Call Center CoPilot: Enhancing Customer Service with Large Language Models (LLMs)

Alfonso Vieyra, Denise Wei, Raha Ghazi, Tom Gao
Donald Bren School of Information and Computer Sciences, University of California, Irvine

Background

- **CoPilot** utilizes a chosen LLM to perform two specific tasks on dialogue input data:
  - **Sentiment Analysis**
  - **Summary Generation**
- **Purpose**: Automating the process of document sentiment and summarization of client engagement to increase overall productivity at call centers.
- **Approach**: Evaluate and analyze various open-source LLMs based on their performance, flexibility, and cost-effectiveness.
- **Models**: Mistral 7B, Gemma 7B, Llama 3, etc.
  - Approximately seven billion parameters.
- **DialogueSum [1]**: Evaluation dataset.

  Included Columns:
  - **dialogue_id**: uniquely identifies each dialogue.
  - **dialogue_text**: full text of a dialogue.
  - **actual_summary**: contains the observed summary of the dialogue.

  Generated Column:
  - **actual_sentiment**: the benchmark sentiment generated using the FLAN-T5-XXLarge model; reviewed to ensure accuracy.

Results

	Sentiment Analysis		Summarization				
	Accuracy	F1 Score	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
Falcon-7B	0.583	0.554	0.264	0.075	0.237	0.278	0.880
Mistral-7B	0.587	0.603	0.267	0.082	0.245	0.325	0.878
Llama3-8B	0.579	0.556	0.287	0.095	0.264	0.341	0.878
Gemma-7B	0.541	0.541	0.119	0.006	0.111	0.162	0.843

Figure 1: Table of LLM Performances: Falcon Scoring Highest in Summarization

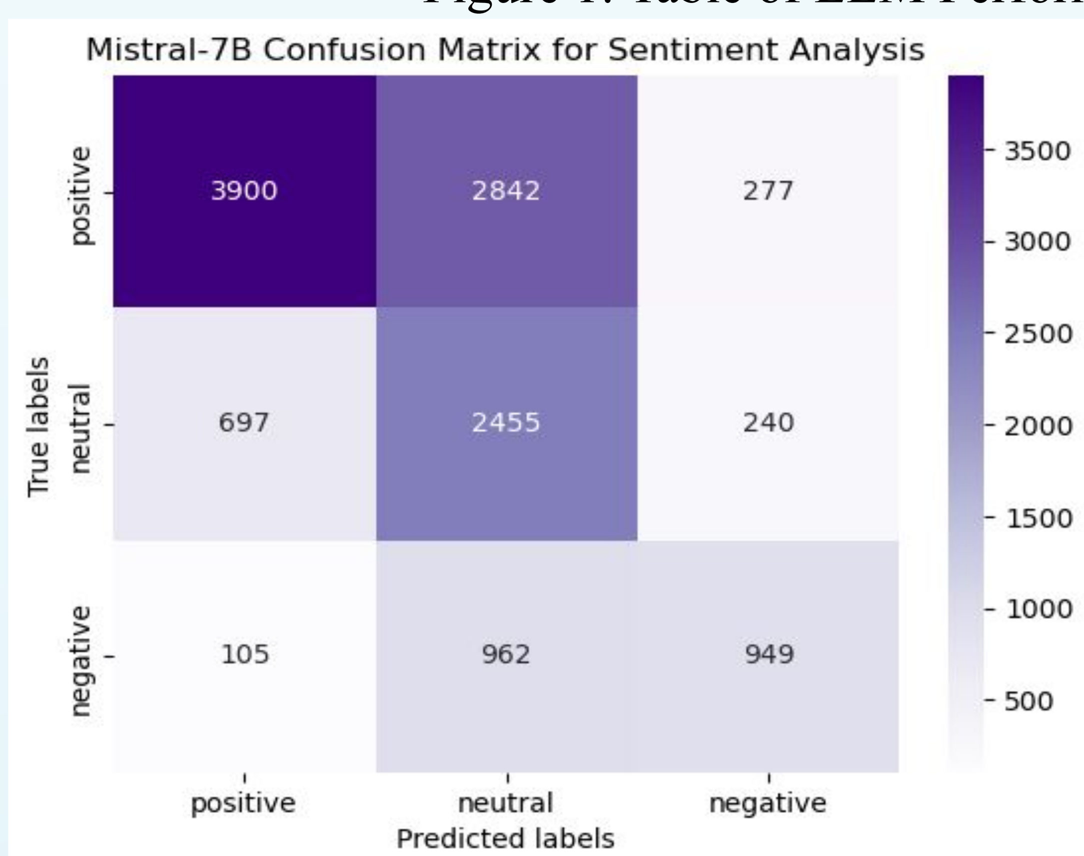

Figure 2: Mistral-7B Confusion Matrix for Sentiment Analysis: Prone to Labeling Positive as Neutral

	Cost-Effectiveness		
	Pre-Task Memory (MB)	Time Taken (seconds)	Memory (MB)
Falcon-7B	26497	0.577	106
Mistral-7B	28649	0.644	237
Llama3-8B	30889	1.001	337
Gemma-7B	32569	0.746	823

Figure 3: Cost-Effectiveness Analysis of Different LLMs: Falcon-7B Demonstrating High Scalability

Evaluation Metrics

Sentiment Analysis

- **Accuracy**: Proportion of correctly classified dialogues out of all dialogues.
- **Precision***: Proportion of true positive predictions out of all positive predictions.
- **Recall***: The proportion of true positive predictions out of all actual positive cases.
- **F1 Score***: Harmonic mean of precision and recall.

*Weighted each class (Positive, Negative, Neutral) by its number of instances

Summarization

Syntax focused:
- **ROUGE**: Recall-Oriented Understudy for Gisting Evaluation (ROUGE).
- **ROUGE-N**: Splits text up into n-grams.
- **ROUGE-L**: Longest Common Subsequence (LCS).

Semantic focused:
- **METEOR [2]**: Metric for Evaluation of Translation with Explicit Ordering.
- **BERTScore [3]**: A metric based on Bidirectional Encoder Representations from Transformers (BERT).

Model Cost-Effectiveness

- **Pre-Task Memory (MB)**: Model memory usage prior to performing sentiment analysis or summarization, which includes memory usage of model weights and preprocessed dialogue inputs.
- **Time Taken (s)**: Time taken to perform sentiment analysis or summarization on a single dialogue input.
- **Memory (MB)**: Additional memory usage to tokenize and perform sentiment analysis or summarization on a single dialogue input.

Conclusions & Next Steps

- **Best Overall Model**: Mistral-7B.
- **Most Cost-Effective Model**: Falcon-7B.
- **Best Sentiment Analysis Model**: Mistral-7B.
- **Best Summarization Model**: Falcon-7B.
- **Mistral-7B** will be selected as the base model used for the CoPilot application and tested in real-world settings.
- Fine-tuning Mistral-7B:
  - **Low-Rank Adaptation (LoRA) [4]**: aiming to refine the model efficiency and applicative precision specifically for call center dialogues.

References

1. Chen, Y., Liu, Y., Chen, L., & Zhang, Y. (2021). Dialogsum: A real-life scenario dialogue summarization dataset. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. https://doi.org/10.18653/v1/2021.findings-acl.449.
2. Lavie, A., & Denkowski, M. J. (2009). The Meteor metric for automatic evaluation of machine translation. Machine Translation, 23(2-3), 105–115. https://doi.org/10.1007/s10590-009-9059-4.
3. Zhang, T., Kishore, V., Wu, F. F., Weinberger, K. Q., & Yoav Artzi. (2019). BERTScore: Evaluating Text Generation with BERT. https://doi.org/10.48550/arxiv.1904.09675.
4. Hu, E. J., Shen, Y., Wallis, P., Zeyuan Allen-Zhu, Li, Y., Wang, S., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2106.09685.

Acknowledgements

We would like to thank:
- Professor Babak Shahbaba and Professor Sharad Mehrotra for their guidance and mentorship.
- Accenture team for the initial project idea and continuous support during the development of this project: Shawna Tuli, Vish Chokshi, Mo Nomeli, Jeff Lu, Ella Liang, Moury Bidgoli, Manish Dasaur, and Marty Hodgett.