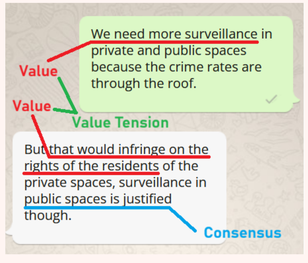
Extracting Value, Value Tensions and Consensus from Deliberations

Can LLMs extract value, value tensions, and consensus points from multi-stakeholder deliberative transcripts?

Ananya Singh Michaël Grauwde Willem Paul-Brinkman

A Deliberative Conversation



Motivation

- Deliberation is important but **hard to scale**
- Human moderators can be **expensive**, possibly **biased** and **time-consuming**
- Normal summaries risk obscuring minority opinions
- LLM comparatively leave groups **less separated**

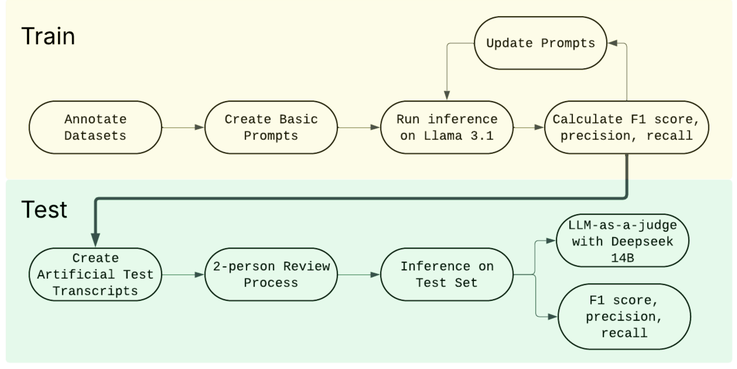
Research Questions

1. **Which LLM most accurately** extracts deliberative constructs from multi-stakeholder transcripts, as measured by metric evaluation and an LLM-as-a-judge evaluation?
2. **Which prompting strategy most reliably** elicits correct and complete deliberative construct extraction, as measured by metric evaluation and an LLM-as-a-judge evaluation?
3. How does the **interaction between model and prompting strategy** affect extraction quality?

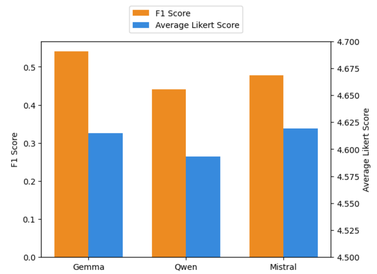
Design Choices

LLM Models	Mistral 7B, Qwen 2.5 7B, Gemma 2 9B
Prompting Strategies	Zero-Shot, Few-Shot, Chain-of-Thought (CoT)
Training Dataset	UK House of Commons (Hansard)
Test Dataset	Artificially created (Public Safety)

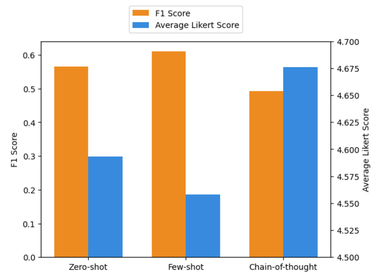
Method



Results



(a) Comparison of average Likert scores (LLM-as-a-judge) and F1 score for the 3 different LLMs



(b) Comparison of average Likert scores (LLM-as-a-judge) and F1 score for the 3 different prompting strategies

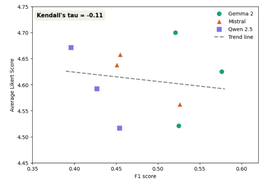


Figure 2: Trends between LLM-as-a-judge scores and metric evaluation across all model prompt combinations

Discussion

- Best F1 score (0.576) achieved by zero-shot Gemma was only marginally better than inter-annotator agreement F1 score (0.5)
- Results aligned with Babatunde et al. [1], which implies CoT's added complexity negatively impacted its results
- Implies the use of LLMs as a triage tool instead of full automation
- High LLM-assigned Likert scores for LLM-as-a-judge might be attributed to leniency bias

Conclusion

- **Most accurate LLM:** Gemma 2 as per F1 score, Mistral as per LLM-as-a-judge
- **Most accurate prompting strategy:** Few-shot as per F1 score, chain of thought as per LLM-as-a-judge
- **Interaction between model and prompting strategy:** no single model-prompt combination dominated across both methods (Kendall's Tau = -0.11), implying the interaction between model and prompt is sensitive to how quality is defined
- **Future work**: Using large LLMs, branching into more prompting techniques, evaluating the scores with a human evaluation study

LLMs are capable of extracting value, value tensions, and consensus points but the final judgment should stay with a human

References

[1] Ibukun Babatunde, Obiabuchi Nnanna, and Mark Klein. Moderating large scale online deliberative processes with large language models (llms): Enhancing collective decision-making., March 2025.

TU Delft