

Roundtable: Cultural, Political & Societal Perspectives on LLMs

- How do cultural differences affect the performance and perception of LLMs?
- Which cultural biases have you observed in your own work?
- What role should political frameworks (e.g., EU AI Act, digital sovereignty) play in shaping NLP research?
- Could we jointly design evaluation datasets that better capture cultural variation?
- What low-effort collaboration could we start on this topic (e.g., survey, shared benchmark)?

Notes / Key Takeaways:

Roundtable: Multilinguality & Representation in LLMs

- What are the most pressing gaps in multilingual coverage today (scripts, dialects, low-resource languages)?
- How do we evaluate true multilinguality vs. translation bias?
- Are new representations (e.g., hyperbolic embeddings, script-aware pretraining) a promising direction?
- Could our two groups pool resources to build a shared multilingual evaluation set?
- What joint experiments would test cross-lab ideas (e.g., same dataset, different models)?

Notes / Key Takeaways:

Roundtable: Alignment, Safety & Generalization of LLMs

- What does alignment mean in practical research terms — aligning to humans, tasks, values, or safety constraints?
- Which methods seem scalable for reasoning models (e.g., RLHF alternatives, constitutional AI, data curation)?
- How do we measure generalization in NLP beyond benchmarks?
- Where do safety concerns overlap with generalization failures?
- Could we run a joint benchmarking effort or hackathon-style project across both universities?

Notes / Key Takeaways:

Roundtable: Human-Centric NLP: Labels, Explanations & Interaction

- How do we handle human disagreement in annotation? Should models learn the variation instead of collapsing it?
- What are the trade-offs between explainability for researchers vs. for end-users?
- How can human feedback loops be made sustainable for PhD-scale research (small teams, limited annotation budgets)?
- Could we collaborate on shared annotation pipelines or a joint user study?
- What tools, datasets, or frameworks do we already have that could be shared?

Notes / Key Takeaways: