Description of Planned Statistics for the Project

Lawrence Fulton (lawrence.fulton@rwth-aachen.de)
Computational Social Systems - RWTH Aachen

August 14, 2025

1 Prompt Sensitivity

1.1 Prompt Generation

For the later research it is important to understand how sensitive the model is with regards to the prompt. For this we will run the experiment with different prompts and then compare the results. The prompts will be created the following, where the content inside the square brackets is the type of prompt, indicating if it is either "system", "user", or "assistant". There are always two different versions per prompt, one for the LLM to generate a new reply for the debate and one for it to generate a value how much the LLM agrees to the statements. The content inside the curly brackets is a placeholder for the actual content that will be inserted later and will be the same for all prompts. The {user_answer} and {assistant_answer} will be the previous replies of the user and the assistant.

Table 1: Overview of Variables for Prompt Sensitivity Experiment

Variable	Example
experiment_scenario	Du diskutierst über die Aussage: Die Förderung von Windenergie soll beendet werden.
background_story	Ich bin 75 Jahre alt und männlich. Ich habe einen Hauptschulabschluss, ein mittleres monatliches Haushalts-Nettoeinkommen und ich bin nicht berufstätig. Ich bin etwas religiös. Politisch-ideologisch ordne ich mich in der Mitte ein. Ich identifiziere mich mäßig mit der Partei SPD und habe SPD gewählt. Ich lebe in Westdeutschland. Ich finde, die Regierung sollte die Einwanderung einschränken und habe keine Meinung dazu, ob die Regierung Maßnahmen ergreifen sollte, um die Einkommensunterschiede zu verringern.
question_prompt	Auf einer Skala von 1 bis 7: Wie sehr stimmst du der Aussage zu: Die Förderung von Windenergie soll beendet werden. Antworte nur mit einer Zahl.

For each prompt versions there will be two different prompts. The first one (Discussion Prompt) will be used to continue the discussion. The second one (Answer Generation) will be

used to generate the numerical answer to the question found in question_prompt. Thus the question_prompt placeholder will always only be found in the Answer Generation prompt. The answer from the Discussion Prompt will then be used in future discussions steps and the answer from the Answer Generation will be used later to analyse the change of opinion over time.

Listing 1: Prompt 1 - Discussion Prompt

```
[System] Scenario: {experiment_scenario}
Background Story: {background_story}
The following is a debate between you and another person. Complete your next
   ↪ reply. Keep the reply shorter than 30 words and in German.
[User] {user_answer}
[Assistant] {assistant_answer}
[User] {user_answer_2}
{...}
```

Listing 2: Prompt 1 - Answer Generation

```
[System] Scenario: {experiment_scenario}
Background Story: {background_story}
The following is a debate between you and another person.
[User] {user_answer}
[Assistant] {assistant_answer}
[User] {user_answer_2}
[Assistant] {assistant_answer_2}
{...}
[User] {question_prompt}
```

For Prompt 2 we will provide the full sentence in the system prompt instead of only providing the keyword ("Background Story" or "Scenario").

Listing 3: Prompt 2 - Discussion Prompt

```
[System] The scenario is the following: {experiment_scenario}
This is your background story: {background_story}
The following is a conversation between you and another person. Complete
   ↪ your next reply. Keep the reply shorter than 30 words and in German.
[User] {user_answer}
[Assistant] {assistant_answer}
[User] {user_answer_2}
{...}
```

Listing 4: Prompt 2 - Answer Generation

```
[System] The scenario is the following: {experiment_scenario}
This is your background story: {background_story}
The following is a conversation between you and another person.
[User] {user_answer}
```

```
[Assistant] {assistant_answer}
[User] {user_answer_2}
[Assistant] {assistant_answer_2}
{...}
User: {question_prompt}
```

The third prompt version we change the tone to be very polite. Yin, Wang, Horio, Kawahara, and Sekine (2024) found that the politeness of prompts can significantly affect LLM performance. Additionally we will modify the structure of the background_story to use the the word "You" over "I", thus resulting in sentences such as "Du bist 75 Jahre alt und männlich. Du hast einen Hauptschulabschluss..."

