Levon Green II

4/6/2024

CS6625

Midterm

PaLM vs Gemini Model Analysis

1. Accuracy

PaLM: The PaLM model was able to answer 20/26 questions correctly. 2/26 of the answers provided by the model were entirely wrong. 4/26 answers were technically correct, but were made up since they were not included in the prompt-answer dataset.

Gemini: The Gemini model was able to answer 22/26 questions correctly. 4/26 questions were completely wrong. The Gemini model did not make up answers like the PaLM model and instead defaulted to the standard, “I am not sure. Please call: 1-334-808-6576 for department questions or ask your instructor for academic questions” message.

2. Precision

PaLM: 24/26 of the answers provided were precise in that it adequately answered the question provided including the answers that were seemingly made up but still technically correct. Two example questions displaying this is when I asked the model about undergrad and graduate credit requirements as well as graduate class sizes.

Gemini: The Gemini Model provided more precise answers in that it more frequently displays the correct answer from the dataset instead of deviating from the supplied data with 22/26 answers being correct. Of these 22 answers, they were closely attuned to the information in the dataset or extremely close to it.

3. Recall (Sensitivity)

PaLM: The PaLM model was able to recall 20/26 answers correctly as provided in the dataset.

Gemini: The Gemini model was able to recall 22/26 answers correctly as provided in the dataset.

4. F1 Score

PaLM: The F1 Score for the PaLM Model is 21.81/26

Gemini: The F1 Score for the Gemini Model is 22/26

Formula used: F1 = 2 \* (p \* r) /(p+r) where p = precision and r = recall

5. Response Time

PaLM: Data was entered into the model manually. Most answers were provided within 1-2 seconds. This response time would be satisfactory for most end users looking to get answers to department questions.

Gemini: Data was entered into the model manually. Most answers were provided within 1-2 seconds. This response time would be satisfactory for most end users looking to get answers to department questions.

6. User Satisfaction

PaLM: The PaLM model is fairly consistent with the answers provided. It seems to be better at interpreting “language” and finding the correct information even when that information is not present in the dataset.

Gemini: The Gemini model is consistent with the data within the dataset. It does not deviate or vary from the answers contained within the dataset. It does miss a few questions that are closely related to the questions in the dataset by providing the default “I am not sure” answer.

7. Error Rate

PaLM: The error rate for the PaLM model was 2/26. Of the 26 questions that I asked, 24 of the answers were relevant to the question being asked even when the model made up the answer. 2 of the 26 answers provided were wrong.

Gemini: The error rate for the Gemini Model was 4/26. Of the 26 questions that I asked, 22 of the answers were relevant to the question that I asked. 4 of the 26 answers provided were wrong.

8. Contextual Understanding

PaLM: The PaLM model performed better when reviewing contextual understanding of the questions presented. For questions such as “How many credits are required to graduate with BS CS?”, it answered, “A total of 120 semester hours are required for graduation, including 30 hours of computer science courses.” While this data was not present in the dataset, the model still provided a relevant and correct answer.

Gemini: The Gemini model did not perform well in terms of contextual understanding. The model only provided relevant answers if the question closely resembled the questions contained in the dataset.

9. Domain Adaptability

PaLM: Questions asked: “What is the most popular programming language?”, “What is Google?”. “What is the hardiness zone for 35749?”

When asking questions that deviate far from the realm of the data in the dataset, the default, “I am not sure” answer is what populates. This holds true for general computer science questions about programming languages which should be within the domain of the assistant. Also general questions about the internet or other topics, that happen to be outside of the domain of the assistant also provide the standard, “I am not sure” message.

Gemini: Questions asked: “What is the most popular programming language?”, “What is Google?”. “What is the hardiness zone for 35749?”

When asking the same bonus questions as the PaLM model, the Gemini model does not deviate from the standard answers within the dataset. Based on the analysis from previous steps, this is expected.

Summary:

Overall both models handle general computer science questions about Troy fairly well. The Gemini model seems to be more useful in cases where a developer does not need the answer to deviate from the provided answers in the data set at all. The use cases for this model would be vital for a medical AI assistant where people may ask questions regarding doctors, general care questions or even medication questions. In these cases, it is important for the model not to deviate from the standard answers in the dataset. This model would also be useful for legal use cases where a legal AI assistant may provide general legal information that has to be precise.

The PaLM model performed better in terms of provided relevant information when it did not know the correct information. While a developer may not want this behavior when the model is being used for something sensitive such as the healthcare or legal field, it can be beneficial for a more general AI assistant where precision and recall scores do not have to be as high.