An Evaluation of the Performance of Large Language Models on Relevancy Classification Tasks

Avinash Chouhan
University of Delaware - Summer Fellows Research
Sensify Lab - SMIDGen Project

August 8, 2024

Introduction

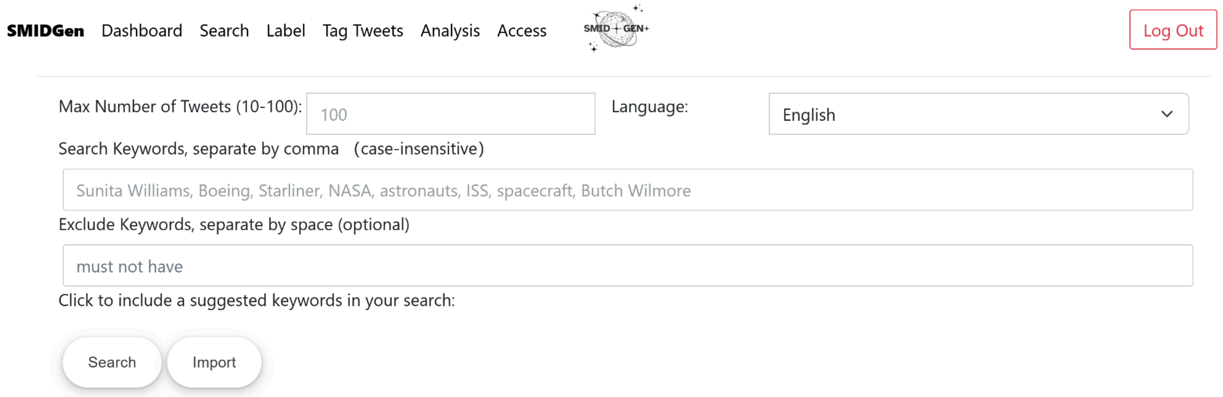
Over the last decade, the growing interest in Artificial Intelligence (AI) has led us to create unique Large Language Models (LLMs) that have impacted our everyday lives in ways we do not even realize. LLMs are algorithmic models trained on immense amounts of data, allowing them to understand text-based prompts and evaluate them with the help of Natural Language Processing (NLP)[1]. Today, LLMs are extensively used in every industry, with tasks ranging from simply summarizing a paragraph to writing complex and beautiful sonnets. There are currently hundreds of LLMs from a variety of different developers, the most prominent ones include OpenAI's ChatGPT, Google's Gemini, Meta's Llama, and Anthropic's Claude[2]. As we continue increasing our utilization of these models and our dependence on them for regular tasks, developers keep training and further developing them to be more accurate and reliable through additional research and larger datasets. While considerable research has already been done on LLMs and their generative AI abilities, this research specifically focuses on evaluating and comparing individual performance on relevancy classification tasks for each of the 4 prominent LLMs.

Methodology

This research aims to be a comparative analysis of the 4 different LLMs, which include ChatGPT -4o mini, Gemini 1.5 Flash, Llama 3, and Claude 3. The test will consist of giving the model a prompt, alongside a dataset to apply the prompt to, and then analyzing the AI-generated data for insights into the model's performance. The dataset consists of 200 tweets, collected using the SMIDGen[3] system while searching for the recent Boeing Starliner failure incident.

The SMIDGen system was used for data collection because of its interface and ability to search for specific keywords across the X[4] platform. The interface allowed advanced search filtering including languages, multiple keywords, and keyword omittance. To collect the dataset a search was conducted using the parameters shown in Figure 1.1.

[1] https://www.ibm.com/topics/large-language-models
[2] https://zapier.com/blog/best-llm/
[3] https://smidgen.cis.udel.edu
[4] Previously Twitter



Max Number of Tweets (10-100): 100

Language: English

Search Keywords, separate by comma (case-insensitive)

Sunita Williams, Boeing, Starliner, NASA, astronauts, ISS, spacecraft, Butch Wilmore

