CS 329X: Human Centered LLMs

Summary and Open Questions in Human Centered LLMs

Diyi Yang

Announcement: Final Presentation

- Dec 5th:
  - 4-6pm PT, **sign up for a slot**
    - https://docs.google.com/spreadsheets/d/1PRP_qEzc4Wgk1IACN8VYSsMVNs-3_sBEy_UKPrrJZuU/edit?usp=sharing
  - You're very welcome to stay for the entire session
  - 5 mins for presentation, 3 mins for QA
  - Make the presentation informative and engaging :)

Announcement: Final Report

- Final Project Report (👉 Dec 10, 11:59 PM PT)
  - No late days

The final paper should be 8 pages long, in ACL submission format and adhering to ACL guidelines concerning references, layout, supplementary materials, and so forth.

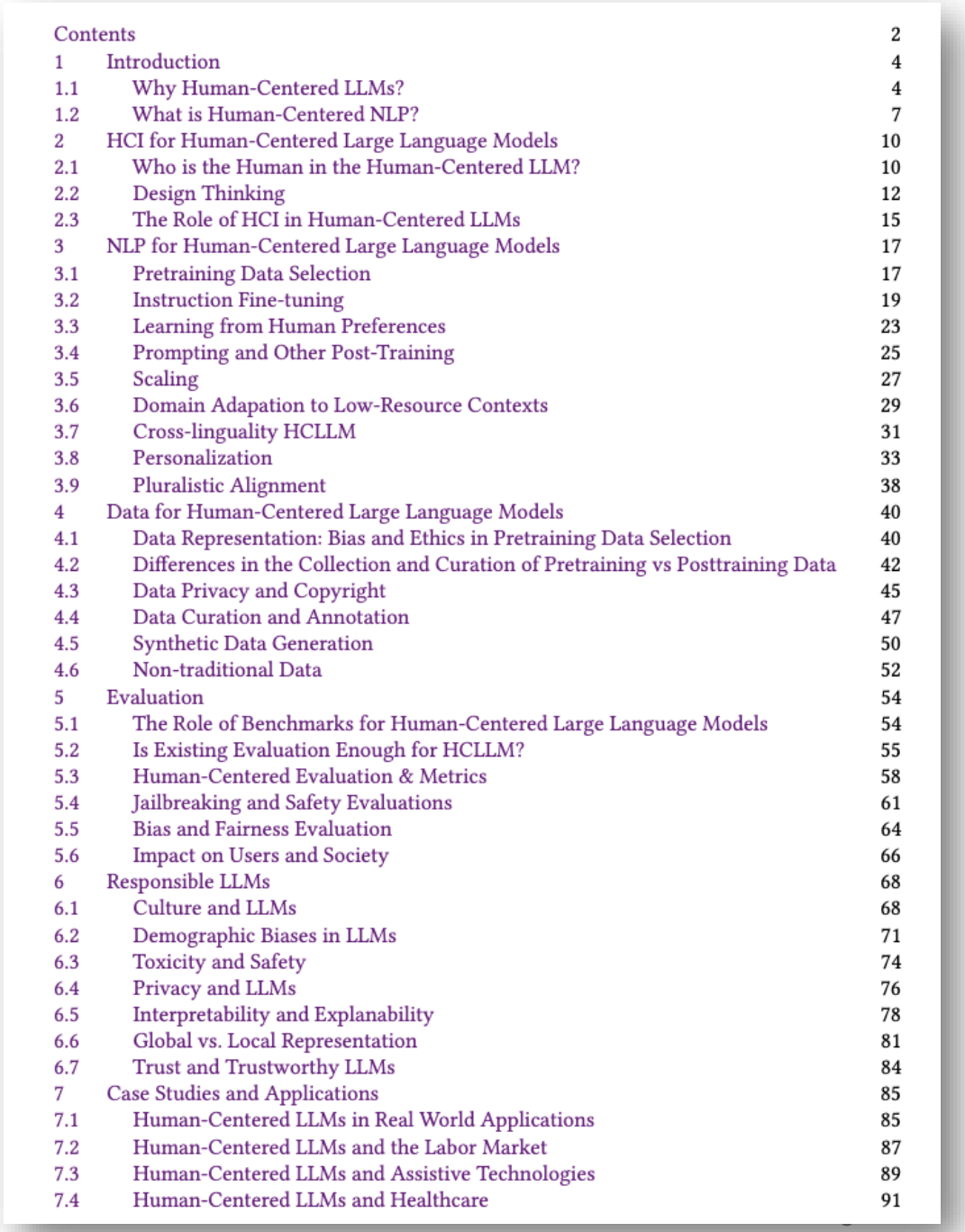
Below are the required components for the final paper:

