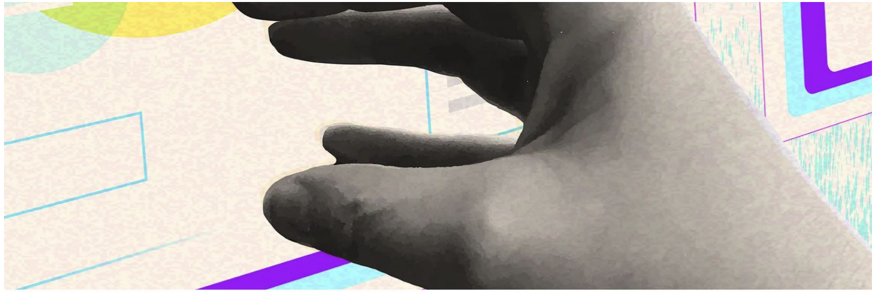
Theory of Mind in Generative Models: From Uncertainty to Shared Meaning

Malihe Alikhani

COLM 2025

Every conversation, whether between humans or with machines, is a negotiation with uncertainty.
We hesitate, clarify, hedge, correct; not because we’re inefficient, but because that’s how understanding emerges. That’s how we co-construct meaning.



CHATGPT & GENERATIVE AI

AI Tutors Can Work—With the Right Guardrails

Research suggests that unrestricted AI use can hinder learning, but AI tutors designed to prompt and prod can be powerful teaching tools. Here’s how to create one.

The New York Times

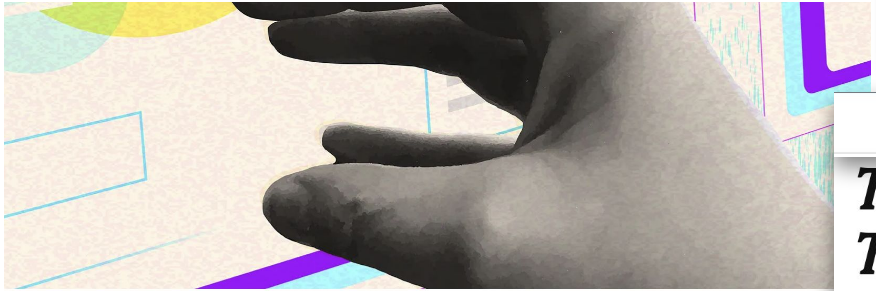
Human Therapists Prepare for Battle Against A.I. Pretenders

Chatbots posing as therapists may encourage users to commit harmful acts, the nation's largest psychological organization warned federal regulators.

The New York Times

A.I. Is Getting More Powerful, but Its Hallucinations Are Getting Worse

A new wave of "reasoning" systems from companies like OpenAI is producing incorrect information more often. Even the companies don't know why.



CHATGPT & GENERATIVE AI

AI Tutors Can Work—With the Right Guardrail

Research suggests that unrestricted AI use can hinder learning, but A tutors designed to prompt and prod can be powerful teaching tools. Here's how to create one.

The New York Times

They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling.

Generative A.I. chatbots are going down conspiratorial rabbit holes and endorsing wild, mystical belief systems. For some people, conversations with the technology can deeply distort reality.

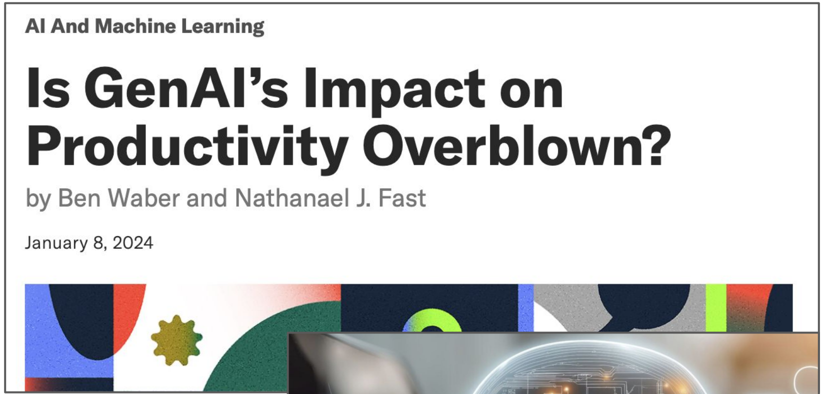
Is AI Making Us More Productive?

Who benefits from AI productivity?

Experienced developers take 19% longer to complete tasks with AI.

People with a BA degree or higher tend to be more productive using AI tools.

44% of people who are hard of hearing find AI tools like ChatGPT inaccessible.

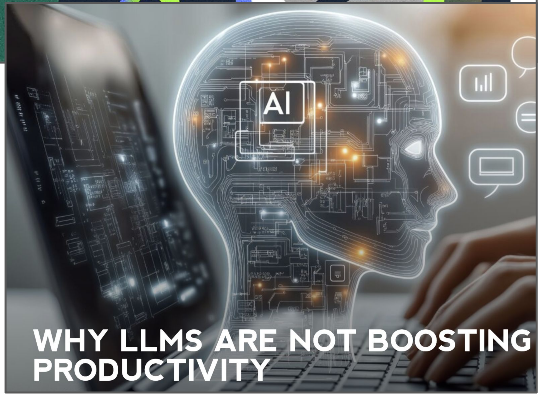


AI And Machine Learning

Is GenAI's Impact on Productivity Overblown?

by Ben Waber and Nathanael J. Fast

January 8, 2024



Why Do LLMs Struggle with Complex Tasks?

Despite their strengths, LLMs still find it hard to complete multi-step, nuanced tasks. They miss the subtle cues and deeper context that human understanding brings.

digitaltrends yahoo!tech

OpenAI's big, new Operator AI already has problems

Fionna Agomuoh

> 20% worse than humans (Open AI)

Similar struggles in Multi WOZ (Xu et al.)

How Do We Communicate to Solve Tasks?

Can you find a place to stay in Boston?

Client

Which neighborhood are you staying?

Travel Agent

Hm. Any is fine. I am visiting NEU.

I think you'd like Back Bay. There's a new listing, but I'm not sure it's quality

Effective Communication

Consider how often, during conversations, we intuitively interpret what others truly mean beyond their literal words:

Common Ground

"Should we eat at the usual place?"

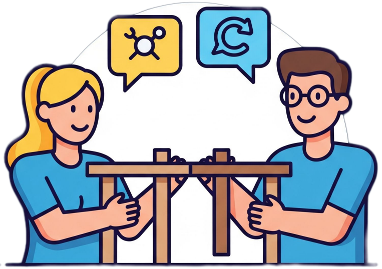
shared understanding of which restaurant is the "usual" one.
(Stalnaker, 1978)

Question Under Discussion

"What time is it?" implying "Should we wrap up soon?"

question guides the conversation.
(Roberts, 1996/2012)

Collaborative Dialogue



"This part goes here.", "Did you finish assembling the legs?"

confirming and clarifying instructions to ensure **mutual understanding and success**.
(Grosz & Sidner, 1990)

In this talk

Part 1| Uncertainty & Theory of Mind

- How to help LLMs externalize their own uncertainty?
- How to understand others' uncertainty with Theory of Mind?
- How adding deliberate friction models user mental states better?
- How can we prevent sycophancy?