Listing 5: Prompt 3 - Discussion Prompt

```
[System] Please imagine the following scenario: {experiment_scenario}
Here is your background: {background_story}
Kindly respond in German to the next message from another person. Please
    ↪ keep your reply under 30 words.
[User] {user_answer}
[Assistant] {assistant_answer}
[User] {user_answer_2}
{...}
```

Listing 6: Prompt 3 - Answer Generation

```
[System] Please imagine the following scenario: {experiment_scenario}
Here is your background: {background_story}
Kindly respond in German to the next message from another person.
[User] {user_answer}
[Assistant] {assistant_answer}
[User] {user_answer_2}
[Assistant] {assistant_answer_2}
[...]
[User] {question_prompt}
```

1.2 Prompt Sensitivity Analysis

1.2.1 Visualisation

To analyse the sensitivity of the model to the different prompts we will run the experiment with each prompt and then compare the results. Since we have seven parties (Die Linke, SPD, Bündnis 90/Die Grünen, FDP, CDU/CSU, AfD, default) and we are testing all parties with all parties this will lead to a total of $\sum_{i=1}^{7} i = 28$ different combinations of parties. For each combination we will run the experiment with each of the three prompts, leading to a total of $28 \times 3 = 84$ experiments. Each experiment will be repeated 5 times which results in a total of $84 \times 5 = 420$ experiments per question. For now I am only applying it to one question but this is easily extendable to all questions. We prompt the LLM to generate us

the answer to the {question_prompt} at every 4th step of the discussion. The LLM will then generate a number between 1 and 7, which is the answer to the question. We will then compare the answers of the LLM for each party and each prompt.

To get a graphical overview over the spread of the results per prompt we have normalised the data for each party at each discussion step. This is done since we expect that members of different parties will have very different opinion to the topics given and this would allow us to compare all of the results. The method used for the normalisation is the Standard Scaler from the 'sklearn' library. The Standard Scaler normalises the data by subtracting the mean and dividing by the standard deviation. This results in a distribution with a mean of 0 and a standard deviation of 1. The results can be seen in the following:

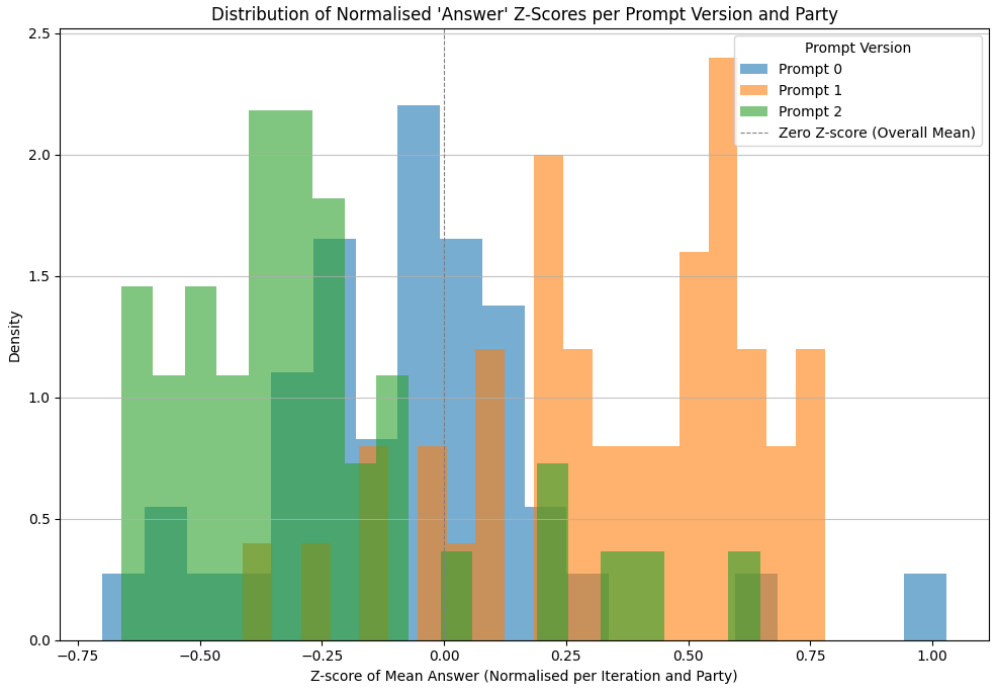

Figure 1: Histogram of the normalised results over each party and prompt. The x-axis shows the normalised values and the y-axis shows the frequency of these values. The different colors represent the different prompts used.

In Figure 2 you can see the non-normalised results for the party "CDU/CSU" for each prompt with regards to the the question "Die Förderung von Windenergie soll beendet werden.". One can observe that there is a clear difference in the responses depending on the prompt used based on this example with it appearing that prompt 1 leads to a higher response than prompt 2.

1.2.2 Statistical Analysis

To analyse the statistical significance of the differences between the prompts we will use a mixed-effects model. The mixed-effects model allows us to account for the fact that we have repeated measures from the same debate sessions. We will use the 'statsmodels' library in Python to fit the model.

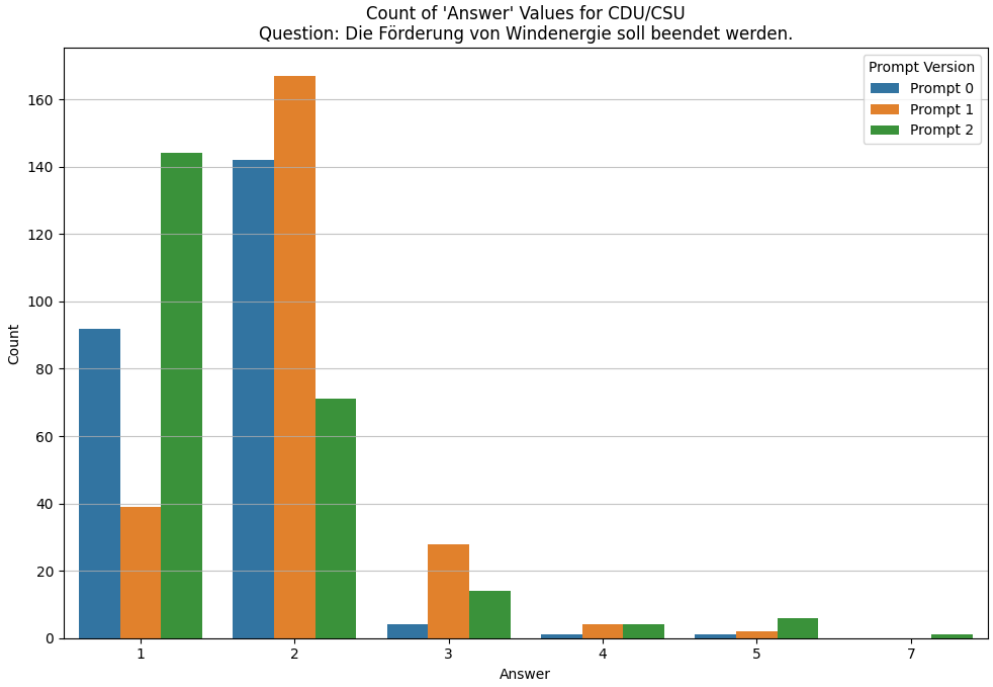

Figure 2: Results for the party "CDU/CSU" for the question "Die Förderung von Windenergie soll beendet werden." for each prompt. The x-axis shows the discussion step and the y-axis shows the response of the party.

In a mixed-effects model, we distinguish between fixed and random effects.

- **Fixed Effects** are the variables we want to investigate directly. Our fixed effects are 'version' (the prompt version), 'party', and 'time', excluding their interactions. This allows us to test whether the prompt 'version' has a significant effect on the 'answer', and how this effect might change depending on the party or over time.
- **Random Effects** account for sources of non-independent variance in the data. In this experiment, multiple answers are generated within the same debate. To account for this clustering, we will include a random intercept for each unique debate session. A debate session is uniquely identified by the combination of the parties debating, the prompt version, and the repetition number.

The model will have the following structure:

$$\text{answer} \sim \text{version} + \text{party} + \text{time} + \text{question\_index} + (1|\text{debate\_id}) \quad (1)$$

where 'debate_id' is a unique identifier for each experimental run. This model structure allows us to test our hypotheses about the prompts while correctly controlling for the nested structure of the data.

1.3 Results

The results of the mixed-effects model are presented in Table 2. The analysis reveals a statistically significant effect of the prompt 'version' on the 'answer' provided by the model.

Table 2: Mixed-Effects Model Regression Results for ‘answer‘

Effect	**Coef.**	**Std.Err.**	**z**	$P > \|z\|$
Intercept	1.455	0.066	22.076	<0.001
party[T.AfD]	2.106	0.081	26.128	<0.001
party[T.Bündnis 90/Die Grünen]	0.444	0.081	5.512	<0.001
party[T.CDU/CSU]	0.663	0.081	8.224	<0.001
party[T.Die Linke]	0.028	0.081	0.352	0.725
party[T.FDP]	0.133	0.081	1.652	0.099
party[T.SPD]	0.361	0.081	4.483	<0.001
version[T.out_1]	0.297	0.053	5.634	<0.001
version[T.out_2]	-0.132	0.053	-2.503	0.012
time	-0.158	0.005	-30.385	<0.001

Note: Group Var = 0.280. The baseline for ‘version‘ is Prompt 1 and the baseline for ‘party‘ is the default party.