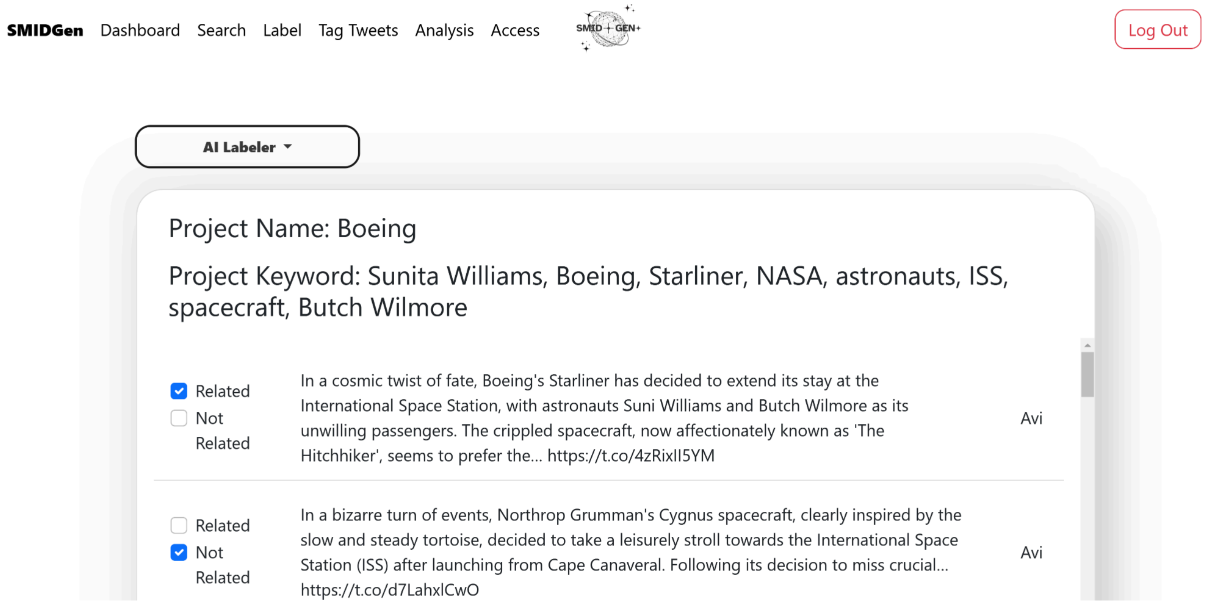
Exclude Keywords, separate by space (optional)

must not have

Click to include a suggested keywords in your search:

Figure 1.1: The SMIDGen search interface with filters for this research's dataset. The most common keywords relating to the topic are included in the search, language is set to English and no keywords have been omitted from the search. Search is performed in 2 batches of 100 due to X's API limits.

After the collection process, the data has to be labeled by a human to create the ground truth values. This was also done through SMIDGen, using its labeling feature shown in Figure 1.2. Once labeling was complete, the data could be downloaded locally with the human-classified labels and the collected dataset.



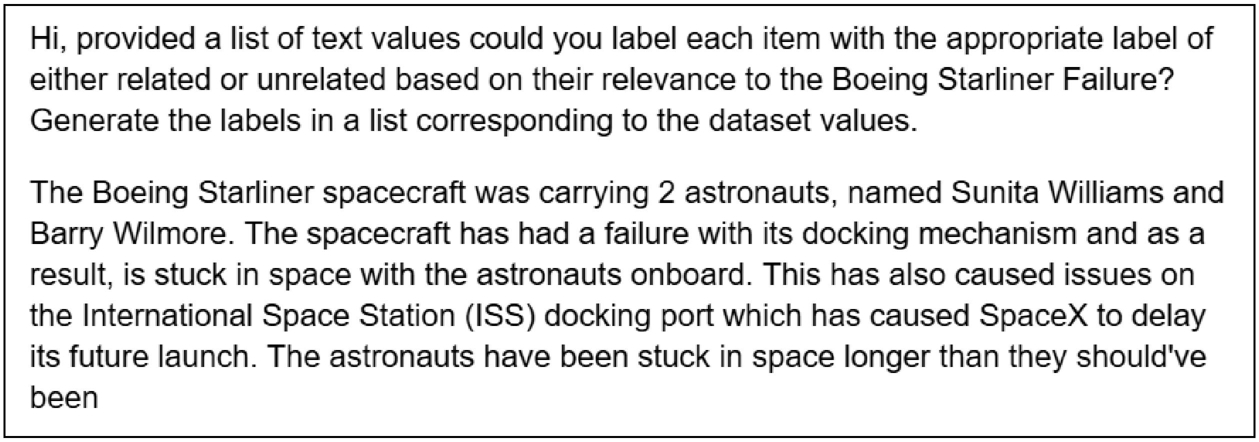
Project Name: Boeing

Project Keyword: Sunita Williams, Boeing, Starliner, NASA, astronauts, ISS, spacecraft, Butch Wilmore

