

Icebreaker: 2mins
Section 1: 5mins
Section 2: 5mins
Discussion: 10mins
Wrap Up Discussion: 10mins

LLMs: Evaluation

Eric Nguyen, Shigu Feng, Daniel Skopichev, Sabrina Lopez, Uttam Rao

Icebreaker: Judge LLM Responses

- You'll see a question and two LLM-generated responses.
- Your task is to read both responses and evaluate which response is better
 - Which is more creative?
 - Which is more objective?
 - Are there different 'vibes' to an LLM?

Icebreaker: Judge LLM Responses

- ChatGPT-4c**
 - Shows vivid metaphors
 - Focuses on storytelling over factual details
 - Emphasizes imagination and creativity, tone
 - Appeals to emotions rather than reason
 - Can be more creative than humans
 - May outperform humans who value creativity
 - Includes dates and historical details (e.g., 476 AD)
 - Lays out facts in a clear, organized manner
 - Appeals to logic, facts, reasoning
 - More reliable for reference-based or rule-based metrics
 - May feel dry or less memorable to human evaluators

Icebreaker: Judge LLM Responses

- ChatGPT-4c**
 - Dramatizes characters (e.g., invisible fish named Dory)
 - Emphasizes themes and lessons
 - Focuses on emotional growth and journey
 - May prioritize personal growth over factual accuracy
 - Strong candidate for human preference evaluation
 - May feel more like a friend or mentor than a neutral AI
 - Less visual overuse
- GPT-4**
 - Provides a straightforward plot breakdown
 - Lays out facts in a clear, organized manner
 - Emphasizes factual accuracy
 - More neutral tones, like a teacher or historian assessing work
 - Less likely to use metaphors or analogies
 - May feel flat or less engaging to human readers

Let's Talk Cookies

Let's say we have the recipe for the base of a cookie recipe. Let's say the base of the cookie recipe is an SOTA LLM.

- Different cookies can be made from the same base to satisfy different preferences.
- How can we understand if this base is good for all the cookie variations, or can it be improved?

Timeline

- What are LLMs being used for?
- How do we evaluate LLMs?
- How can we trust LLMs?
- What are the key definitions?

What are LLMs being used for?

Foundation Models

- Foundation models** - big model trained on broad and generally unlabeled data and is suitable for many downstream tasks
 - Based on self-supervision + DNNs (long existed)
 - E.g., GPT-3 with 175B parameters, adapted via prompt engineering → text generation
- Other key definitions**
 - Emergent** - requires human emerge, not hard coded
 - Interpretable** - consolidation of methods for building ML systems across wide range of applications
 - One model can be reused everywhere

Foundation of Foundation Models

Icebreaker: Judge LLM Responses

- You'll see a question and two LLM-generated responses
- Your task is to read both responses and evaluate which response is better
 - Which is more creative?
 - Which is more objective?
 - Are there different 'vibes' to an LLM?