The model indicates that the prompt version has a significant impact on the generated answers. Compared to the baseline (Prompt 1), ‘version[T.out_1]‘ (Prompt 2) is associated with a significant increase in the answer value by 0.297 ($\beta = 0.297$, SE = 0.053, $z = 5.634$, $p < 0.001$). Additionally, ‘version[T.out_2]‘ (Prompt 3) is associated with a significant decrease in the answer value by 0.132 ($\beta = -0.132$, SE = 0.053, $z = -2.503$, $p = 0.012$). These results confirm that even slight variations in prompt wording can lead to statistically significant differences in the model's output. As expected, the ‘party‘ and ‘time‘ variables were also significant predictors.

Thus we can conclude that the model is sensitive to the prompt variations, and this sensitivity can be statistically quantified using a mixed-effects model. The results suggest that the choice of prompt can significantly influence the model's responses, which is crucial for interpreting the outcomes of our experiments.

2 Change of opinion over time

Seeing that the prompt version has a significant effect on the ‘answer‘ generated by the model, we now have to take this into consideration when analyising the change of the opinions over time and are not able to just focus on one prompt as which would have been sufficient if the prompt had no effect.

The research question we are interested in is "Does discussion cause agents' opinions to significantly shift from their starting points and converge towards more uniform stances?". To answer this question we will have to analyse the change of the opinions over time with the focus on two different aspects: Firstly, the change of the opinions over time, quantified by calculating the distance of the opinions at each time step to the initial opinion. Secondly, the convergence of the opinions over time, quantified by calculating the variance of the opinions at each time step. These metrics are question agnostic, allowing us to compare the results across different questions and prompts.

2.1 Distance to Initial Opinion

To analyse the change of opinion over time, we will calculate the distance of the opinions at each time step to the initial opinion. This will allow us to see how much the opinions have changed over time. The distance is calculated as the absolute difference between the initial opinion for each agent in each debate and the opinion at each time step. Since we calculate the difference from the initial opinion, and the difference between a number and itself is 0 we will ignore timestep 0 for this analysis (though still added in the figure). A plot of the distance to the initial opinion over time can be seen in Figure 3.

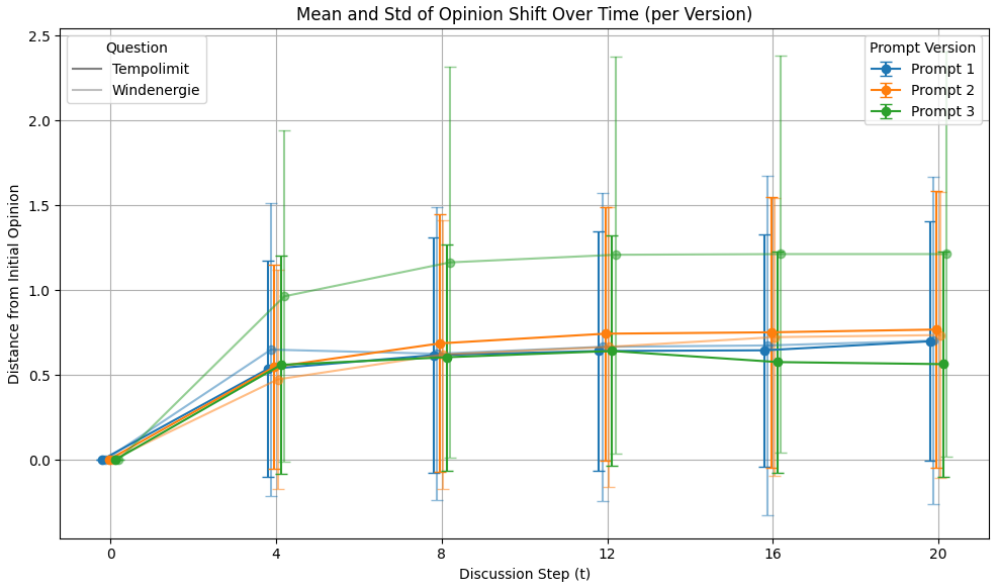

Figure 3: Distance to Initial Opinion Over Time

We again use the mixed-effects model to analyse the change of opinion over time. The model will have the following structure:

$$\text{distance} \sim \text{party} + \text{time} + (1|\text{debate\_id}) \tag{2}$$

where ‘distance‘ is the distance to the initial opinion, ‘party‘ is the party of the agent, and ‘time‘ is the time step with the same random effect structure as before. There is a significant effect of time onto the answer ($\beta = 0.008$, SE $= 0.003$, $z = 0.008$, $p < 0.004$).

We withhold the variable ‘version‘ for now since the categorical value ‘version‘ is one-hot encoded. Thus by adding the variable the first version will be considered the baseline. The interactions between time:version thus only show the differences between the versions and the baseline version. Therefore if we want to see if time is significant we have to subdivide the data into the different versions and then analyse the significance of time for each version separately. Applying another mixed-effect model on the reduced dataset with only data from one prompt version we will get three further results with all being significant (p=[0.01,<0.001, <0.001]) and positive effect sizes $\beta = [0.008, 0.018, 0.013]$.

Thus we can conclude that the distance to the initial opinion is significantly affected by the time step, meaning that the opinions shift over time. Though the effect sizes are fairly small, indicating that the opinions do not change drastically over time.

2.2 Variance of Opinions

To analyse the convergence of the opinions over time, we will calculate the variance of the opinions at each time step. This will allow us to see how much the opinions are converging towards a common stance. The variance is calculated as the average of the squared differences from the mean opinion at each time step. This reduces the dataset quite drastically since we now aggregate over time steps. A plot of the variance of the opinions over time can be seen in Figure 4.

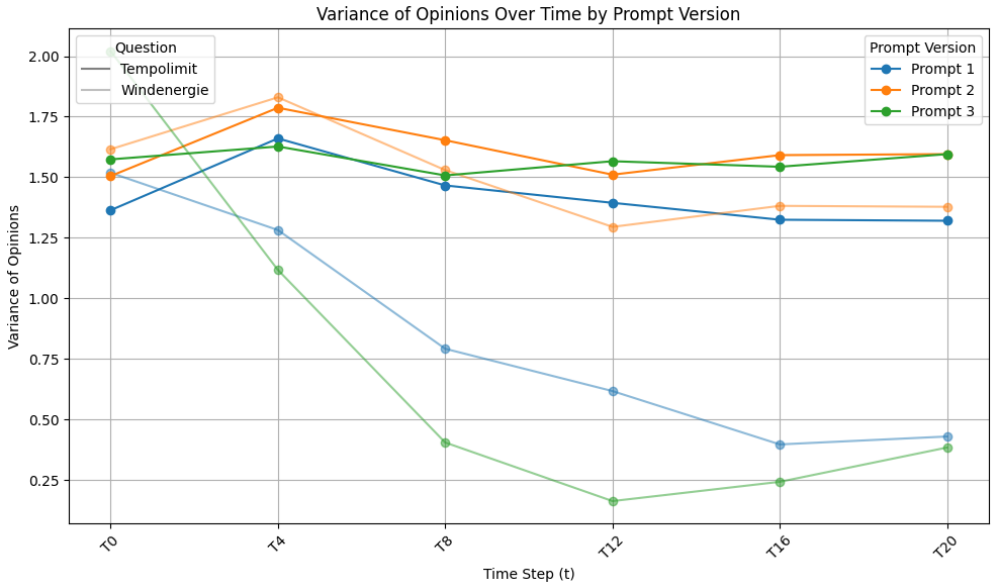

Figure 4: Variance of Opinions Over Time

Since we are firstly interested in the convergence of the opinions over time we create the following mixed-effects model:

$$\text{variance} \sim \text{time} + (1|\text{prompt}) + (1|\text{question}) \tag{3}$$

where ‘variance‘ is the variance of the opinions, and ‘time‘ is the time step. The random effects are the prompt version and the question index. The model shows a significant effect of time onto the variance ($\beta = -0.028, SE = 0.008, z = -3.437, p = 0.001$). This indicates that the variance of the opinions decreases over time, even only slightly. These results have to be interpreted with caution, as the effect size is very limited right now (3 prompts * 2 questions).

References

Yin, Z., Wang, H., Horio, K., Kawahara, D., & Sekine, S. (2024). Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance. In *Proceedings of the second workshop on social influence in conversations (sicon 2024)* (pp. 9–35).