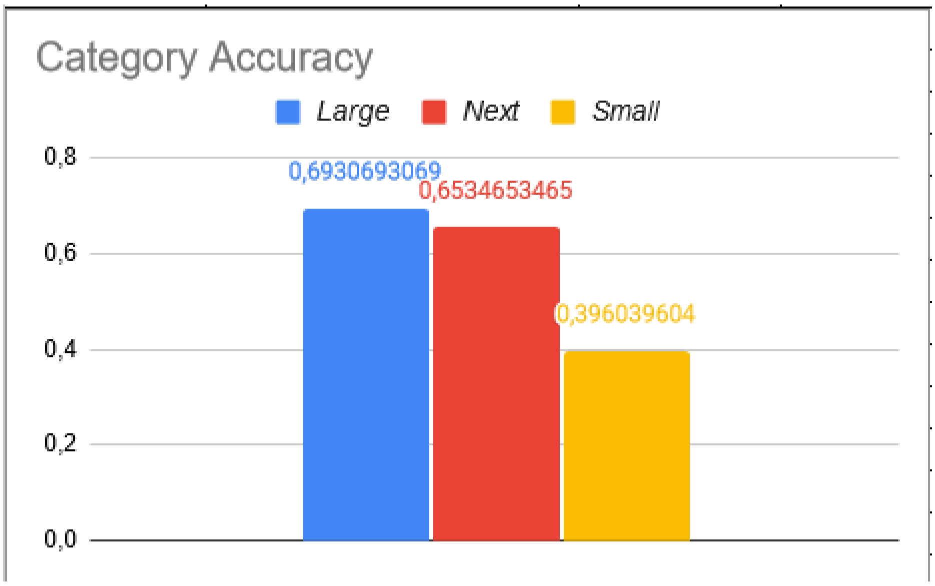
Category

The accuracy scores are quite similar, once we take the reworked score from ChatGPT as it took some time to understand how to work with the models. ChatGPT has the highest accuracy score with 0,86, followed by Mistral (0,76) and then Claude (0,75).

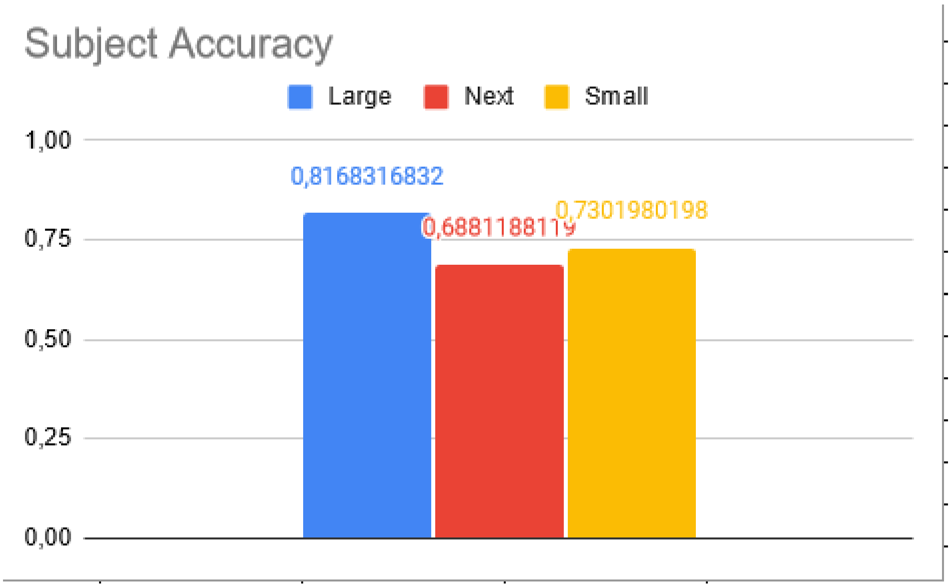
The differences in Mistral were not surprising, as the large option did largely better than its two counterparts.



Subject

For the classification, according to the main subject(s) of the texts, Mistral Large performs better than both ChatGPT(0,69) and Claude (0,68).

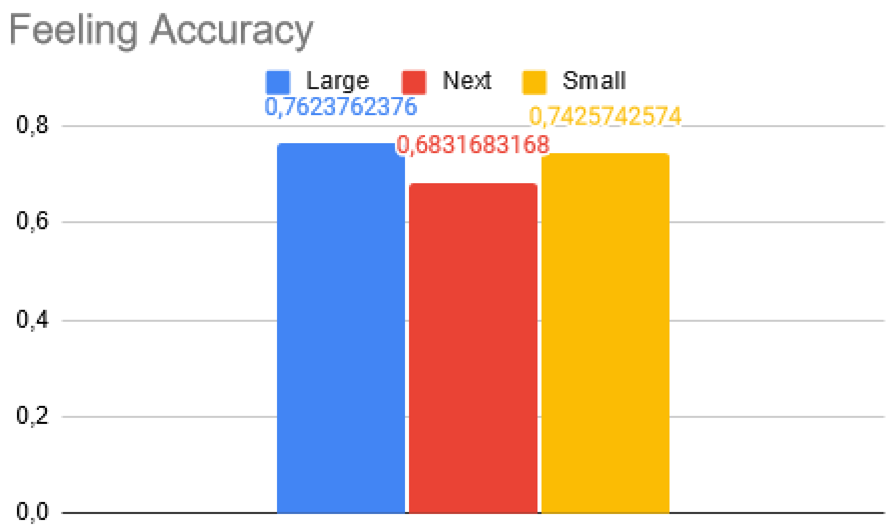
Between the three different Mistral models, the "small" performs surprisingly better than the "Next" model. However, the Large remains the most accurate of the three.



