Datathon25 - Sentient Challenge

Overview

We worked under the following restrictions:
- Use our ODS framework,
- Use any model on [Fireworks](#) ≤ 90B parameters (total parameters),
- Time constraints of 24 hours.

Our approach focused on improving the given framework in order to achieve the best results with the given constraints. We experimented with different model architectures, prompting strategies, data augmentation using Wikipedia and evaluating the final answers of different models using majority vote.
To evaluate the available models for accuracy we tested for the Frames Benchmark.

Model comparison

- Qwen (QwQ-32B)

QwQ-32B served as our initial benchmark model, which achieved 59.7% Accuracy, providing a baseline for comparison with later attempts.

- Llama 3.3 (Llama 3.3 70B Instruct)

Llama 3.3 allowed us to improve the accuracy significantly (to 63.96%) and was used for the next phase of improvements.

Prompting strategies

Next, we tested minor but targeted prompt engineering techniques. These changes focused on improving clarity, ensuring consistent structure, and guiding the model's reasoning process. This led to a marginal increase to 64.03%, indicating diminishing returns from prompt-level tweaks alone.

Wikipedia tool

Imported langchain wikipedia tool and converted it to smolagents tool. Tried various parameters but couldn't go above 64% using LLama 3.3

Reranking

Activation of the Jina reranker improved the accuracy to 67.72%.

Majority voting
Our final innovation consisted of summarizing/evaluating the answers of multiple models using majority vote. Due to time constraints we mainly combined models we got from earlier testing. This step could be improved by using an additional agent to find the most accurate answer across different models. Using different variations of models from above we got the final accuracy of 69.42%.