accenture

Copilot: Call Center Assistant

Alfonso Vieyra, Denise Wei, Raha Ghazi, Tom Gao

Agenda

- UCI Relationship
- AI Copilot Context
- Framework and Approach
- Model Evaluations
- Future Opportunities

UC Irvine Relationship

UC Irvine Data Science Capstone

- Accenture sponsors the undergraduate-level Data Science Capstone every Winter Quarter.
- Accenture produces project ideas and students “build” the project in 10 weeks.
- Students gain hands-on, real-world experience in addition to soft skills.
- AI for Good initiative.

Our Team

Martin Hodgett

Manish Dasaur

Mo Nomeli

Vishrut Chokshi

Shawna Tuli

Jeffrey Lu

Ella Liang

Moury Bidgoli

Context

Challenge

Agent Burnout

Long Customer Service Wait Times

Slow Traditional Call Analysis Methods

Solution

Call Center CoPilot Prototype is an AI-powered app that utilizes a hand-picked LLM to perform two specific tasks:

- **Sentiment Analysis:** identify and track customer sentiment.
- **Summarization:** generate concise summaries of key points.

Value Proposition

Increase CSAT Scores

enhance customer satisfaction and loyalty

Reduce Average Call Handle Time

improve efficiency

Decrease Agent Burnout

automate repetitive tasks

Project Approach

Prepare and Clean Dataset

- Record dialogue, sentiment, and summarization.

Evaluate open-source LLMS

- Analyze performance, flexibility, and cost-effectiveness.

Create Web App

- Utilize Heroku to deploy and manage web app.

Fine-tuning with LORA

- Fine-tune on training and evaluate on testing.

Academic Journal Publication

- Detail results and findings in a comprehensive paper.

Data Source & Description

Source: HuggingFace.co

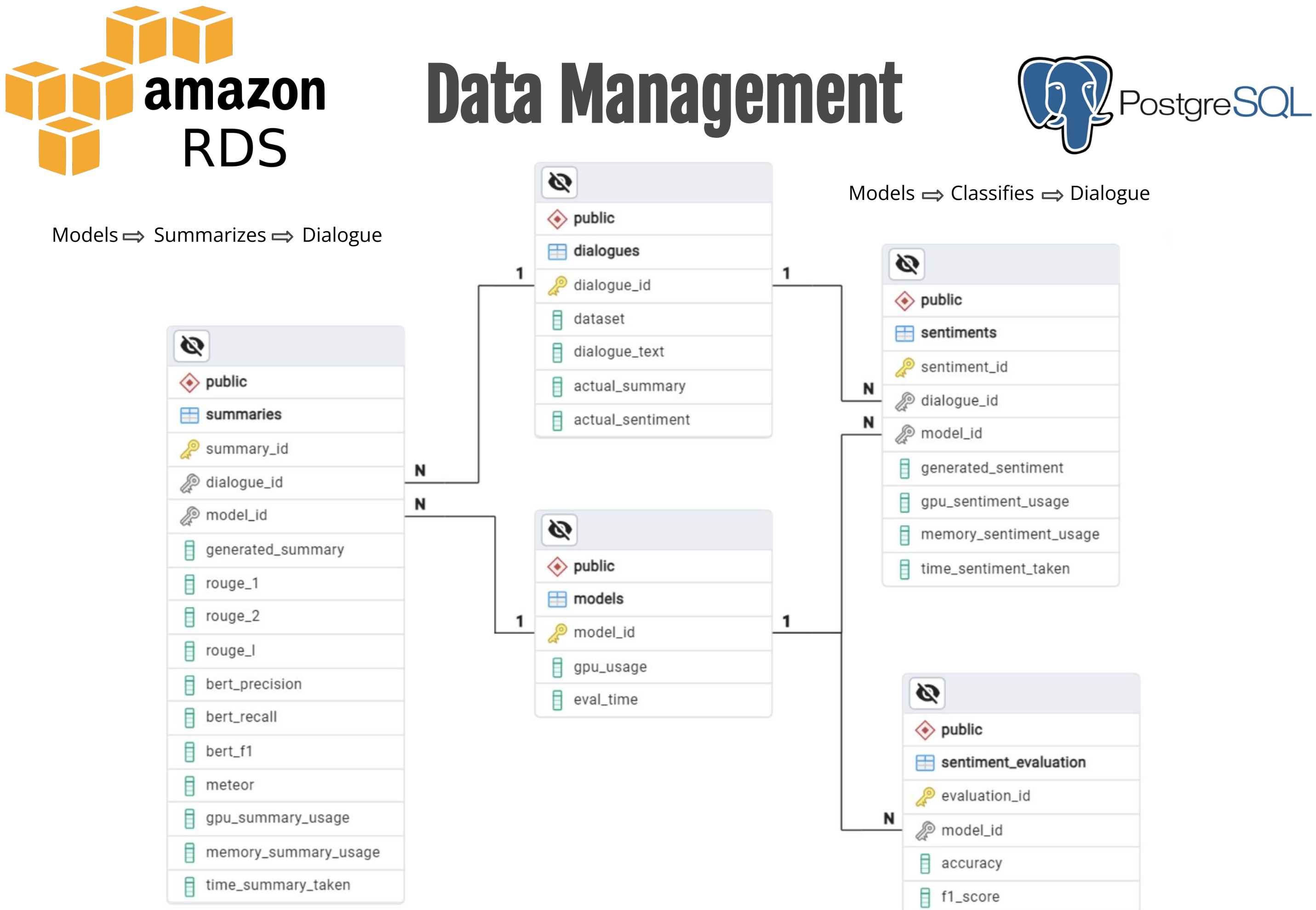
Columns - ID, Dialogue, Summary, and Topic:

- **dialogue_id** corresponds to the *unique* dialogue of each row.
- **dataset** is a categorical variable that identifies which data set the row belongs.
- **dialogue_text** contains the interaction between entities.
- **actual_summary** contains the user summary.

Generated Column - Sentiment:

- Used **FLAN-T5-XXLarge** to create the sentiment column.
- Manually reviewed a random sample of the data to ensure accuracy.

Data Management



Metrics Used

Syntax focused:

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

Semantic focused:

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

Example of ROUGE-1

Original (Reference) Text: “The man is walking down the street in the city”.

Summary (Generated) Text: “A man walks down the road.”

ROUGE-1:

- **Reference 1-grams:** {“The”, “man”, “is”, “walking”. . . “city”}.
- **Generated 1-grams:** {“A”, “man”, “walks”, “down”, “the”, “road”}.
- **Common 1-grams:** {“man”, “down”, “the”}.
- **Precision (P):** 3 common 1-grams / 6 total generated 1-grams = 3/6.
- **Recall (R):** 3 common 1-grams / 10 total generated 1-grams = 3/10.
- **F1 Score (Harmonic Mean):** 2PR/(P + R) = 0.375.

Example of ROUGE-2/L

Original (Reference) Text: “The man is walking down the street in the city”.

Summary (Generated) Text: “A man walks down the road.”

ROUGE-2:

- **Reference 2-grams:** {“The man”, “man is”, “is walking”. . . “the city”}.
- **Common 2-grams:** {“down the”}.
- **F1 Score:** 2PR/(P + R) ≈ 0.143.

ROUGE-L:

- **Longest Common Subsequence:** 3 - “man”, “down”, “the”.
- **F1 Score:** 2PR/(P + R) = 0.375.

Example of METEOR

Original (Reference) Text: “The man is walking down the street in the city”.

Summary (Generated) Text: “A man walks down the road.”

METEOR:

- Align words based on matches, synonyms, and stemming.
  - Exact matches: “man”, “down”, “the”.
  - Synonym matches: “street” - “road”.
  - Stem matches: “walking” - “walks”.
- **Precision (P):** 5 words aligned / 6 words in generated.
- **Recall (R):** 5 words aligned / 10 words in reference.
- **F1 Score:** $2PR/(P + R) \approx 0.519$.

Example of BERTScore

Original (Reference) Text: “The man is walking down the street in the city”.

Summary (Generated) Text: “A man walks down the road.”

BERTScore:

- Generate contextual embeddings of each word using the BERT model.
- Calculate cosine similarity:
  - “A”:“The” = 0.9, “man”:“man” = 1.0, “walks”:“walking” = 0.95, etc.
- Precision, Recall, and F1 Score are all calculated the same as before.

Justification for Metrics

- **ROUGE:** A standard and intuitive metric in NLP, it easy to compute and has various variants which allows for flexibility.

