Aalto University
MS-E2177 - Seminar on Case Studies in Operations Research

Inclus: Interim report

Milana Begantsova, Project Manager

Rami Echriti

Kalle Johansson

Jaakko Vauhkonen

inclus

1 Changes in the objective

The project's objectives and scope have remained unchanged from the original project plan. We are trying to improve the quality of the LLM outputs and LLM's ability to interpret given data through two different approaches. These are still the quantitative-to-qualitative, and qualitative-to-quantitative approaches as described in the project plan.

The secondary objective of providing considerations to avoid AI hallucination is still within the scope of the project, as well as providing a use case summary for Inclus' risk assessment. One method to explore hallucination is model fine-tuning through OpenAI's API.

2 Project Status

2.1 Current progress of the project

We started our project by conducting background research to gain a thorough understanding of the topic. Initially, we followed the original schedule outlined in our project plan, meeting as a team once a week. However, we soon recognized the need to increase our meeting frequency to 2-3 times a week due to the creative nature of prompt engineering that ended up requiring more brainstorming with the entire group. Additionally, losing one team member has increased the workload for the remaining team members.

Besides our initial project plan, we did not manage to make any progress regarding fine-tuning. This was due to the lack of access to that part of the LLM. However, this issue has now been resolved. In our meetings we have focused on prompt engineering and finding related papers for the literature review. While we have successfully addressed the tasks outlined in our project plan, staying on the schedule has proved challenging due to complications arising from prompt engineering. Progress on the prompt engineering part is discussed in more detail in the following sections.

2.1.1 Discussion on "quantitative-to-qualitative" approach

We have advanced in a "quantitative-to-qualitative" approach. Employing the LLM model, we translated comments into numerical values on a scale of 0-5 for each dimension. Subsequently, we compared the values provided by the LLM model to the actual data. While the majority of the results were accurate, there were some outliers that did not align with the data.

To refine the process, we supplied the LLM with examples of comments representing specific values for each dimension. Utilizing these examples as context, the model evaluated the actual comments. This adjustment appeared to enhance the results, although we have yet to conduct sufficient benchmarking to substantiate these improvements.

We have also been attempting the LLM to list all the risks that were provided with the most number of comments. Initially, the outcomes appeared inconsistent with the actual data. However, after experimenting with various prompts, that is, with prompts with different words choice, instructions and output formats, we managed eventually to achieve meaningful results. It was observed that requesting the LLM model to provide results in JSON format (JavaScript Object Notation, which is essentially a way to format data [1]) generally led to more accurate outcomes. Nevertheless, some outliers and miscalculations continued to appear when validating the results manually.

2.1.2 Discussion on "qualitative-to-quantitative" approach

We have made progress with the "qualitative-to-quantitative" approach, which focuses on turning qualitative data (such as answer comments) into quantitative data. The main reason for that is that given comment we can get its numerical representation and then train a traditional machine learning model. For the prompt, we set GPT-4 a role as risk analysis and natural language processing expert. There is a set of restrictions in the prompt to ensure better result:

1. Do not add any additional text except for the required output format
2. Treat each comment equally
3. If the assistant is not sure of the answer, it can return the text 'None'

4. Each comment should be accessed next features with regards to associated risk type: cohesion, objectivity of author, argumentative, top 10 keywords related to risk and not common words, dimension of comment (impact/likelihood), semantic of the comment, if is author is expert in risk analysis, does the comment fully describes risk - 'fullness'.
5. The output is fixed to asterisk separated list of features for each comment:
   i. comment*cohesion*expert*dimension*list of keywords*semantic*objectivity*argumentative

Some contextual information about the company itself is also provided. After all contextual information is given, we list example comments along with associated risk types. From our experiments, it is important to provide risk type too, since some comments do not have enough information.

The question itself is posed like "Find embedding for each comment in the provided list (for all 20)" with a given output format. Currently, it takes 3-4 minutes to process 20 comments.

We haven't come up with an idea of how to validate the accuracy of the embedding and set of useful features is not in the final form yet. At least it seems that that no halucination happens with key words.

2.2 Ongoing tasks

Currently, our focus has been directed on prompt engineering and acquiring meaningful outcomes from the LLM-model. Challenges encountered with prompt engineering included data hallucination and inconsistency in results. Across various iterations, the LLM-model would alter results, failing to yield accurate outcomes.

To address this problem, we are implementing several strategies. Firstly, we are improving the prompts by adding additional constraints and limiting the output format and the number of words in the output. Moreover, we are expanding the model's understanding by providing it with more background information and examples of the expected output and behavior.

2.3 Next tasks

We will continue with the two main approaches and we will start the validation part of our results. Our current models are quite slow and take over 20 minutes to run and we will try to find ways to make them faster.