1. *Introduction* (3 points)
2. *Related Work* (3 points)
3. *Methods* (7 points)
4. *Results* (10 points)
5. *Discussion/Conclusion* (1 point)
6. *Ethical Consideration*: Please write an explicit discussion section of any potential ethical issues, such as around the ethical implication of the project, the use of the data, and potential applications of your work. Here are some recommendations from ACL's ethics guideline: *"Ethical questions may arise when working with a variety of types of computational work with language, including (but not limited to) the collection and release of data, inference of information or judgments about individuals, real-world impact of the deployment of language technologies, and environmental consequences of large-scale computation."*
7. *Authorship statement*: At the end of your paper (after the 'Acknowledgments' section in the template), please include a brief authorship statement, explaining how the individual authors contributed to the project. You are free to include whatever information you deem important to convey. For guidance, see the second page, right column, of this guidance for PNAS authors (p. 12). We are requiring this largely because we think it is a good policy in general. This statement is required even for singly-authored papers, because we want to know whether your project is a collaboration with people outside of the class. *Only in extreme cases, and after discussion with the team, would we consider giving separate grades to team members based on this statement.*
8. *References*

Class-Level Report

- Thanks for your hard work!
- If you want to be involved as a co-author for a high-quality report, sign up to contribute

https://forms.gle/hQ7UymXxBDAJrxSz8



CS329X: HCLLM

Diyi Yang

Rose E. Wang

Caleb Ziems

- **Website:** http://web.stanford.edu/class/cs329x

Why CS 329X: HCLLM

- **Both NLP and HCI** perspectives in the age of LLMs
  - NLP people know the standard method of data preparation, training, evaluation, and deployment.
  - HCI people know ways to mimic natural use scenario, collect human feedback, design interactions...
  - Both are needed for human-centered LLMs
- **Different aspects** from language, vision, robotics, health, education, social science...
- **Expectation: research seminar** with a few deep-dive lectures

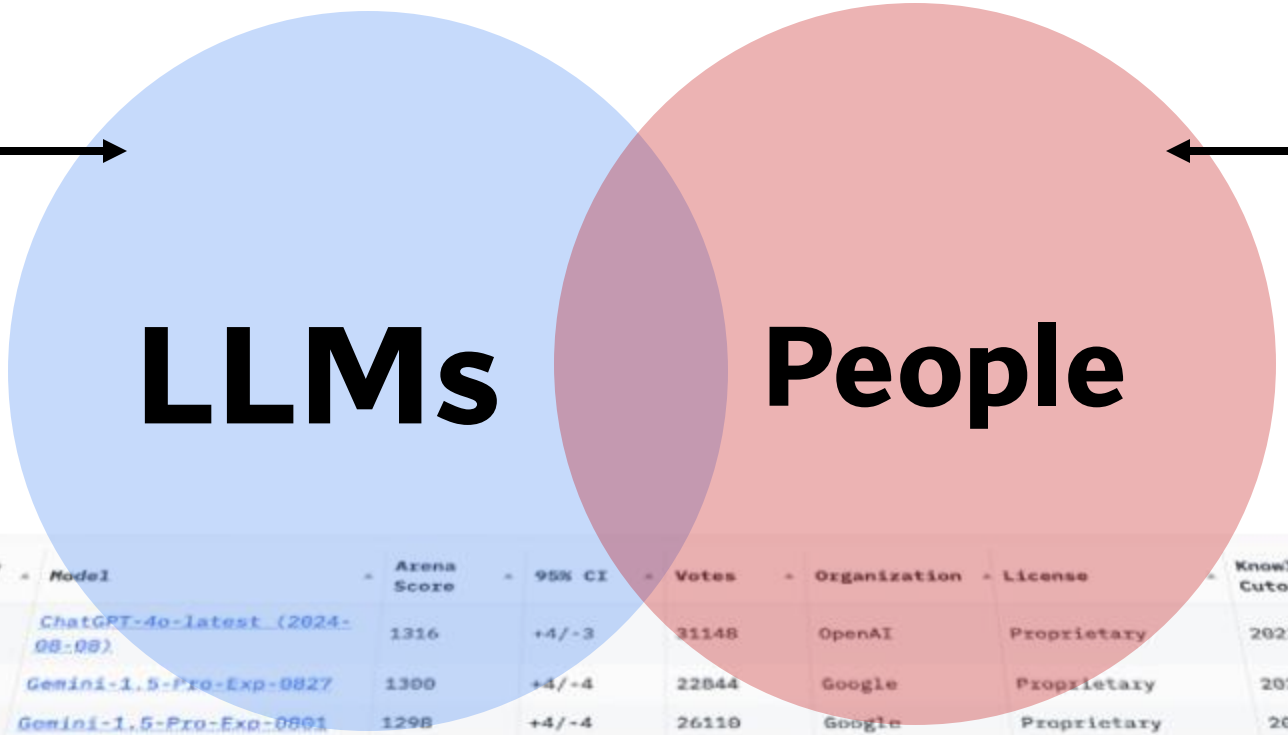
What is Human-Centered NLP?

Human-centered NLP involves

designing and developing NLP systems in a way that is attuned to the needs and preferences of humans, and that considers the ethical and social implications of these systems.

- It involves multiple development stages
- It needs to be optimized for humans

Reasoning
Benchmarking
Robustness
Generalization
Verification
Infrastructure
Efficiency
Scalability
Interpretability
...



ank* UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
	ChatGPT-4o-latest (2024-08-08)	1316	+4/-3	31148	OpenAI	Proprietary	2023/10
	Gemini-1.5-Pro-Exp-0827	1300	+4/-4	22844	Google	Proprietary	2023/11
	Gemini-1.5-Pro-Exp-0801	1298	+4/-4	26110	Google	Proprietary	2023/11
	Grok-2-08-13	1294	+4/-4	16215	xAI	Proprietary	2024/3
	GPT-4o-2024-05-13	1285	+3/-2	86306	OpenAI	Proprietary	2023/10
	GPT-4o-mini-2024-07-18	1274	+4/-4	26088	OpenAI	Proprietary	2023/10
	Claude 3.5 Sonnet	1270	+3/-3	56674	Anthropic	Proprietary	2024/4
	Gemini-1.5-Flash-Exp-0827	1268	+5/-4	16780	Google	Proprietary	2023/11
	Grok-2-Mini-08-13	1267	+4/-4	16731	xAI	Proprietary	2024/3
	Meta-Llama-3.1-405b-Instruct	1266	+4/-4	27397	Meta	Llama 3.1 Community	2023/12

Personality
Social Factors
Culture and Value
Privacy
Ethics
Fairness
Interaction
Trust
Positive Impact
...

Human-centered LLMs should be in every stage

What we have covered: Foundational Basics

- Foundational Basics (Week 1 to Week 5)
  - ➢ The Ultimate Crash into NLP and HCI
  - ❖ Learning from human preferences
  - ❖ Personalization vs. collective opinion in preference tuning
  - ❖ Data, data and data
  - ❖ Design thinking + natural language as the new user interface
  - ❖ Enabling human-AI interaction
  - ❖ Evaluating human-AI interaction

What we have covered: Cutting-Edge Topics

- Cutting-Edge Topics (Week 5 to Week 10)
  - ❖Culture and values in LLMs
  - ❖Risks, trust and safety
  - ❖Creativity and productivity

What we have covered: Hot-take debate

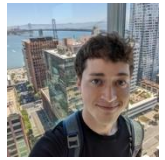
- As an artistic tool, does Generative AI simply *shortcut* or truly *augment* human creativity?
- Many new papers use LLMs to simulate human behavior. What long-term impact will these methodologies have on social science?
- In light of risks around misinformation etc, which conversational style is societally more beneficial for general-purpose chat-style LLMs to adopt?
- A move towards pluralistic systems is quite likely to decrease overall usage bias in users.

We had 6 Guest Lectures!

Rose R. Wang: Human-AI Interaction in **Education**

Ryan Louie: Human-AI Interaction in **Mental Health**

Megha Srivastava: Human-AI Interaction in **Robotics**



Caleb Ziems: Human-AI Interaction in **Social Science**

Hao Zhu: LLMs for social simulation

Julia Kreutzer (Cohere For AI): Open-source multilingual LLMs

Ultimate Crash to LLMs and Prompting

- ✓**Transformers and Large Language Models**
- ✓**Prompting**
  - ✓Zero-shot, few-shot
  - ✓Chain-of-thought, tree-of-thought, graph-of-thought
  - ✓Answer engineering
- ✓**Optimization and Calibration**
  - ✓Sensitivity and inconsistency
  - ✓Output biases and calibration
  - ✓Optimization via DSPY

Learning from Human Feedback

- ✓ Different type of human feedback
- ✓ Learning from human feedback
  - ✓ Dataset updates (weak supervision, data augmentation)
  - ✓ Loss function updates (unlikelihood learning)
  - ✓ Parameter space updates (parameter efficient fine-tuning, model editing)
- ✓ RLHF
- ✓ DPO
- ✓ Limitations of human feedback

Local vs. Global Preferences

- ✓ **Constitutional AI and Collective CAI**
  - ✓ Constitutional AI
  - ✓ Collective Constitutional AI
  - ✓ Alignment with both Local and Global Preferences
- ✓ **Pluralistic Alignment**
- ✓ **Preference Tuning**
  - ✓ Group preference optimization
  - ✓ Demonstrated feedback
  - ✓ Interactive learning from user edits

Design Thinking

- ✓**Design Thinking**
  - ✓Motivation: why designs on top of LLMs are important
  - ✓Design Thinking:
    - ✓Double Diamond
    - ✓Problem Reframing
    - ✓Prototyping
    - ✓Interview and Think Aloud Studies
- ✓**Natural Language As the New Interface**
- ✓**Interface and Interaction**

Enable Human-AI Interaction

- ✓ **Ways to Enable Human-AI Interaction**
  - ✓ Different types of human-LLM interaction
  - ✓ LLM-empowered agents
  - ✓ **Learning from human feedback ++**
    - ✓ Constitutional Maker
    - ✓ Group preference optimization
    - ✓ Demonstrated feedback
    - ✓ Learning from user edits
- ✓ **Human-AI Interaction Case Studies**

Evaluate Human-AI Interaction

- ✓**How, What, Who and When**
- ✓**Ethics and Rethink Evaluation**

How are we evaluating?

Methods Quant. | Qual. *Types* Intrinsic | Extrinsic *Metric* Validated | New

What is being evaluated?

Modules Model module | HCI module (UX) | End-to-end *Goal* Utility | Satisfaction | ...

Who is evaluating?

Humans Lay users | Domain experts *Automated* LLM

When do we evaluate?

Duration Instant | Short-term | Long-term

Culture and Values in LLMs

- ✓Culture often leads to diverse interpretations
- ✓Cultural differences shape communication dynamics
- ✓Culture also matters in interpreting visual semiotics
- ✓Existing LLMs show unintended culture alignment
- ✓LLM simulations of sociocultural groups produce caricatures
- ✓Building LLMs that are culturally aware is greatly needed

Safety and Risks in LLMs

- ✓ Discrimination, Hate Speech, and Exclusion
- ✓ Information Hazards
- ✓ Misinformation Harms
- ✓ Malicious Uses
- ✓ Human-Computer Interaction Harms
- ✓ Environmental and Socioeconomic Harms

Social Implication of LLMs

- ✓ Personalized education
- ✓ Companion and support
- ✓ LLMs for science discovery
- ✓ Transformation to workforce
- ✓ Malicious uses
- ✓ Many other unknowns

Many Open Questions

- Low-resource language and dialects
- Alignment
- Evaluation and interpretability
- Global representation
- Trust
- Safety in LLMs and their applications
- Human-AI collaboration and collective intelligence
- Copyright, data and privacy
- Lots of cool applications in LLMs + societally important domains
- …

Final Thoughts

1. What is human centered NLP

2. How to build human centered NLP
   - Data, formulation, and technical challenges
   - Interdisciplinary methods

3. What does the “progress” look like for human centered NLP

4. What does HCNLP bring to society & vice versa