- **METEOR:** Is able to account for some nuance, semantic similarity, and word order while also looking at exact matches like ROUGE.

- **BERTScore:** Uses contextual embeddings from the BERT model which allows it to detect more nuance and is more sensitive towards context within the text.

Metric Assumptions

- **ROUGE** score assumes we do not care about semantic similarity.
- **METEOR** assumes the accuracy of stemming and the synonyms it detects in the aligning process.
- **BERTScore** assumes we care more about semantic similarity than word overlap and lexical similarity.

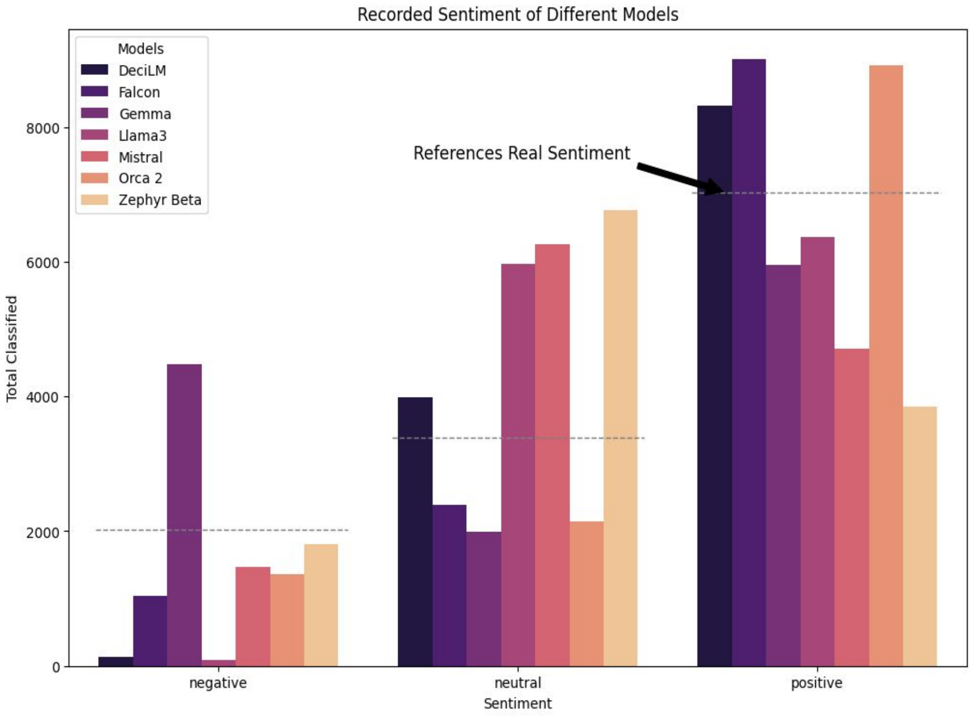
Sentiment Analysis

Sentiment Analysis

Definition: *The process of identifying and classifying text by some measurable means.*

Our Classification Labels:	**Benefits:**
😖 Negative 😐 Neutral 😊 Positive	- Determine overall state of a customer - client interaction. - Contrast negative versus positive interactions to achieve optimal format. **Pitfalls:** - Incorrect classification.

Label Distribution



Key Takeaways:

- Three models under/over classify negative labels.
- DeciLM has most similar distribution of neutral classification.
- Zephyr Beta has least similar distribution of sentiment.

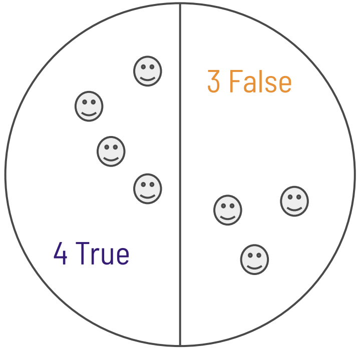
Metrics Used For Sentiment

Accuracy Score

Definition: This score measures the number of times a model has correctly labelled a dialogue overall.