☑ Related ☐ Not Related	In a cosmic twist of fate, Boeing's Starliner has decided to extend its stay at the International Space Station, with astronauts Suni Williams and Butch Wilmore as its unwilling passengers. The crippled spacecraft, now affectionately known as 'The Hitchhiker', seems to prefer the... https://t.co/4zRixlI5YM	Avi
☐ Related ☑ Not Related	In a bizarre turn of events, Northrop Grumman's Cygnus spacecraft, clearly inspired by the slow and steady tortoise, decided to take a leisurely stroll towards the International Space Station (ISS) after launching from Cape Canaveral. Following its decision to miss crucial... https://t.co/d7LahxlCwO	Avi

Figure 1.2: Showcasing the labeling interfacing of SMIDGen, where collected data can be labeled with relevant labels. Additional features included multi-user access and customized labels for a more practical environment.

After the human-labeled dataset is collected, the next step is to generate labels for the data using LLMs. First, a prompt is engineered to precisely asks the model to complete a task. While also providing sufficient information for the model to understand the task. The prompt selected for this research is shown in Figure 1.3.



Hi, provided a list of text values could you label each item with the appropriate label of either related or unrelated based on their relevance to the Boeing Starliner Failure? Generate the labels in a list corresponding to the dataset values.

The Boeing Starliner spacecraft was carrying 2 astronauts, named Sunita Williams and Barry Wilmore. The spacecraft has had a failure with its docking mechanism and as a result, is stuck in space with the astronauts onboard. This has also caused issues on the International Space Station (ISS) docking port which has caused SpaceX to delay its future launch. The astronauts have been stuck in space longer than they should've been

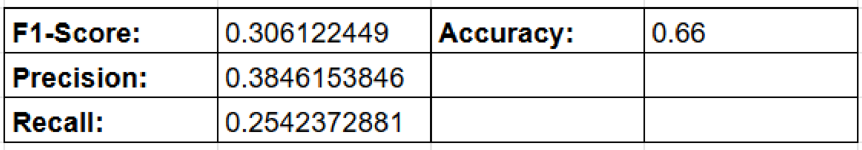
Figure 1.3: The created prompt for the research asking LLMs to classify data based on the topic of the Boeing Starliner and its description.

The initially collected dataset had to be cleaned removing like counts and similar metrics, while only saving the text content of each tweet in a list. After generating the labels with each LLM, the performance evaluation[5] of each model could be started.

Results

In the dataset of 200 total tweets, as classified by a human there were 59 related tweets. The average word count of the dataset was 27 words and the Flesch Reading Ease test[6] had an average rating of 41.5. This means the dataset was fairly complex and needed to be evaluated at an advanced level for better comprehension.

The Google Gemini labeled data yielded 15 correctly predicted positives and 117 correctly predicted negatives. Calculations for performance are shown in Figure 2.1.



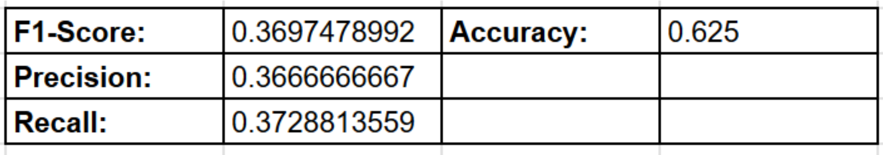
F1-Score:	0.306122449	**Accuracy:**	0.66
Precision:	0.3846153846		
Recall:	0.2542372881		

Figure 2.1: Performance Metrics for the Gemini 1.5 Flash LLM

The OpenAi ChatGPT labeled data yielded 22 correctly predicted positives and 103 correctly predicted negatives. Calculations for performance are shown in Figure 2.2.

[5] All the computational work was done through Python in jupyter Notebooks, with the help of libraries such as Pandas, Numpy, MathPlot, Scikit Learn, and TextStat.