We will try to change the temperature of the LLM model and try to improve the results. Changing the temperature is expected to affect how much the LLM model hallucinates data. We will start working with fine-tuning the LLM method. Previously we had problems with the LLM server and it did not allow us to do this but now the problem seems to be fixed and we can start to work with this approach.

We will start working on the literature review and study more about data hallucination and the "Overloading the model with too much information" effect. We will continue working with our benchmark questions and try to build a model that can answer those questions correctly.

3 Changes in the initial project plan

One of our low-probability risks was realized two weeks after presenting the project plan. Our previous project manager decided to drop the course and stop working on this project, which caused a significant disruption in progress.

We had to redistribute a large number of the current tasks in progress, including choosing a new project manager. This also affected the distribution of future tasks. Our leaving team member was in a responsible role working on the quantitative-to-qualitative approach. Their sudden departure from the team meant continuing work on this task was difficult. Completed tasks, mainly those regarding the project plan and the backgrounds of the project, have been omitted.

We also had a temporary problem with the LLM server, which prevented us from utilizing the fine-tuning of the model. This complicated progress, but this issue was quickly resolved with help from Inclus.

3.1 Updated schedule

Due to the change in the team members, the progress of the project has been delayed. An updated schedule is presented in Figure 1. This schedule is greatly adjusted from the previous one, as we have fallen behind on

progress. The implementation category of the tasks has also been segmented more clearly to represent work on each of our project objectives. We also increased the amount and emphasis of our literature review to better align with the timing of our tasks.

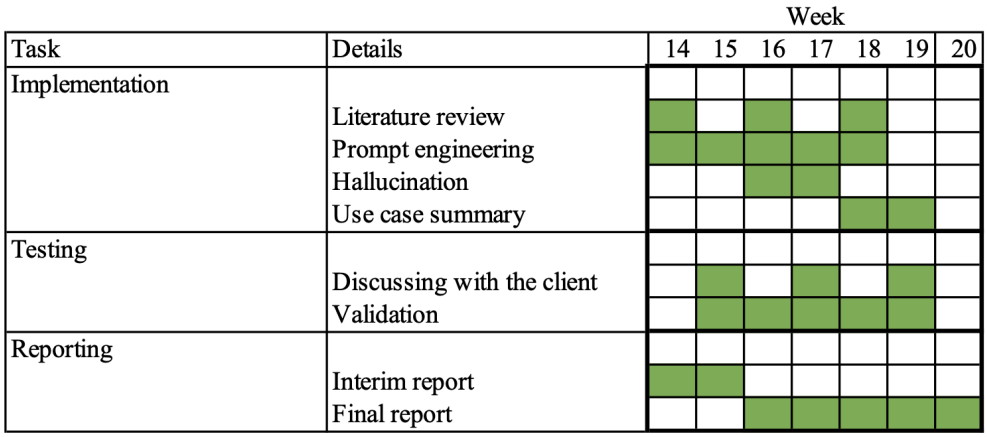

Figure 1: The updated schedule for the project.

3.2 Updated risk management plan

In general, the change in the team composition due to led to an increased workload for the remaining team members, but the workload has remained manageable. The number of weekly meetings has been increased to ensure active contribution to the project and to improve the pace at which we move forward. This helps with staying on schedule and keeping team members motivated.

The risk assessment table itself remains mostly unchanged from the initial project plan. Further inactivity or team members leaving the project would now have an even larger impact on our ability to finish this project. Conversely, the likelihood of more members quitting is now extremely low. The updated assessment table is presented below.

Risk	Effect	Probability	Impact	Prevention
Insufficient scoping	An excessive workload and the client is unhappy about results	Medium	Medium	The team engages in regular meetings with the client and effectively strategizes and plans the project.
Unable to find literature for the project	Not enough justification of the LLM outputs' validity. The team does not have enough expertise about the topic.	Medium	High	Try to find sources outside the commonly used platforms
Team member inactivity	**Increased workload for other team members**	**Very Low**	**Very High**	**Regular meetings and All team members contribute to every part of the project, if possible**
The client stops communicating	Feedback for progress is harder to get	Low	High	Try to maintain active communication through the project and complete all assignments well before the deadline to endure possible communication delays.
Problem with the LLM-server	**Testing becomes significantly more difficult**	**Medium**	**High**	**Try to complete all tasks well before the deadline so possible delays can be endured**
Information under the NDA is leaked to outside parties	Client is harmed and access to their resources can be restricted or the team faces potential legal consequences as specified by the contract	Low	High	Using only Inclus's GPT access point and getting Inclus's approval for presentations that potentially include sensitive data

Table 1: The updated risk assessment table for the project. Realized risks are bolded.