Part 1:

When looking at the research questions, it is too broad and vague. This means it needs to be narrowed down to be more specific. One option is: *are key concepts of a story produced by an LLM closer to the books or the movies?* In this way, it is more specifically focused on certain aspects of the story and it makes it easier to answer the sub questions that will come up with the main research questions. For example, comparing certain aspects of the script and the book with each other, such as a character or a plot device. In order to answer the above research question, data collection is necessarry.

For this research, books that were adapted into movies were researched. For this research the movie Twilight was picked, a movie adaptation of the first book in the Twilight series, written by Stepehenie Meyer. The second movie, Twilight New Moon and its movie counterpart was also used. Next, the movie script and the book were closely compared with each other. This was done by comparing certain scenes of the movie with its book counterparts. By doing this, it is possible to compare certain characters and plot devices with each other, which then could be analyzed to answer the research question. However, measuring the tendency of movies versus books is not the easiest. One way to measure the tendency is by clustering, where the results are analyzed to find underlying patterns or commonalities, which could show the (non)existing similarities between the movie and the book. In order to run this experiment, probing will be used with ChatGPT and DeepSeek. ChatGPT will be used because it is one, if not the biggest, LLM-model that is used currently. DeepSeek will be used to compare their use of data to ChatGPT, especially since ChatGPT is created by a Chinese company and they could have a different source for their data. The probing questions will be as followed:

- Textual similarity probing: extract sentences from the big and movie to compare their similarity
- Knowledge probing: these are questions such as multiple choice and fill in the blank questions to compare the book and movie versions.
- Close reading probing: an event that happens in the book, does it happen similarly in the movie?

The questions:
What makes Bella think Edward is a vampire?
How does Bella confront Edward about being a vampire?
How does the confrontation happen in the book?
How does Bella confront Edward about being a vampire in the movie?
Bella gets a papercut, what does Jasper do? A: he goes crazy B: he does nothing C: Jasper is able to bite her.
In the book, months?
In the movie, months?

However, the project does have limitations and pitfalls. LLMs are capable of having a training bias, which means that the answers they produce may not be accurate, unless they are specifically probed into that direction. This makes it harder to judge their accuracy. Another

problem is that it is hard to judge visual representations i.e. the movie may choose to show the audience through a visual instead of through dialogue. Another problem is the legal issues surrounding the project. Since copyrighted scripts and books require a specific type of handling, the legal department will be involved in this process to make sure everything is handled legally. The legal department has the resources to deal with strict copyright laws, which can also differ per country. As the movies and books are produced in the United States, but the research is conducted in the Netherlands, there may be other issues around copyright legality since it can differ from country to country. The legal department is able to help with specific issues with these copyright laws. It can also help with ethical questions around the legality of the project, since it is using copyrighted work. For annotation, as it is a Digital Humanities project, having an annotator could be very useful for this project as well. The annotator is able to help with labeling the data, as well as making sure the dataset is correct. They could also help with annotating the semantic alignments between the movies and books, as well as annotating the cluster analysis between the books and movies. By annotating the data, the data can also be used for machine learning.

For this experiment, a pdf of the two Twilight books and movies were closely compared with each other to find scenes that were deemed important. Then, ChatGPT was used to test it by using close reading probing and knowledge probing. Next, ChatGPT was compared with DeepSeek by using the same probing questions used with ChatGPT. After doing this, the answers from the probing questions were compared. One thing I found was that the ChatGPT version was more accurate when it came to answering the probing questions. When asked the questions without any specifications if it was the book or the movie, the answer would always be about the books. It seems that this LLM has a preference for the books. This is probably because it is easier to add text to a database, rather than a movie. However, when probed about the movie, the answer was correct. This was different with DeepSeek, which would give the wrong answer when asked *Bella gets a papercut, what does Jasper do? A: he goes crazy B: he does nothing C: Jasper is able to bite her.* It falsely identified the scene as being from the first book, when it actuality it is a scene from the second book, New Moon. When asked *In the book, months ?,* it falsely assumed that the question was a follow up question to the paper cut scene, when in actuality it is a different scene in the New Moon book and movie. ChatGPT did get this right, even when probed by the follow up question *in the movie months?*. Overall, DeepSeek seemed to also favor the books over the movies. DeepSeek did seem to directly quote scenes from the books more often than ChatGPT did. However, when looking at the final answers, it seems that LLM's have a bias towards books instead of movies, and only answer questions about the movies when probed.

Part 2:

The few-shot prompting strategy appeared to perform better than the zero-shot approach. In zero-shot classification, the model was given lyrics and simply asked to determine the genre without prior context. This led to inconsistent predictions, sometimes returning descriptive explanations rather than just the genre name. The lack of clear guidance on expected outputs made it harder for the model to generalize.

However, in the few-shot prompting approach**,** I provided multiple labeled examples before asking for a genre classification. This helped steer the model towards the desired output format and gave it a clearer understanding of genre distinctions. This made it easier for the LLM to give a more accurate reading. However, it seemed that this model did struggle with a few genres, for example genres that were less common. For example, the Japanese lyric was classified as Electronic, even though it is a pop song. It also wrongly classified a Rock song as Pop. However, the few shot prompting approach is preferred, since feeding it more information also gives more information back, compared to the few shot prompting.

When looking at the recal, F1 and precision scores, The poor precision, recall, and F1-scores suggest that the predicted genres did not match the actual genres in many cases. This suggests that the model's understanding of genre classification based on lyrics alone is limited. Especially with non-English lyrics, the LLM seems to have poor grasp on identifying lyrics by genre if a song is not in English. Some genres also have overlap, such as in lyrics or musical background. This also makes it harder for the LLM to pick up on the subtle differences in order to classify genres.

In order to improve this design, more few-shot-prompting approaches are necessary to help the language model learn to distinguish between genres based on lyrics. For example, asking to only give the genre names, and have more explicit questions, such as "answer with Pop, Rock, Country or Hip-Hop. Also feeding the machine more lyrics that are non-English can play a major role of its understanding of different genres.