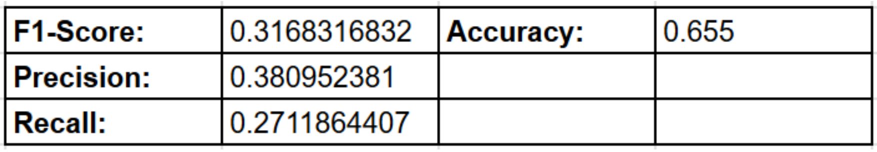
[6] The Flesch Reading Ease score is a measure of text readability. It assigns a number between 0 and 100 to a piece of text, with higher scores indicating easier readability.



F1-Score:	0.3697478992	**Accuracy:**	0.625
Precision:	0.3666666667		
Recall:	0.3728813559		

Figure 2.2: Performance Metrics for the GPT -4o mini LLM

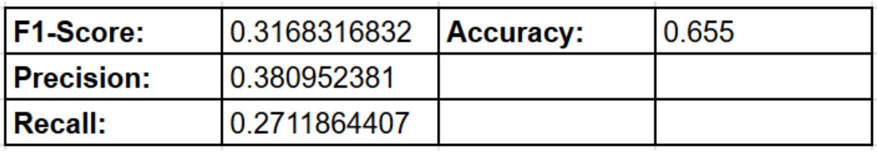
The Meta Llama labeled data yielded 16 correctly predicted positives and 115 correctly predicted negatives. Calculations for performance are shown in Figure 2.3.



F1-Score:	0.3168316832	**Accuracy:**	0.655
Precision:	0.380952381		
Recall:	0.2711864407		

Figure 2.3: Performance Metrics for the Llama 3 LLM

The Anthropic Claude labeled data yielded 16 correctly predicted positives and 116 correctly predicted negatives. Calculations for performance are shown in Figure 2.4.



F1-Score:	0.3168316832	**Accuracy:**	0.655
Precision:	0.380952381		
Recall:	0.2711864407		

Figure 2.4: Performance Metrics for the Cluade 3 LLM

No major differences were analyzed between the 4 LLMs. They all performed similarly, with slight variations in their performance metrics.

Discussion

The results were very interesting and unexpected. Based on the advanced level of the current models and their ability to accurately comprehend language, there was an expectation that the LLMs would accurately label the data, matching the human-classified labels. However, the LLMs performed poorly with the F1 score in the 0.3 range. This showed that with the given prompt, the model could not accurately comprehend the task or the topic and therefore labeled items incorrectly. In Figure 3.2 we can see a graphed comparison of the 4 different scores for each model.

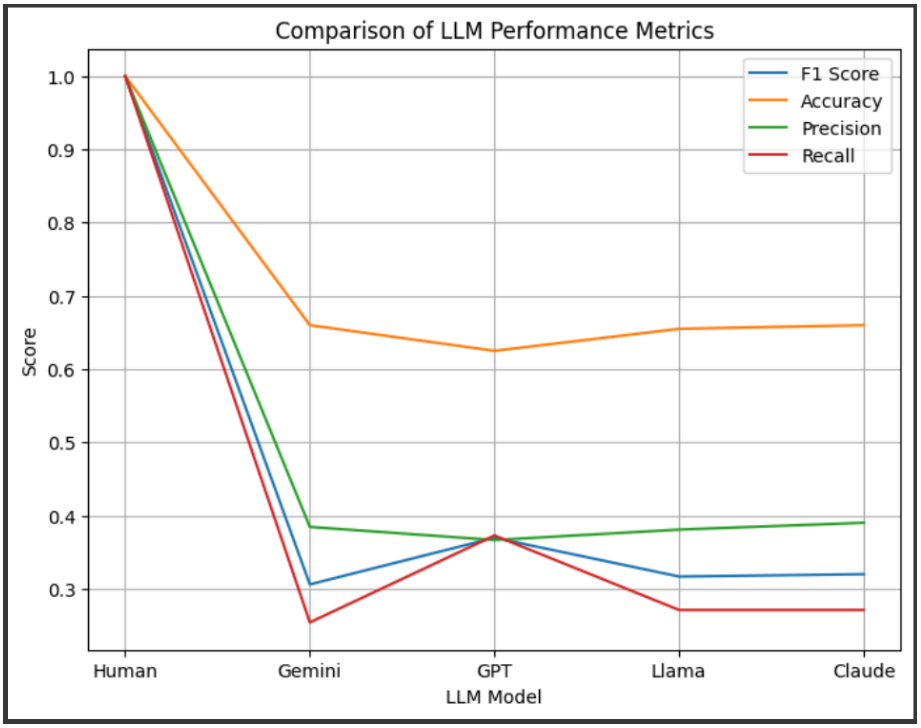


Figure 3.1: Here we can see the graphed individual performance of LLMs. Comparing the points for each of the 4 lines, we can see that GPT performed the best out, however, its precision and accuracy were also slightly lower in comparison.

