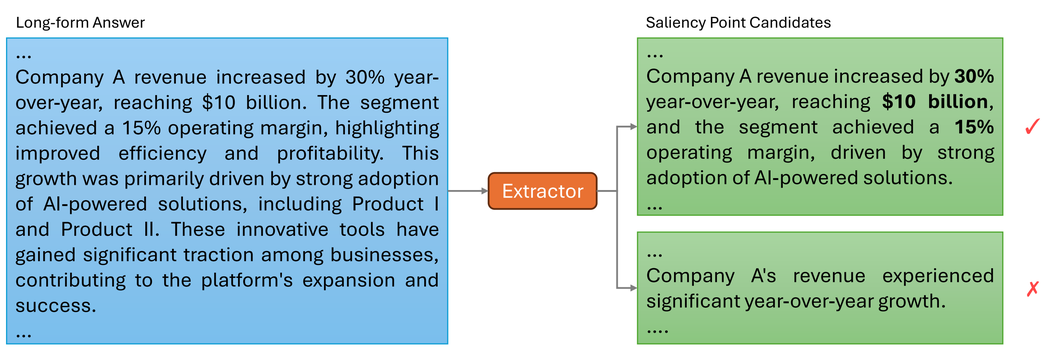

Figure 1: An example of saliency point extraction from long-form answer

- Identifies potential alignments between these points
- Scores each matched pair based on predefined criteria

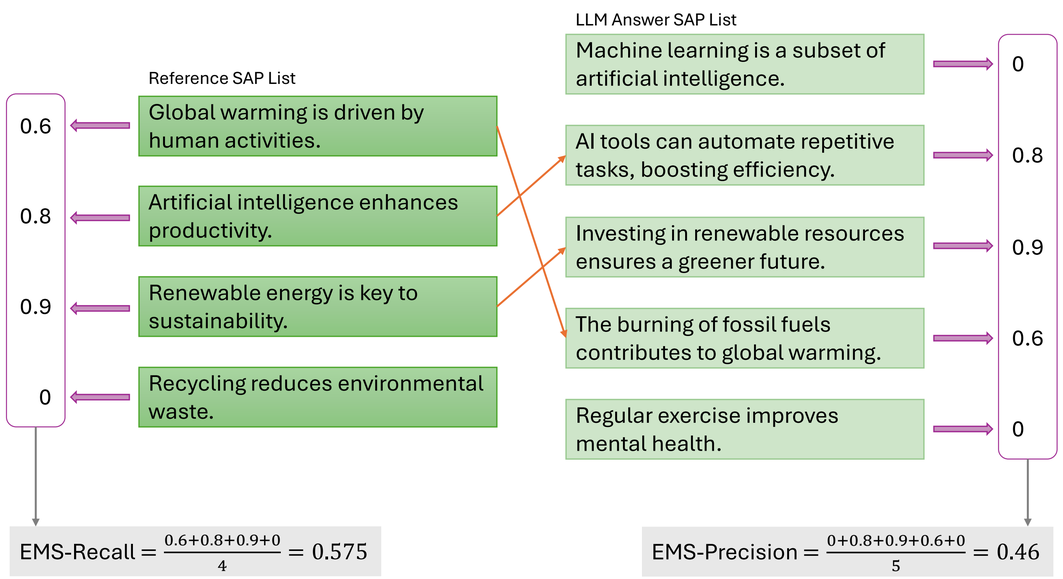

Figure 2: Illustration of matching and scoring procedure in EMS evaluation pipeline.

Dataset

- Use the earnings call transcripts from the third quarter of 2024 for the 10 largest S&P 500 constituents to construct a financial Document-Based QA dataset specifically focusing on long-context and long-form answers.
- Analyze the Q&A sections of several earnings call transcripts to identify common themes and question formats used by analysts. We then use these insights to design five questions that cover commonly referred topics such as revenue growth, operational challenges, strategic initiatives, financial forecasts, and competitive positioning.
- Prompt GPT-4o and Mistral Large to answer each question individually and then generate a consolidated answer by GPT-4o.

Results

- Both RAGChecker and our proposed EMS metrics show more nuanced assessments of the response quality.
- The upward trend of the evaluation figures is identified as the LLM model size increases from 1 bilion to 90 billion.
- Conventional evaluation metrics such as BLEU, ROUGE and BERTScore fail to capture this trend with different model sizes.
- Based on the well-established understanding that larger models offer superior performance, our proposed EMS evaluation is more effective in long-form financial analyses.

Metric/LLM		Llama-3.2-1B	Llama-3.2-11B	Llama-3.2-90B
BLEU		0.02	0.04	0.03
ROUGE	Precision	0.23	0.21	0.22
	Recall	0.14	0.21	0.19
	F1	0.17	0.20	0.20
BERTScore		0.84	0.85	0.85
RAGChecker	Precision	0.45	0.63	0.63
	Recall	0.15	0.31	0.35
	F1	0.21	0.40	0.44
EMS$^{(ROUGE)}$	Precision	0.08	0.17	0.17
	Recall	0.07	0.12	0.13
	F1	0.07	0.13	0.15
EMS$^{(BERTScore)}$	Precision	0.38	0.63	0.65
	Recall	0.34	0.42	0.47
	F1	0.35	0.49	0.54
EMS$^{(LLM)}$	Precision	0.23	0.45	0.47
	Recall	0.21	0.30	0.36
	F1	0.21	0.35	0.40

Table 1: Evaluation results of Ten largest S&P 500 constituents over all questions

Disclaimer

This paper is intended solely for informational purposes and the technique is not a business practice of American Express. Earnings call transcripts are used for demonstration of our evaluation framework and analysis on them is not intended to constitute investment research, advice, or recommendations.