Feeling

For the feeling categorisation, both ChatGPT (0,83) and Claude (0,78) are showing higher accuracy rates than Mistral Next (0,77). The most mistakes for each model were to differentiate from a neutral/negative take on a subject. Also, a lot of "positive" words have been taken too literally, so it is possible to see that as soon as a "good" or a positive word was in a message, it would be categorized as "positive" even if it was, in fact, a neutral message.

In Mistral, Large still possesses the highest accuracy, although again, followed by Small and then Next. This is surprising, as I initially assumed that Next would perform better than small, but that was the case for only one of the categorization tasks.



Categorizing according to their own categories

As a step, I also asked each of the models to categorize the text snippets according to the categories they thought would fit best. Claude and ChatGPT ended with somewhat similar categories, often joined, such as PvP, Player strategies and meta discussions (ChatGPT), PvP balance/meta (Claude), but some interesting results game back. Claude for example categorized a few texts in "Gaming/New World (Amazon Game)" and where Mistral categorized most of the text snippets into "Gameplay Experience/Matchmaking" which is technically not wrong, simply a very large categorisation.

General feeling towards the game

As a finishing touch, I asked the model to resume the general feeling of the player base towards the game. They all answered similarly as they all worked with the same data. The same problems were put forward, balance, bugs, lack of content, as well as a similar hope for the game to get better as it does have a good combat system and it only needs better care for end-game content.

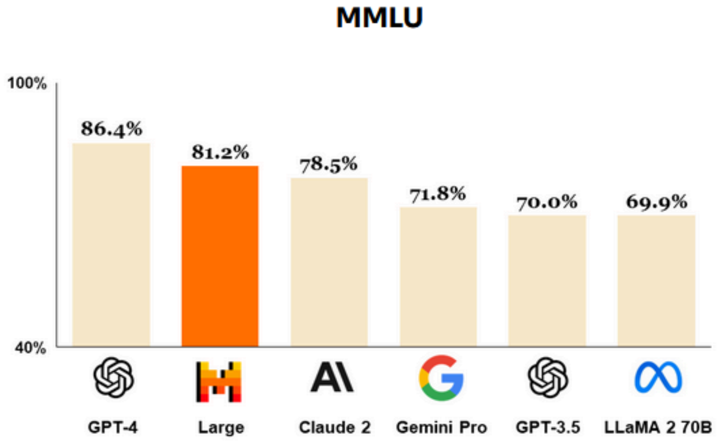
Working with the Models

Working with the different models for the same categorisation should've been similar enough, but the work method had to be slightly modified and tweaked for each model, to make the work easier or even possible. For example, in ChatGPT, it was very important, otherwise it would start coming up with its own category or start ignoring the initial prompt. It was also important to give the different texts in small doses, as a too long message was simply not accepted by the model. On the contrary, Claude could deal with all the data at once, but the free version is limited to a few messages every few hours, so it was important to do as much as possible in the limited messages available. For Mistral, it could take the long messages, but then would sometimes stop doing its task in the middle of the message, but sometimes would accomplish everything as needed. For Mistral, re-using the same chat and only writing the new task worked nicely, unlike the other models, where it was necessary to start a chat per task.

One thing common to all three models (as well as all three mistral) was the fact that they would skip some texts and not categorize them. I am not sure of the reason, as the data was pre-worked as to make clear when a new text was coming: "the text is put in between brackets",,, and end with three commas. The texts were also sent as individual paragraphs in the text itself. I then had to go through all the texts that were skip and re-ask the models to categorize those specifically, which could sometimes take up to five tries to achieve all the needed data.

Mistral

Mistral was a model that I didn't know existed, before a friend recommended I take a look at it. It is made by, mostly, an european team and is still in development, like the other models. I decided to give it a try, as we had not talked about it in the class. It offered three different kinds of models, Large for high complexity tasks, model to be said that rivals with ChatGPT 4, Next, a prototype, and Small for rapidity and efficiently accomplishing smaller tasks. The model performs quite well, as seen with the above results and can be a good alternative to ChatGPT.


MMLU
GPT-4
Large
Claude 2
Gemini Pro
GPT-3.5
LLaMA 2 70B