As we can see LLMs have significantly lower scores than the human-classified labels. There could be various reasons why this was the case, perhaps the prompt was not intricate enough for the model to clearly understand the topic, or it could be an issue with how the dataset was presented to the model. Looking at the prediction accuracy in Figure 3.2 gives us a clearer look at how each model performed. There were slight variations in the accuracy percentages across LLMs however the GPT model was the highest. The prediction accuracies were once again very low, remaining below 40%. While this research could not thoroughly analyze the reason behind the low ratings, future research could learn from this experiment and try to better understand why that was so. Figuring out better prompt generation and dataset formats could help increase the model's understanding of the task and overall classification performance.

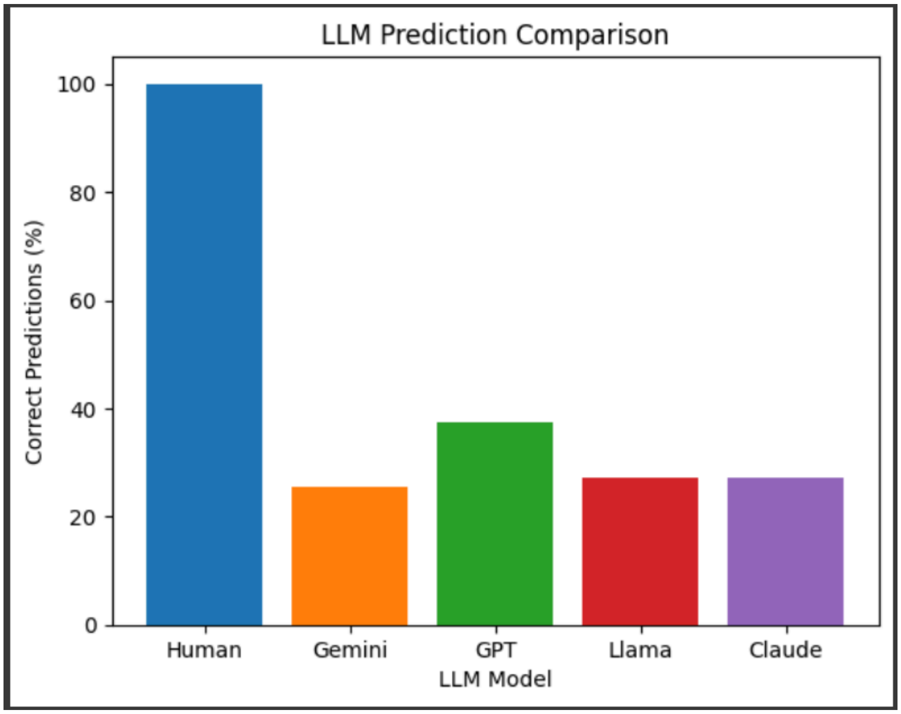


Figure 3.2: In the histogram comparing the prediction accuracy we can more clearly see that GPT was the most accurate with the others only differing slightly.

Conclusion

In conclusion, evaluating the performance of LLMs is very complex, with multiple factors affecting the model and its generated content. From our results, we can understand that while all the LLMs performed relatively similarly, OpenAi's ChatGPT rated higher than the others with an almost 16% higher F1 score, it also picked the most true positives from the dataset. This shows that GPT is the best for relevancy classification tasks. This study could not effectively find the underlying reasons for the overall low performance of the models and therefore requires further research. The dataset, prompt guidelines, and provided code[7] in this study serve as a foundation for future studies to uncover more of the complexities influencing LLMs. A better understanding of these complexities will help advance LLM performance and aid in the development of more robust and accurate generative AI models.

[7] https://github.com/avinashc1047/Summer_FellowsRsrch.git