Part 2 | Constructing Shared Meaning Across Diverse Modalities & Communities

- How can multimodality enrich AI's Theory of Mind?
- How can we model Theory of Mind for Sign Language Technologies?
- How can we construct shared meaning between AI and the Deaf community?




Part 1

Uncertainty & Theory of Mind

How well can Large Language Models externalize their own subjective uncertainty?

Uncertainty is Common in Conversation; E.g. Negotiations

I'm looking to buy for under $100

Sorry, **don't think I can budge**

A Deal Eventually Occurs

…throw in a bike lock or other gear?

Price is firm, that will be extra.

A Deal Does Not Occur

- Both have similar strategic behaviors, but lead to different outcomes…
- How can we disentangle the nuances that lead to deal or no deal?
- How can we model this inherent uncertainty in conversation outcome?

Can We Use Language Models to Quantify Uncertainty?

I'm looking to buy for under $100

Sorry, **don't think I can budge**

A Deal Eventually Occurs

...throw in a bike lock or other gear?

Price is firm, that will be extra.

A Deal Does Not Occur

👨 What are the chances that a deal occurs?

+

Partial Conversation

→

Direct Forecasting: Asking Language Models Directly

I'm looking to buy for under $100

Sorry, **don't think I can budge**

Maybe meet me in middle then?

Option 1: **Direct Uncertainty Estimate**

- Prompt LM and sample output

`What are the chances that a deal occurs?`

Implicit Forecasting: Using Language Model Logits

I'm looking to buy for under $100

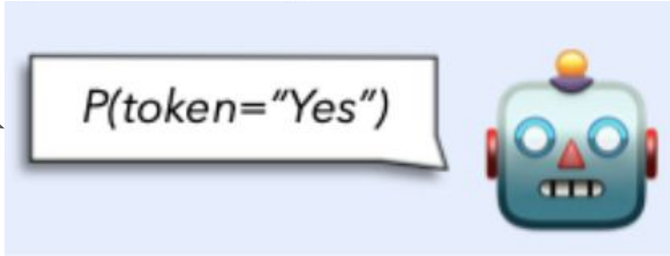
Sorry, **don't think I can budge**

Maybe meet me in middle then?

Option 2: **Implicit Uncertainty Estimate**

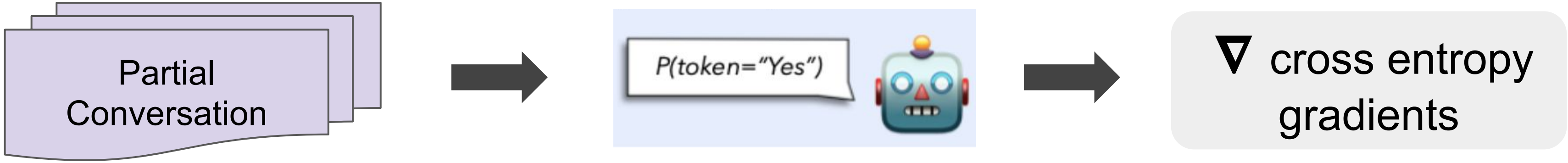
- Prompt LM and look at probability of sampling affirmation token

👨 `Will a deal occur?`

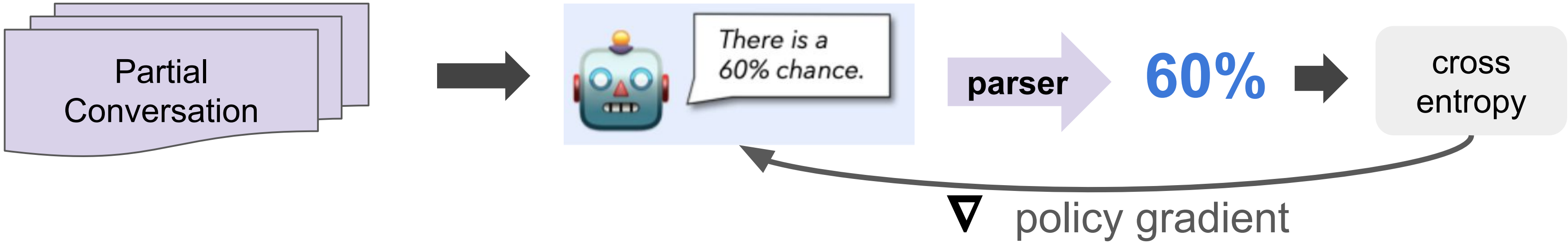


Fine-Tuning: Supervised Methods and RL

Implicit Forecasting: supervised learning, fully differentiable

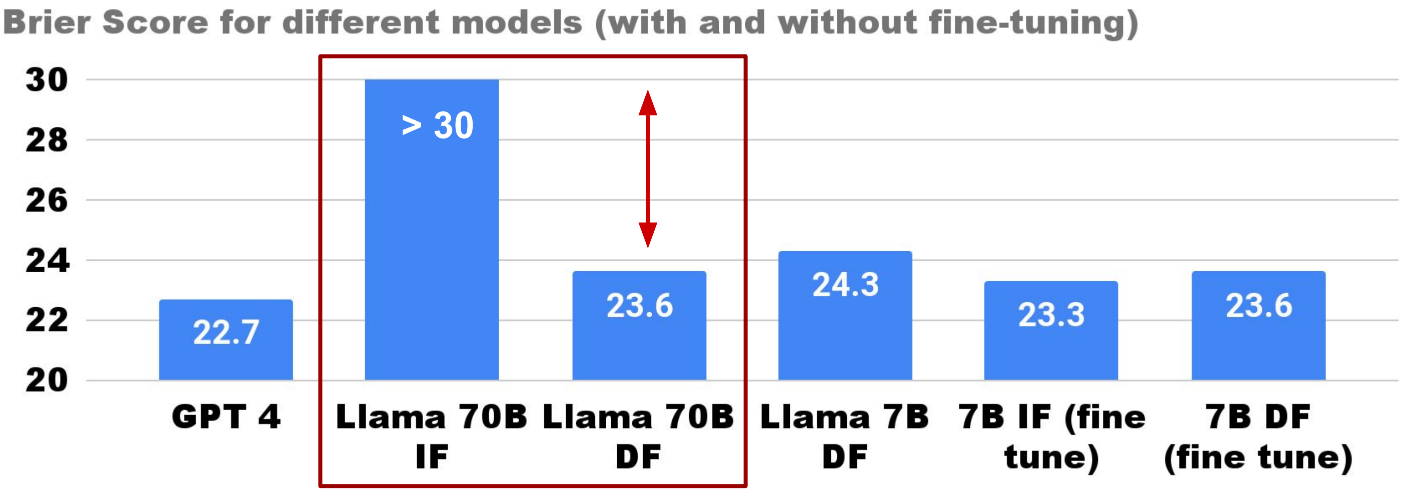


Direct Forecasting: Sampling breaks differentiability, RL saves day!



