Benchmark LLM over Financial Document

Yichen Huang[1] Taichen Zhou[1] Cong Chen[1] Ruolan Lin[1] Longxiang Zhang[1]
Adam Kelleher[1] Lilli Ann Rowan[2] Indraneel Biswas[2] Micheal Threlfall[2]

Data Science Institute
Columbia University

[1]Columbia University Data Science Institute [2]Fidelity

Fidelity Investments

Goal

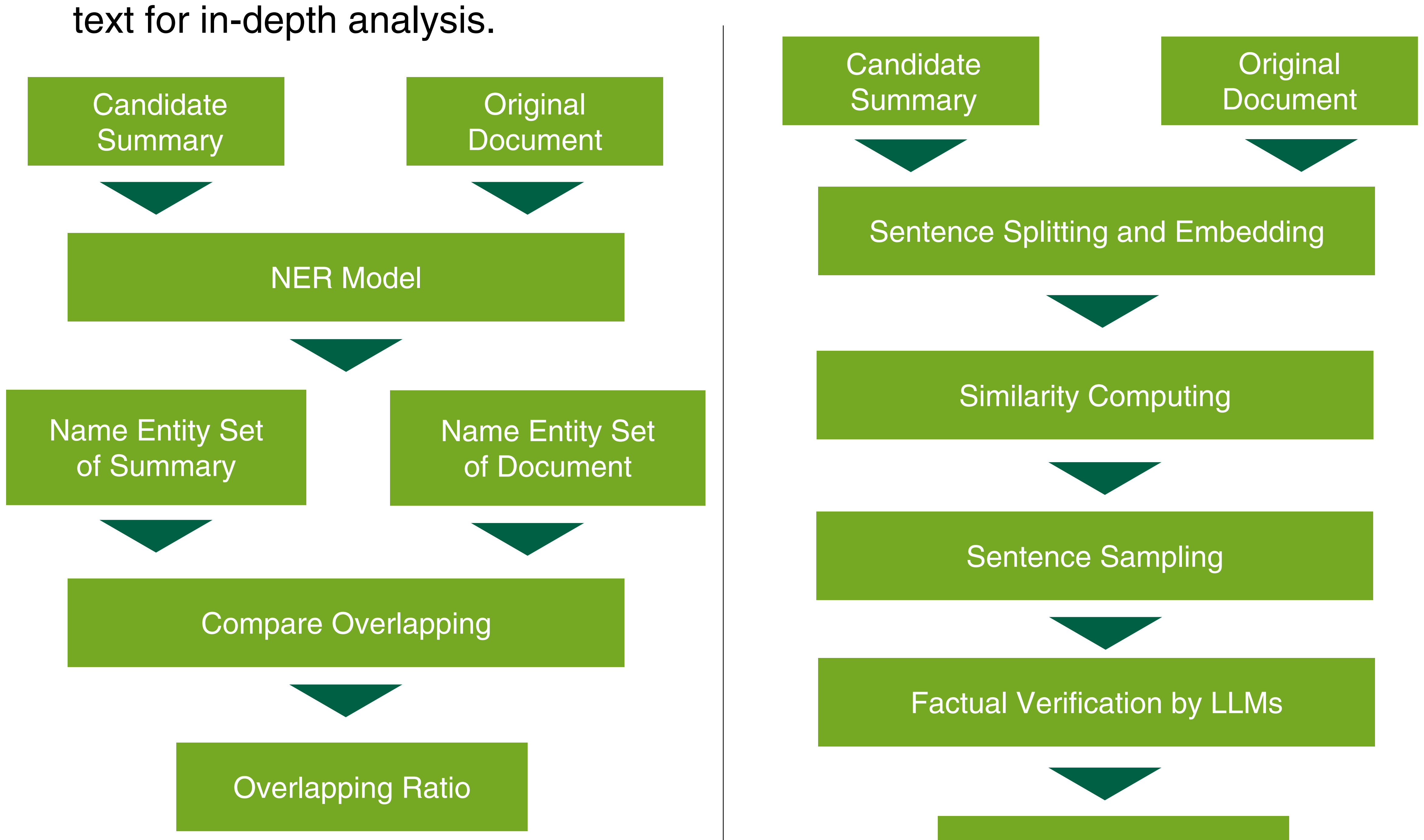
Our project introduces a metric designed to evaluate the quality of textual summaries. This metric is pivotal in fields like finance, where precise information synthesis is critical.

- **Quality Discrimination**: Distinguishes effectively between superior and inferior summaries, ensuring clear differentiation in their factual accuracies.
- **Factual Accuracy Measurement**: Detects and quantifies any factual deviations, assigning lower scores to less accurate summaries.
- **Detail-Oriented Assessment**: Provides comprehensive evaluations, focusing on how well the summary captures the essence and details of the original text.

This metric is not merely a tool for evaluation; it's a step towards enhancing the integrity of information processing in sectors where factual accuracy is non-negotiable.

Frameworks

- **Named Entity Comparison:** Extract and compare financial-related named entities in texts. Analyzes and visualizes named entity accuracy and presence in summaries versus original texts.
- **Sentence-Level-based Summary Checking:** Applies LLMs to check the consistency between the summary and the original text sentence by sentence. Highlights and identifies inconsistencies between the summary and the original text for in-depth analysis.



Candidate Summary
Original Document
NER Model
Name Entity Set of Summary
Name Entity Set of Document
Compare Overlapping
Overlapping Ratio

Candidate Summary
Original Document
Sentence Splitting and Embedding
Similarity Computing
Sentence Sampling
Factual Verification by LLMs
Score

Figure 1. Flowcharts of NER Comparison Model and Sentence-Level-Based Summary Checking Model

Result

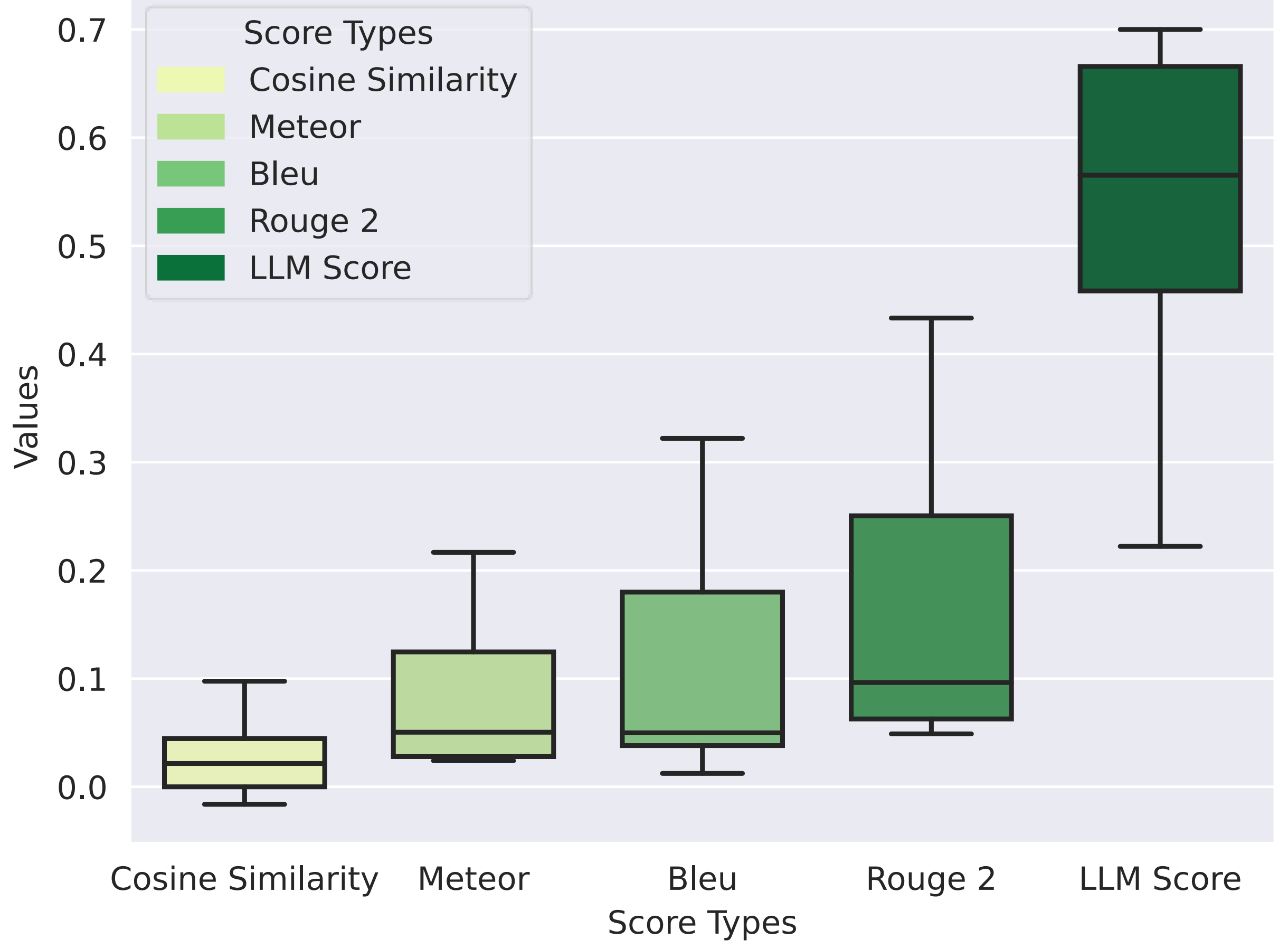


Figure 2. Comparison among different metrics

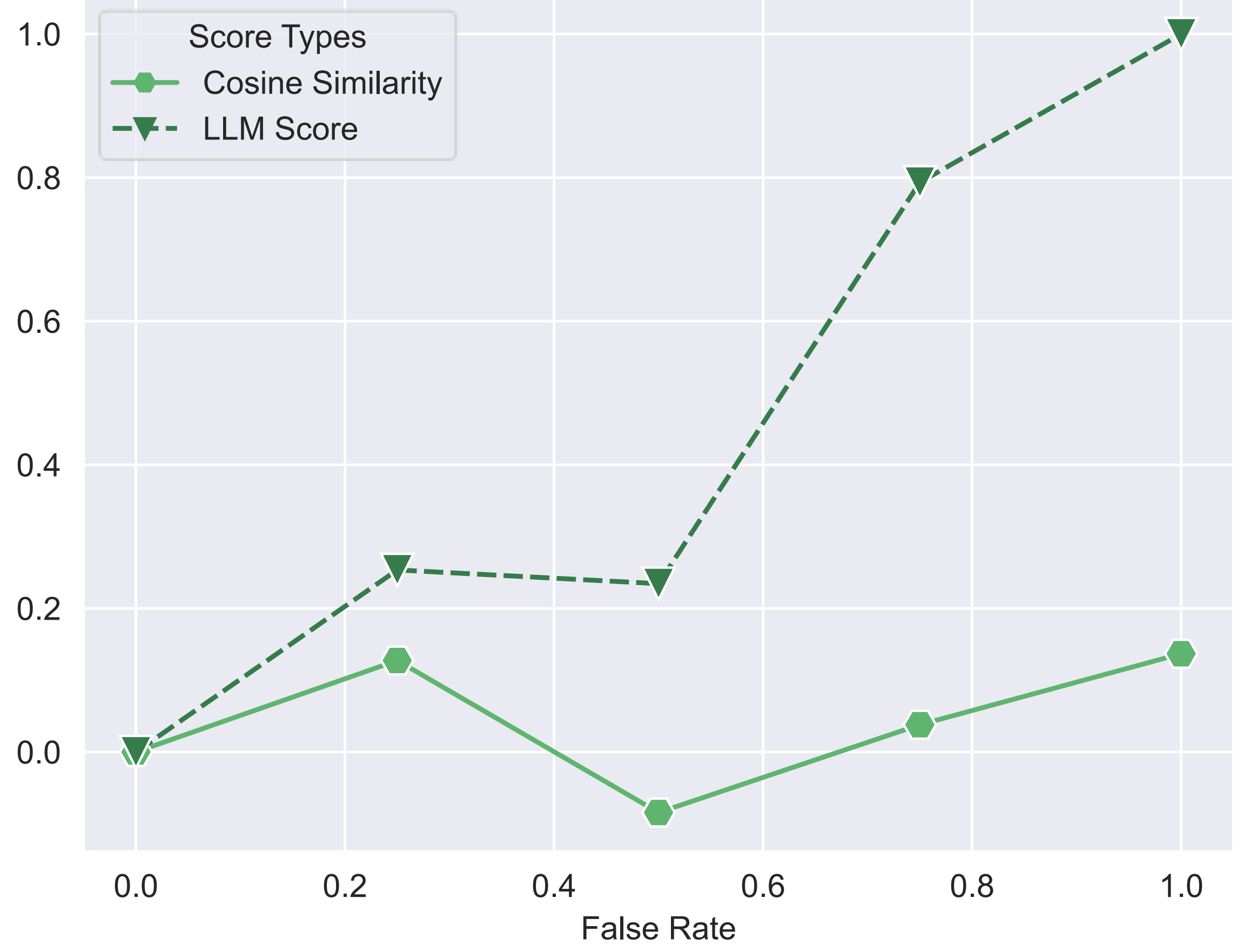


Figure 3. Metric trends with the falsified ratio increase

Metric	Good Summary Score	Falsified Summary Score	Percentage
LLM Score	0.6897	0.3318	**53.00%**
Cosine Similarity	0.8832	0.8382	5.12%
Meteor	0.3063	0.2811	8.24%
Bleu	0.1015	0.0900	10.46%
Rouge 2	0.1512	0.1258	16.03%

Table 1. Comparison between more metrics

Figure 2 and Table 1 reveal that common metrics fail to effectively distinguish between original and falsified summaries. In contrast, our methods demonstrate a clear capacity to discern these differences.

Figure 3 shows that cosine similarity proves inadequate in differentiating between the original and falsified summaries, whereas our methods exhibit a distinct advantage in highlighting these discrepancies.

Conclusion

Our approach consistently surpasses the baseline metrics in the three critical properties we initially identified. This significant outperformance indicates that our methods are markedly more effective than traditional baseline metrics in identifying summary falsification within PDF files.

Currently constrained by budget limitations, we are unable to deploy our model on a more extensive dataset. Moving forward, our team is committed to tirelessly exploring new metrics to evaluate and quantify the quality of summaries, focusing on various dimensions including consistency, variability, and others.

Acknowledgments

We are deeply indebted to Fidelity for sharing the datasets. We are extremely grateful to Columbia DSI for providing us with GCP computational power.