Example:

Accuracy = 4 / 7 = 57.1%



F1 Score

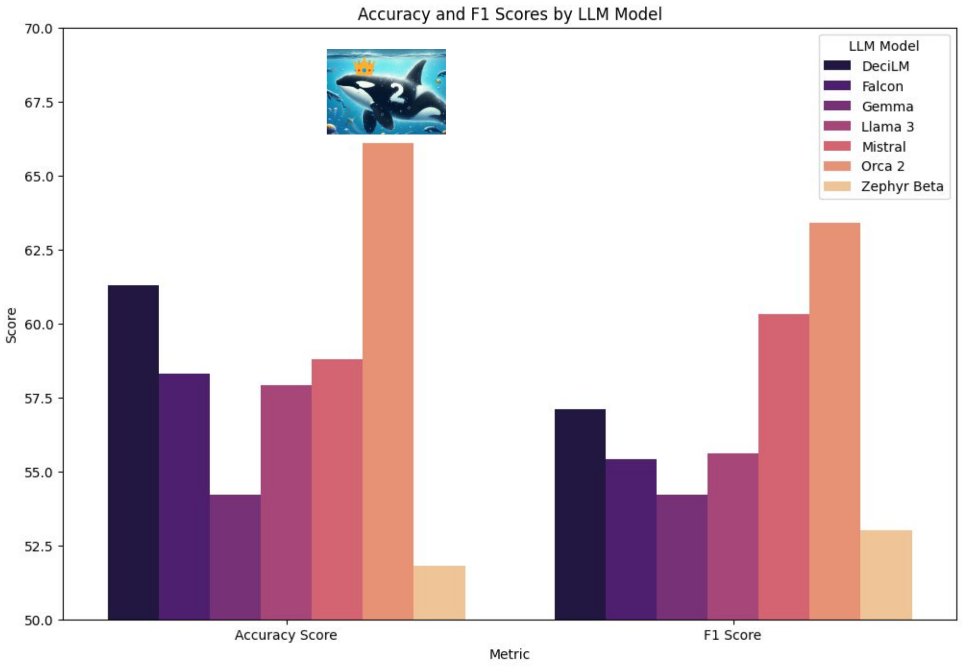
Definition: This score measures the harmonic mean of *precision and recall*.

What is precision and recall?

Precision: ratio of correctly predicted positive observations to the total predicted positives.

Recall: measures the quantity of the actual positives captured by the predictions (how many actual positives are predicted as positive).