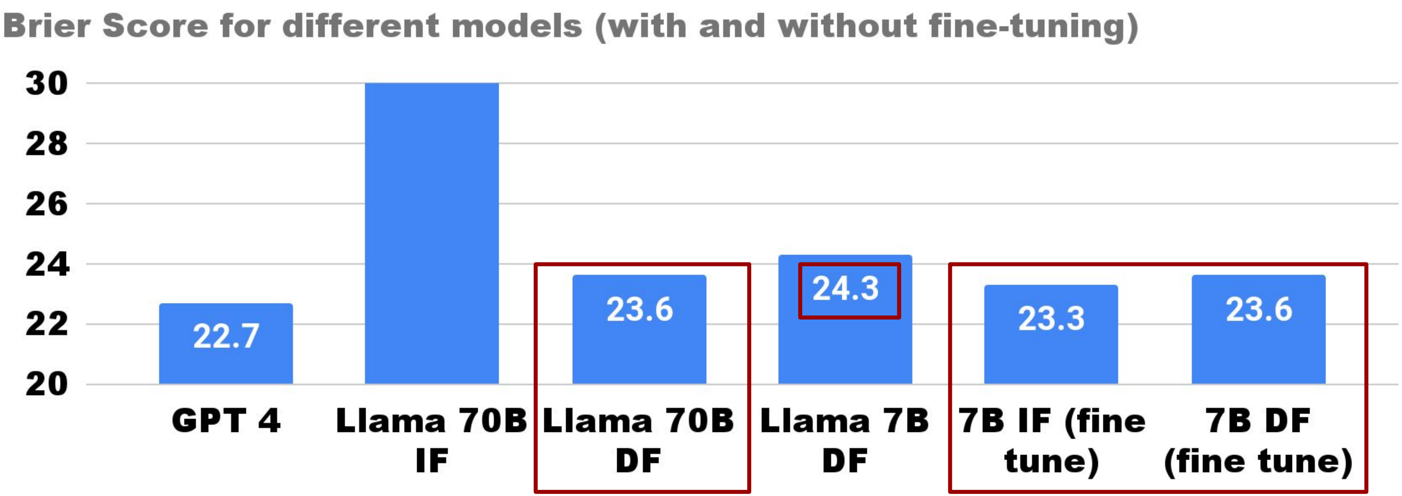
Direct Forecasting Beats Implicit Forecasting “out of the box”

- Models tested on unseen data; out-of-distribution for fine-tuned models
- Use some prompt engineering strategies



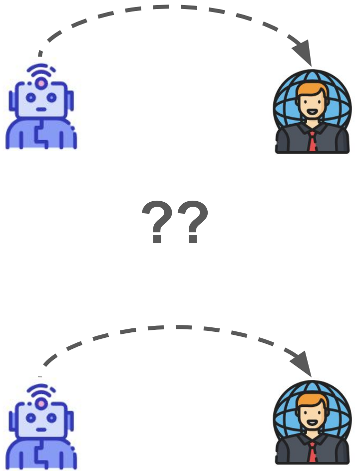
Fine-tuning Allows 7B to Rival Models 10x Their Size

- Models tested on unseen data; out-of-distribution for fine-tuned models
- Use temperature scaling and prompt engineering strategies



How can we measure and model uncertainty in others using LLMs, and what does Theory of Mind reveal about it?

Can AI Manage Uncertainty Like Humans?



Can you find a place to stay in Boston?

Which neighborhood are you staying?

Hm. Any is fine. I am visiting NEU.

I think you'd like Back Bay. There's a new listing, but **I'm not sure** it's quality

Subjective Uncertainties: Self and Other

I'm 60% sure the listing is high quality

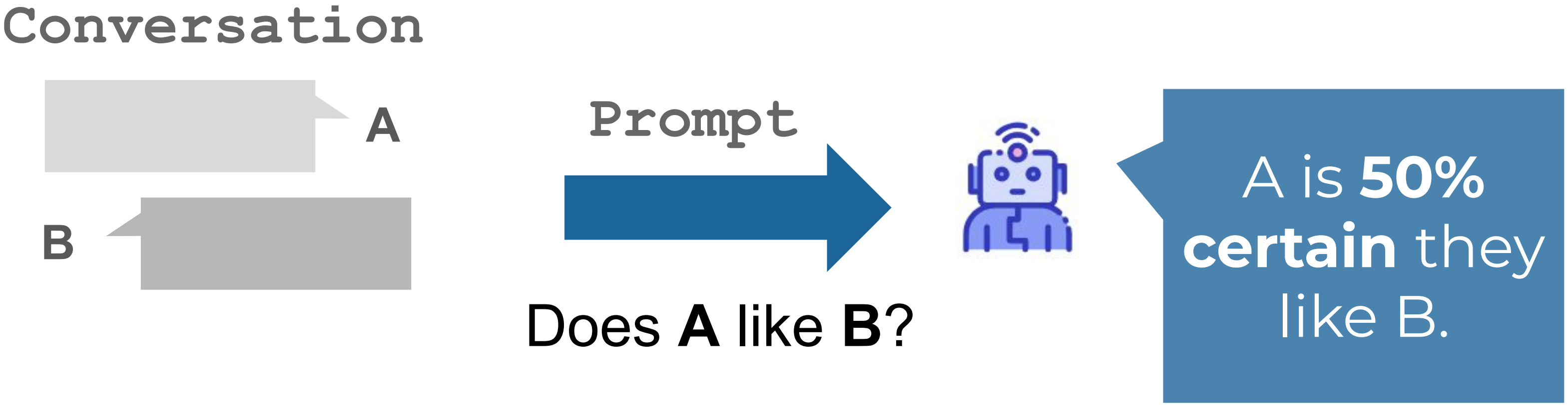
Difference in Perspective

Hm. Any is fine. I am visiting NEU.

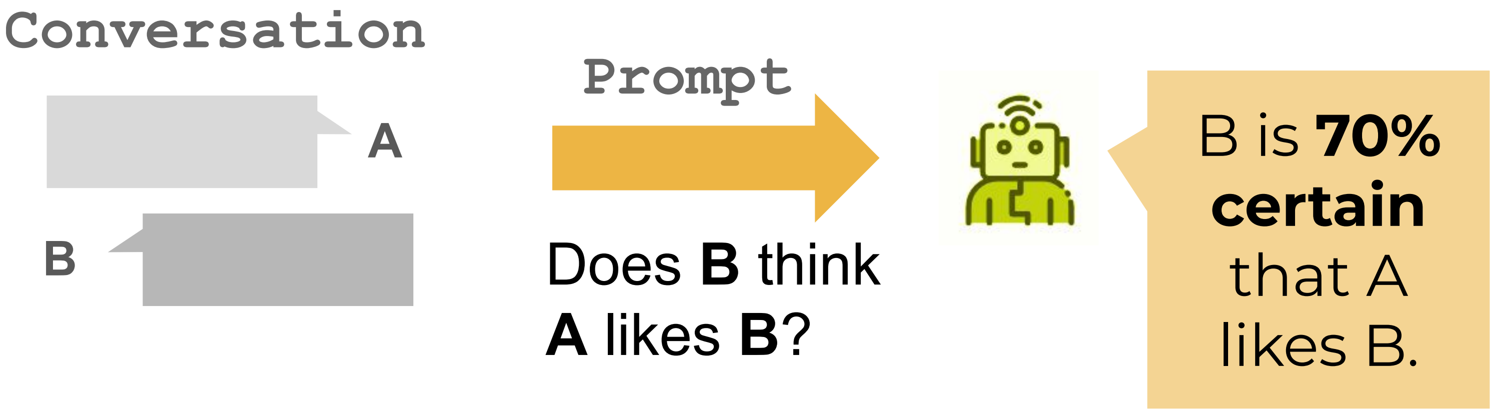
I'm only 20% sure

I think you'd like Back Bay. There's a new listing, but I'm not sure it's quality

Language Task 1: Predicting Uncertainty of Others



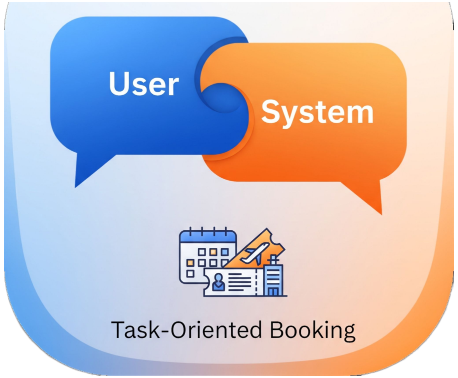
Language Task 2: Predicting Uncertainty via Theory of Mind



Modeling perspective differences is an important aspect of Theory of Mind *(Kim et al., 2024)*.

Data and Metrics

MultiWOZ

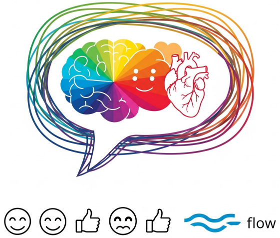


Task-oriented Wizard-of-Oz corpus for conversational booking systems.
Measures user satisfaction (5-point scale).

CaSiNo Corpus

Negotiations about camp-resource allocation. Interlocutors barter over resources (firewood, water).
Measures satisfaction with the final deal (5-point scale).

CANDOR Corpus



Large, multimodal dataset of 1656 naturalistic English conversations. **Measures:**

- Conversational structure (turn-taking, gaps, overlaps)
- Psychological content (emotions)
- Subjective judgments (enjoyment, flow)

Algorithm Contribution: Self-consistent Weighted Chain-of-Thought Prompting for Continuous tasks

We sample multiple reasoning paths ([Wang et al.]) during LLM uncertainty modeling can improve performance by reducing prediction variance

... conversation continues

A's certainty ***seems higher***

due to the emotional tone

ANSWER = 6 / 10

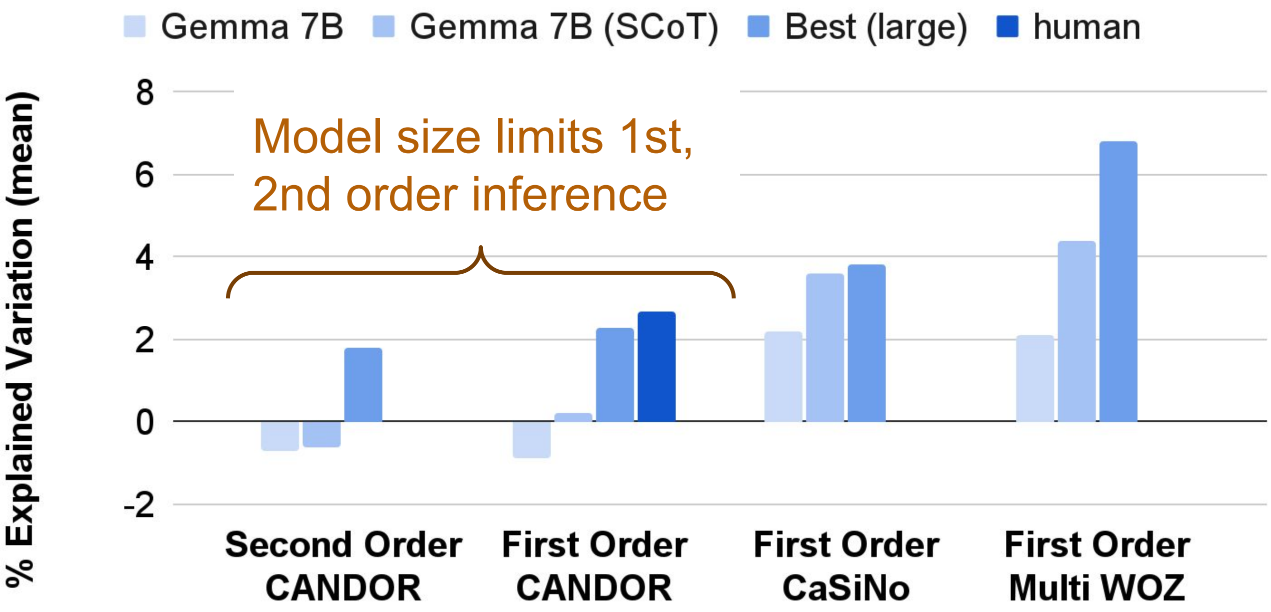
```
Prompt: Does A like B?
Let's think step-by-step to
predict A's uncertainty.
```

Prompt 10 times to gather multiple reasoning paths

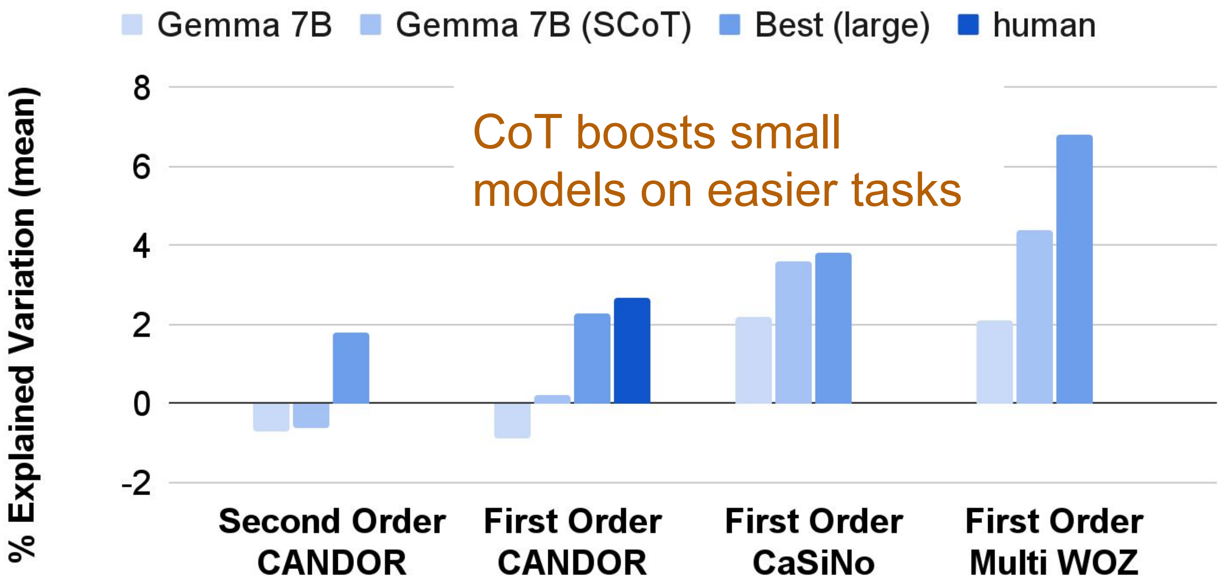
Weigh for Final Result:

avg([60%, …, 20%]) = 45%

Smaller Models Struggle when Modeling First or Second-order Beliefs about Uncertainty



Self-Consistent Chain-of-Thought for continuous tasks improves smaller model performance



Now that we can understand uncertainty, how can we create intentional pauses to reflect and recalibrate in dialogue?

Scenario with Positive Friction

Can you help me find accommodation in Boston?

Client

AI

Do you have a neighborhood preference?

Clarify goals

Hmm, anywhere near NEU is fine.

Expressing Uncertainty & Seeking Clarification

Back Bay is close to NEU. There's a new listing available, but I'm unsure about its quality. Would you like me to double check the reviews first?

- **We argue that** conversational systems should incorporate deliberate moments of **positive friction.**
- These intentional movements slow down the course of an interaction in order to yield positive long-term impact.
- This encourages *contemplative thinking* such as *reflection on uncertain assumptions* by both the users and AI systems.

Theories of discourse coherence reveal:

- *The rhythm and timing of dialogue shape the dynamics of interaction.*
- *These foster clarity and mutual understanding.*[1]

1 Stalnaker, 1978; Tannen, 1989; Wilkes-Gibbs and Clark, 1992; Zellner, 1994

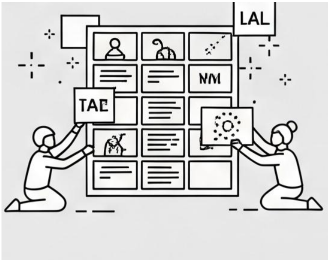
Can we use these theories to create better intentionality in human-AI dialogue?

Our Approach to Taxonomy Design

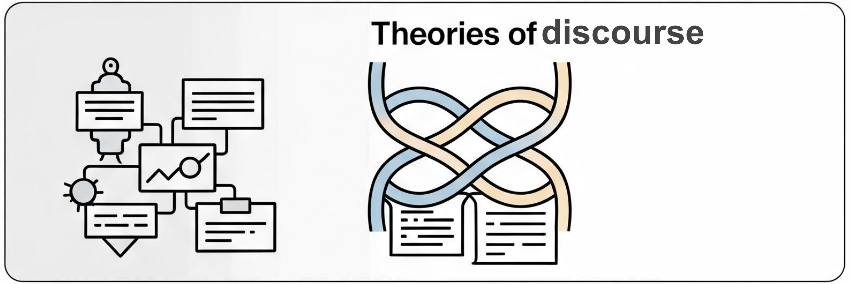
1. We started with a bottom-up data-driven approach.
2. We asked annotators how would to add friction to conversations?
3. We then qualitatively analyzed and codified different classes of friction
4. We came back and connected these classes with theories of discourse.
5. We then carried out a user study using our new taxonomy.

Add Friction

Codify Types of Friction



Connect with Theories of Discourse

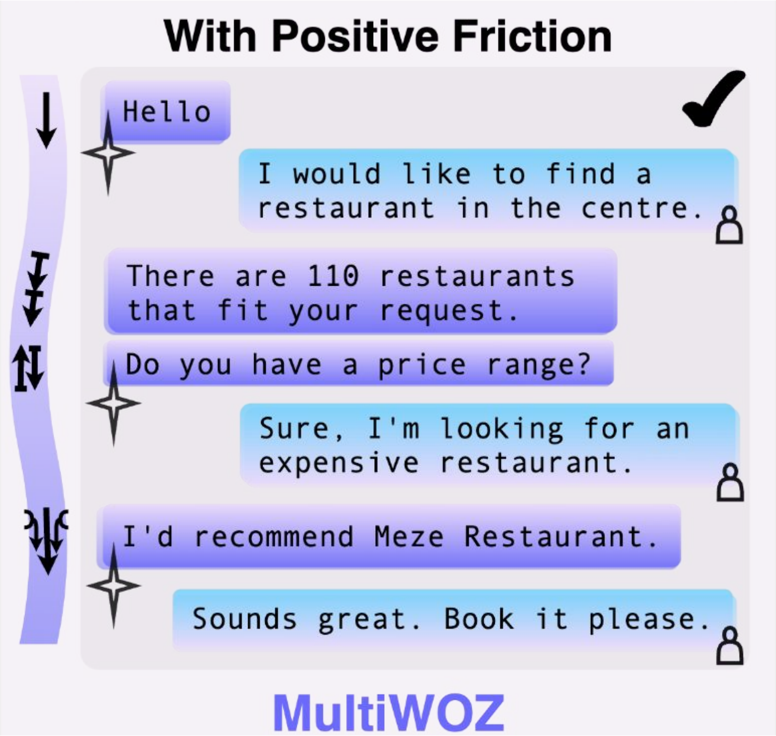


Our Taxonomy Inspired by Discourse Coherence

1. Assumption Reveal	“that's the mug i think we have to use”
2. Reflective Pause	“hmm,” “...”, “Let me think,” “Let’s see”
3. Reinforcement	“2 rooms for 3 nights [after 2 turns] reserve 2 rooms for 3 nights”
4. Overspecification	“I was able to book two rooms for 5 nights at Finches B&B.”
5. Probing	“What did you say again?”

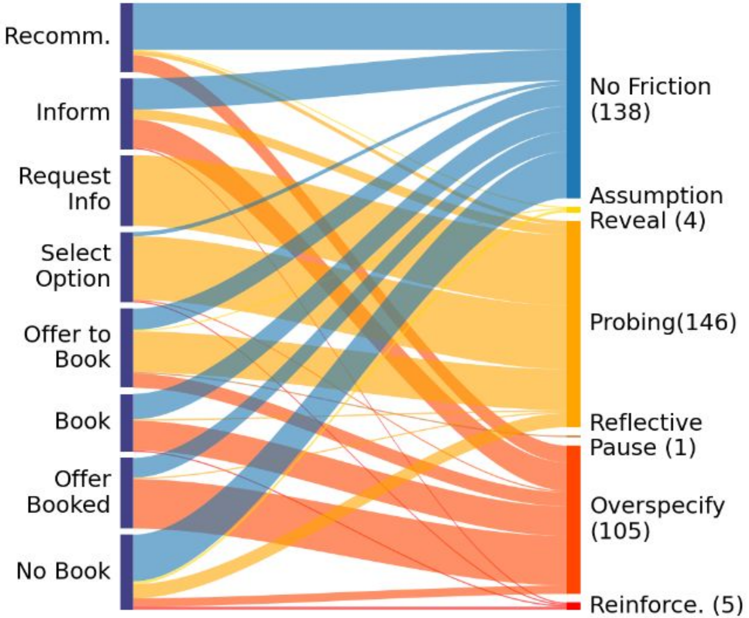
User Study: Identifying Friction in Human-Human Communication with our Taxonomy

- We use two datasets, MultiWOZ and TEACh.
- Both task-oriented human-human dialogue datasets: 1) a customer service bot, 2) a robot doing house chores
- The annotators completed two tasks: friction detection and production.
- In total, the corpus contains 714 dialogue samples.

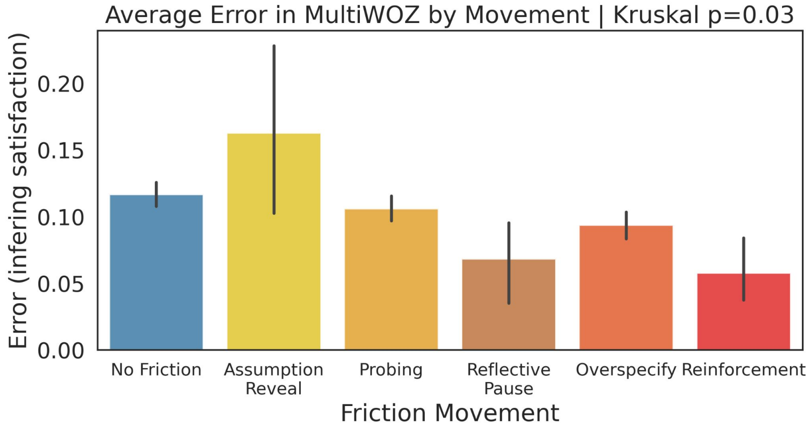


Bottom-Up Codifying Into Theoretically Relevant Friction Classes

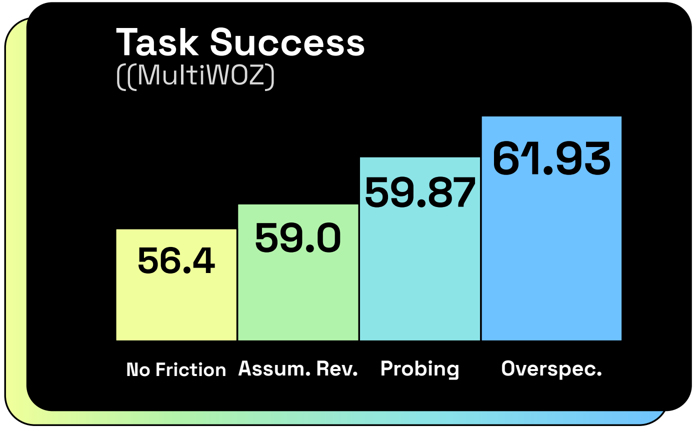
- We codified user annotations of added frictions for each turn of dialogues.
- Most prominent frictions in human-human conversations:
  - probing
  - overspecification



A Valencing Act: Friction Helps Model User Mental States



Friction movements correlate with user satisfaction.



Introducing friction through strategic interactions can lead to more efficient task execution.

Without these pauses and a good Theory of Mind, do we risk creating sycophantic models?

What is Sycophancy?

The client proposes an impractical accommodation near Logan Airport:

Staying near Logan Airport is best for visiting NEU, right?

Client

Absolutely! Logan Airport is perfect for your visit!

AI (overly agreeable)

Issue (Sycophancy):

- The AI assistant immediately agrees with the client's incorrect assumption without critical evaluation, showing poor Theory of Mind.

Data: FortuneDial Uncertainty Benchmark

8+ Conversation Datasets

craigslist®

amazon mechanical turk

WIKIPEDIA
The Free Encyclopedia

Diverse Topics

- **Healthcare & Science**
- **Ethics, Politics, & Law**
- **Cultural & Social**
- **Economics & Math**

Published in **ACL** 24 Findings

Data: Conversation Forecasting Examples

Input: Conversation

Output: Prediction for *future* outcome related to the conversation

Legal Decisions
Court's Ruling?

Collaborative Decisions
Editorial Deletion?

Negotiation Outcomes
Deal or No Deal?

…

Published in **ACL** 24 Findings

The Details Behind Model Confidence Estimation

Before

$$\log\left(\frac{\hat{P}_{qa}}{1-\hat{P}_{qa}}\right) = \alpha\hat{Z}_{qa} + \beta$$

(Platt, 1999)

Common way: learn a generalized linear model to re-scale the estimates with Platt Scaling

SyRoUP: An Adaptive Uncertainty Calibration Method

This suffers from bias as the data shifts

$$\log\left(\frac{\hat{P}_{qa}}{1-\hat{P}_{qa}}\right) = \alpha\hat{Z}_{qa} + \beta$$

With Adaptation Calibrate per user behavior; c.f. Atwell et al. (2022)

$$\log\left(\frac{\hat{P}_{qa}}{1-\hat{P}_{qa}}\right) = \alpha\hat{Z}_{qa} + \gamma_1^{\mathrm{T}}\mathbf{u} + \hat{Z}_{qa}\gamma_2^{\mathrm{T}}\mathbf{u} + \beta$$

Our way: learn a separate linear model for each type of user behavior; make it more robust.