Evaluation Metrics



Best Performer: Orca 2

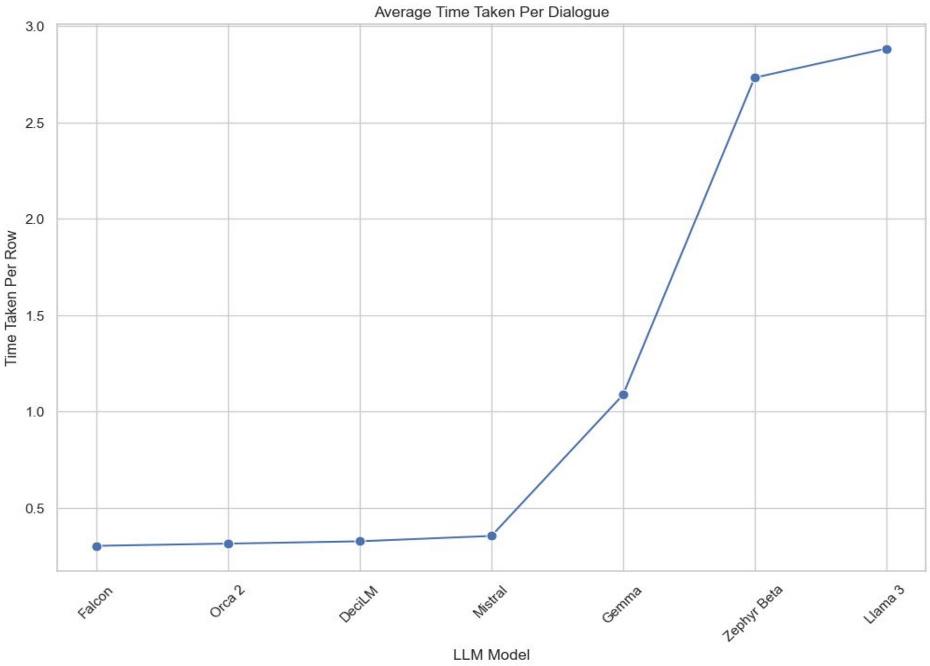
66.1%

Highest Accuracy Score

63.4%

Highest F1 Score

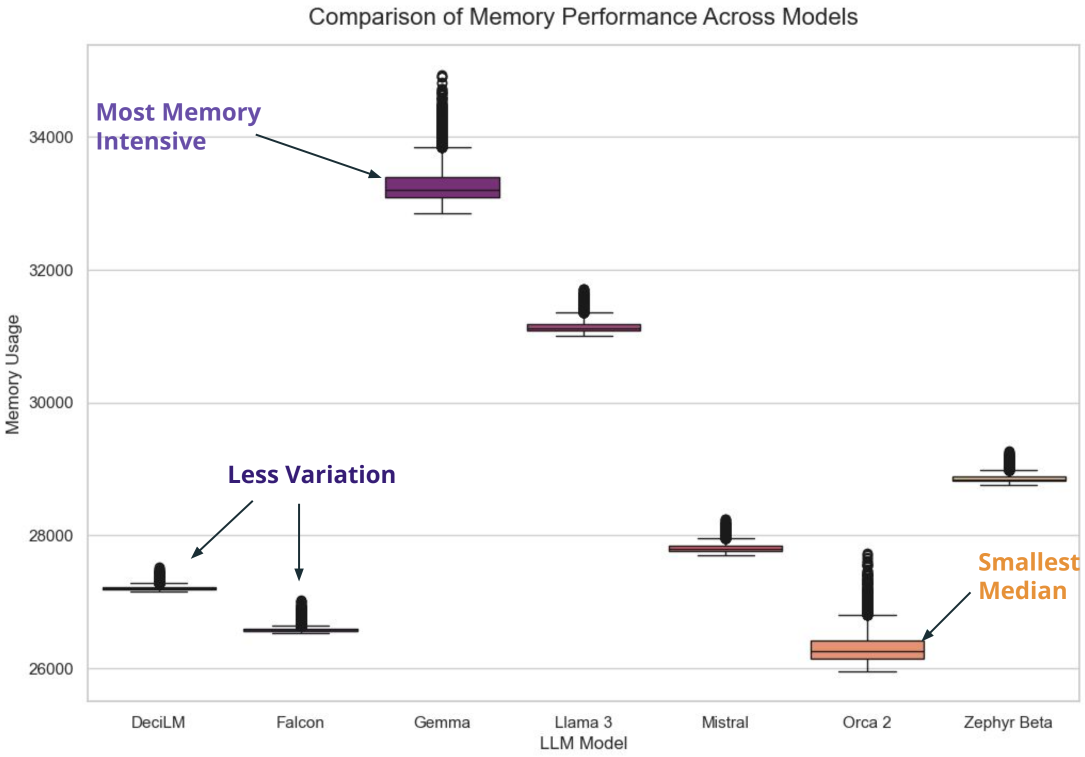
Time Analysis



Fastest: Falcon at 0.3 sec