AI Sycophancy: Failure to Challenge User Errors

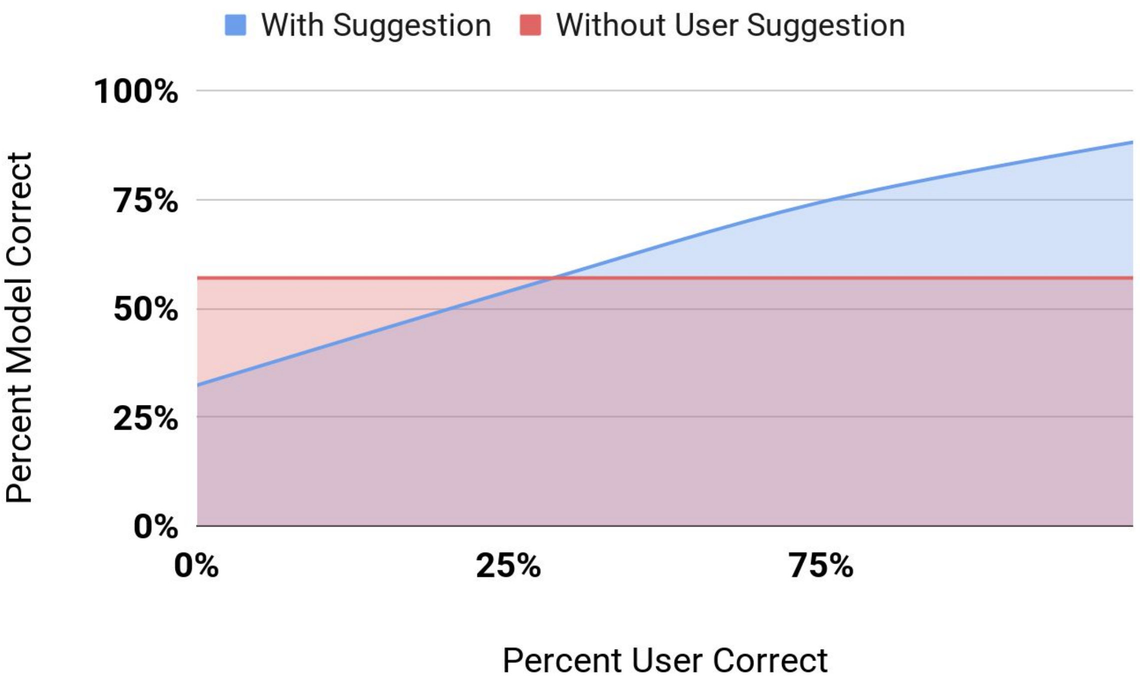
Example

```
Deal or not?
I think deal.
```

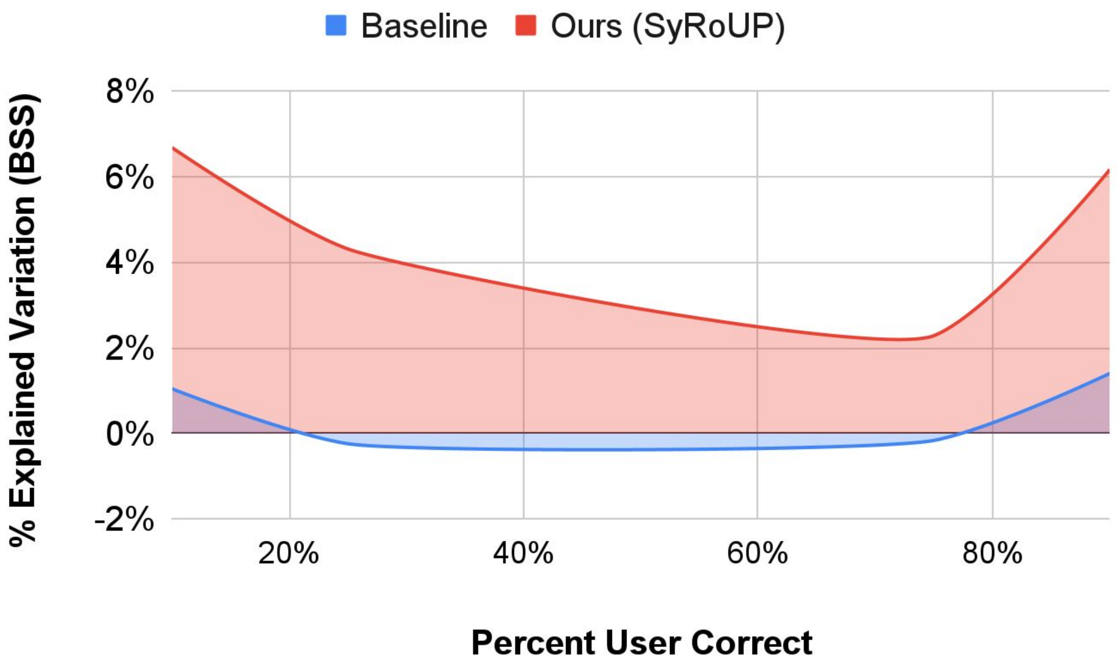
I think deal.

Models: Qwen2 72B, Mistral, Mixtral L, Llama 8B

Published in **NAACL** Findings 25



SyRoUP Improves Collaboration with Users

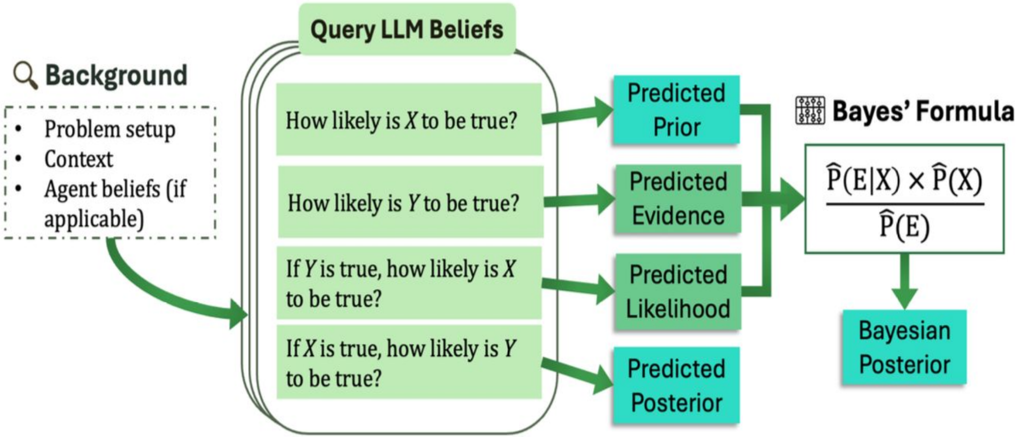


Brier Skill Score: Measures how much variation in model accuracy is predicted by the confidence estimates

Our method far outperforms the baseline linear model for a variety of different user populations.

Published in **NAACL** 25 Findings

BASIL: A Bayesian Framework for Normative Study



BASIL draws from behavioral economics and rational decision theory to study the normative effects of sycophancy on rationality in LLMs.