Slowest: Llama 3 at 2.9 sec

Memory Usage



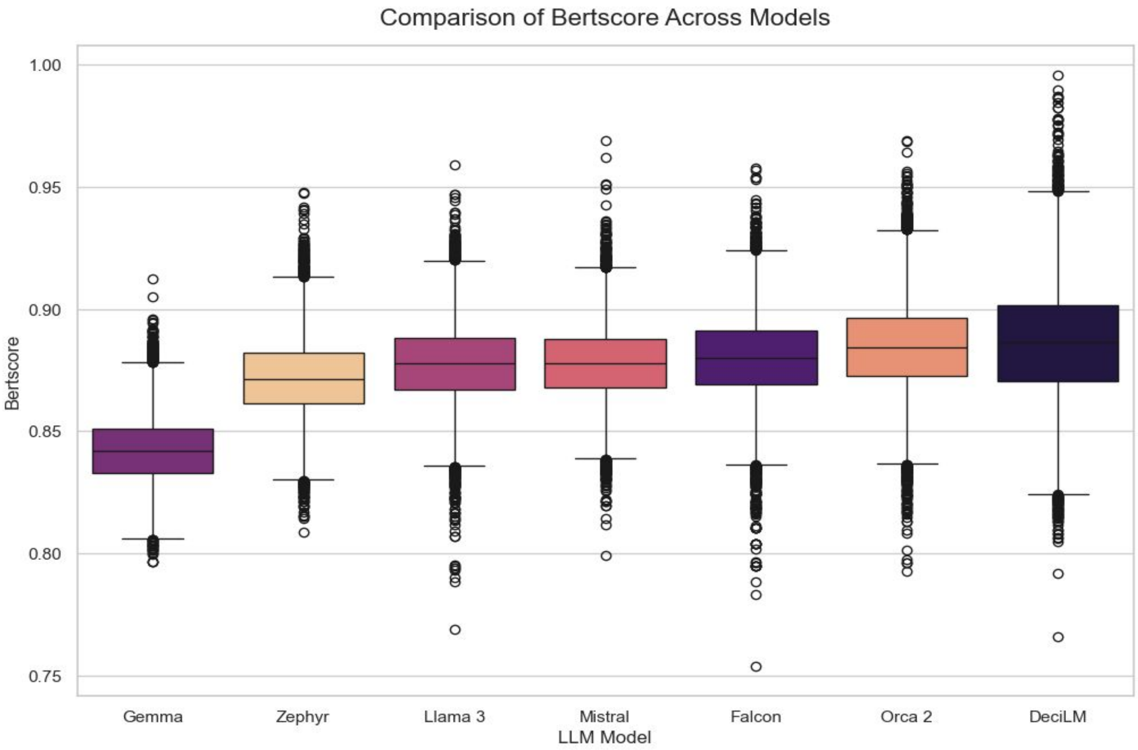
Best Sentiment Performer:

Orca 2

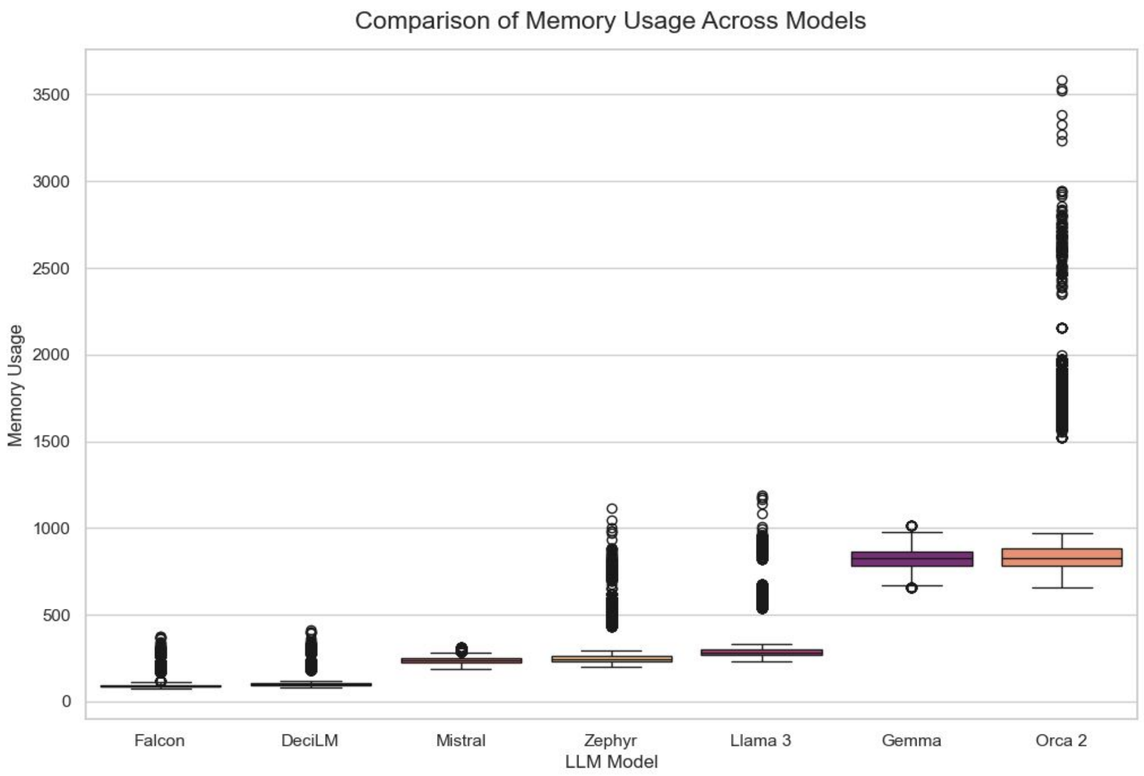
Highest Accuracy Score
Highest F1 Score
Least Memory Usage

Summarization

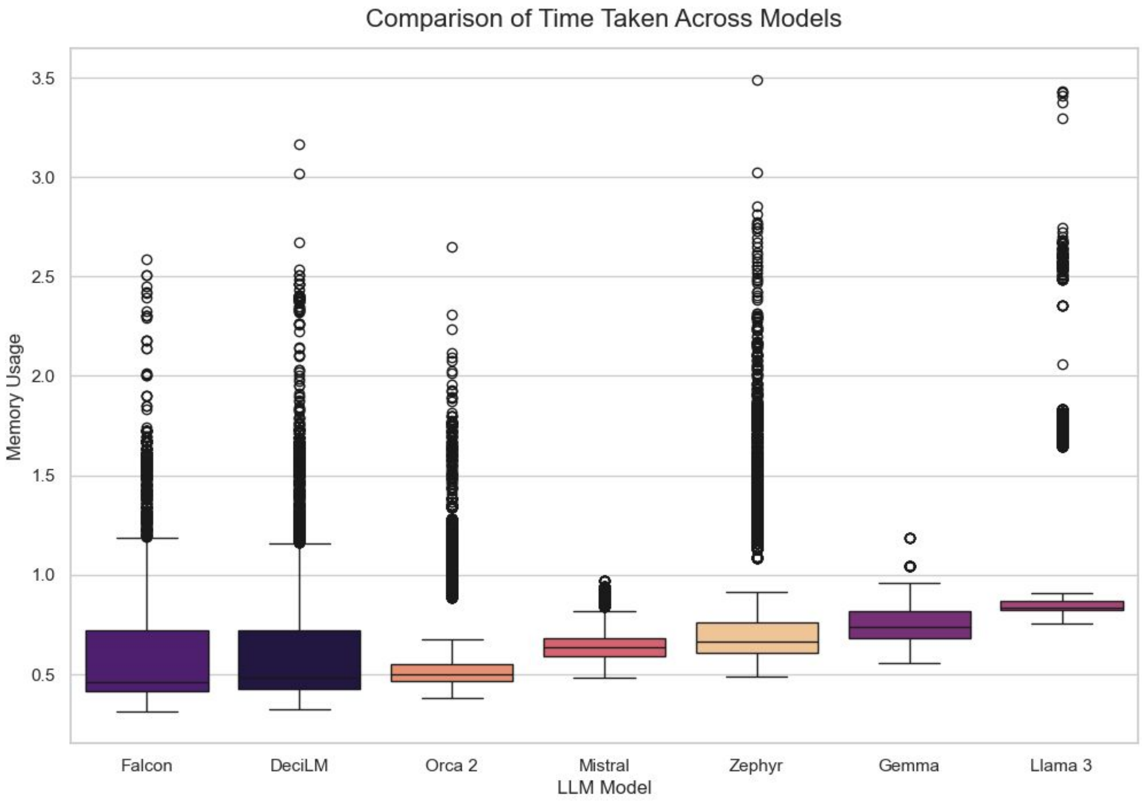
Comparison of Bertscore Across Models



Comparison of Memory Usage Across Models



Comparison of Time Taken Across Models



Final Touches

Fine-Tuning

- Fine tune final model on training split of DialogSum and reevaluate on testing split.

Web App Design

- Improve the design of the web app and implement additional functionalities.

Final Evaluation

- Perform a final evaluation on our selected model using actual call center transcripts/logs.

Potential Future Changes

Pre-Call Features

- Use transcripts/summaries from previous calls with a customer to briefly inform an agent before starting the call with a customer.
    - Utilizing LLMs Information Extraction, the Call Center CoPilot could highlight key points and takeaways from the customer's previous calls.

Performance Metrics

- Show agents metrics based on length of calls, time between calls, etc.
- Can be used by agents or supervisors for performance evaluation.

Categorization of Calls

- Categorize calls based on topics in the corresponding industry for ease of access in the future.

DEMO

- https://copilot-ecdfb807803e.herokuapp.com/

ANY QUESTIONS?