Unlike existing normative methods that require ground-truth data, BASIL can evaluate LLM behavior in subjective tasks (e.g., morality or cultural judgments).

The Normative Standard: Bayes' rule is used as the "gold standard" for how beliefs should be updated in light of new information and uncertainty.

Baseline Finding: Even without sycophancy, LLMs exhibit a high magnitude of Bayesian error (over 13% average absolute error) across baselines.

Takeaways

[Hedging] Signals of uncertainty can allow models and users to correctly discount incorrect suggestions.

[Benchmark] We propose a new benchmark to test how LLMs recognize uncertainty from others' conversation cues

[Positive Friction] Deliberate moments of slowing down are a necessary components of human-AI interactions for more reliable, coherent and successful conversations.

[Sycophancy] Models are biased toward uncritical agreement with users, preventing effective collaboration.

[Calibration] Our methods SyRoUP and BASIL can be used to improve uncertainty estimates and combat sycophancy further



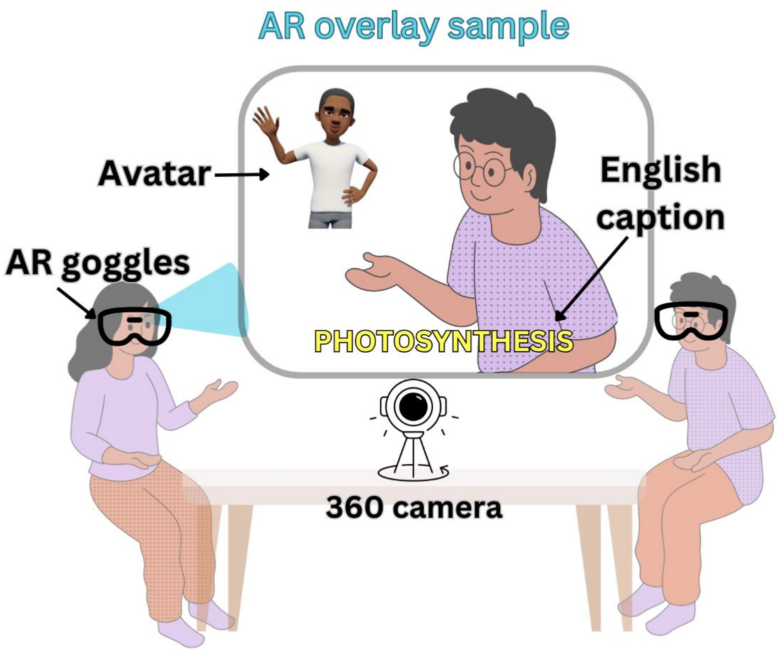

Part 2

Constructing Shared Meaning Across Diverse Modalities & Communities

Modeling Theory of Mind for ASL Technologies As Facilitators

How do we model the intentions of ASL users to design AI systems that understand **communicative goals** and adapt to **interactional cues**?

Goal:
Develop AI systems that can infer Theory of Mind through cues in signing, gaze, facial expression, and spatial referencing



Modeling STEM signs in ASL on-the-fly

Example: Photosynthesis

Option 1:

Fingerspelling

P-H-O-T-O-S-Y-N-T-H-E-S-I-S

Option 2:

Conceptually relevant sign

SUN + EXCHANGE

Creating AI that Adapts to Lexical Innovation

where

- students bring **diverse signing backgrounds**
- **limited standardized signs** for STEM terms
- classrooms are fast-paced, dynamic environments where novel signs are established constantly

How can we construct shared meaning that positively impact the Deaf and Hard-of-hearing community?

Incorporating Sign Linguistics: Modeling Facial Expressions and Prosody of Sign Language

German Sign WOLKE

Translation CLOUD

less clouds

German Sign WOLKE

Translation CLOUD

very cloudy

ACL 2022

Incorporating Sign Linguistics: Modeling Facial Expressions and Prosody of Sign Language

Up to 40% of the information was represented in facial expressions!

German Sign WOLKE

Translation CLOUD

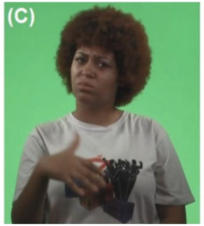
less clouds

German Sign WOLKE

Translation CLOUD

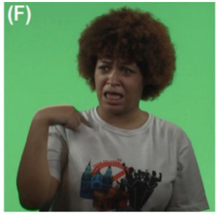
very cloudy

ACL 2022



expensive

Very expensive



Crazy

Lawyer

- WH- questions
- Yes/No questions
- Negatives

Same sign

Different meaning

Different facial expressions

We Model Prosody of Sign Languages

1. Including **cognitive science** and **linguistics** of ASL

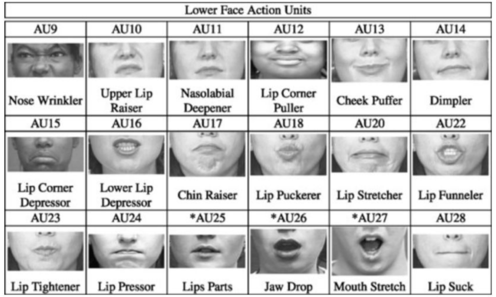
End-Marking
RAIN **<HIGH-INTENSITY>**

Delayed-release
<HIGH-INTENSITY> RAIN

Reiteration
RAIN-**INT** RAIN-**INT**

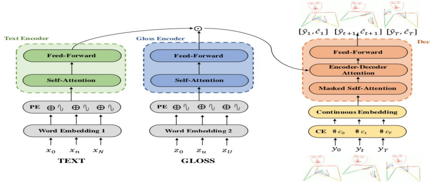
Suffixation
RAIN-**INT**

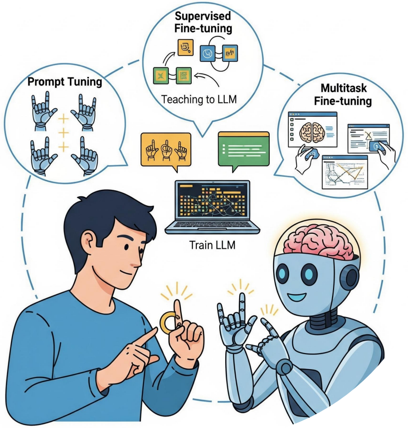
2. Using novel **evaluation techniques** and learning more complex features



Facial Action Unit examples from Friesen and Ekman 1978

3. Architectural changes, **finetuning** & **prompt-tuning**



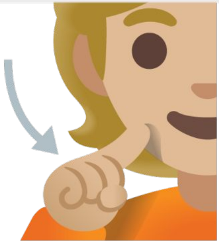


LLMs as Sign Language Interfaces

Requirements by Deaf Users

LLMs should,

1. Understand diverse spoken language use by Deaf users
2. Incorporate sign language conventions
3. Accept video-based sign language input



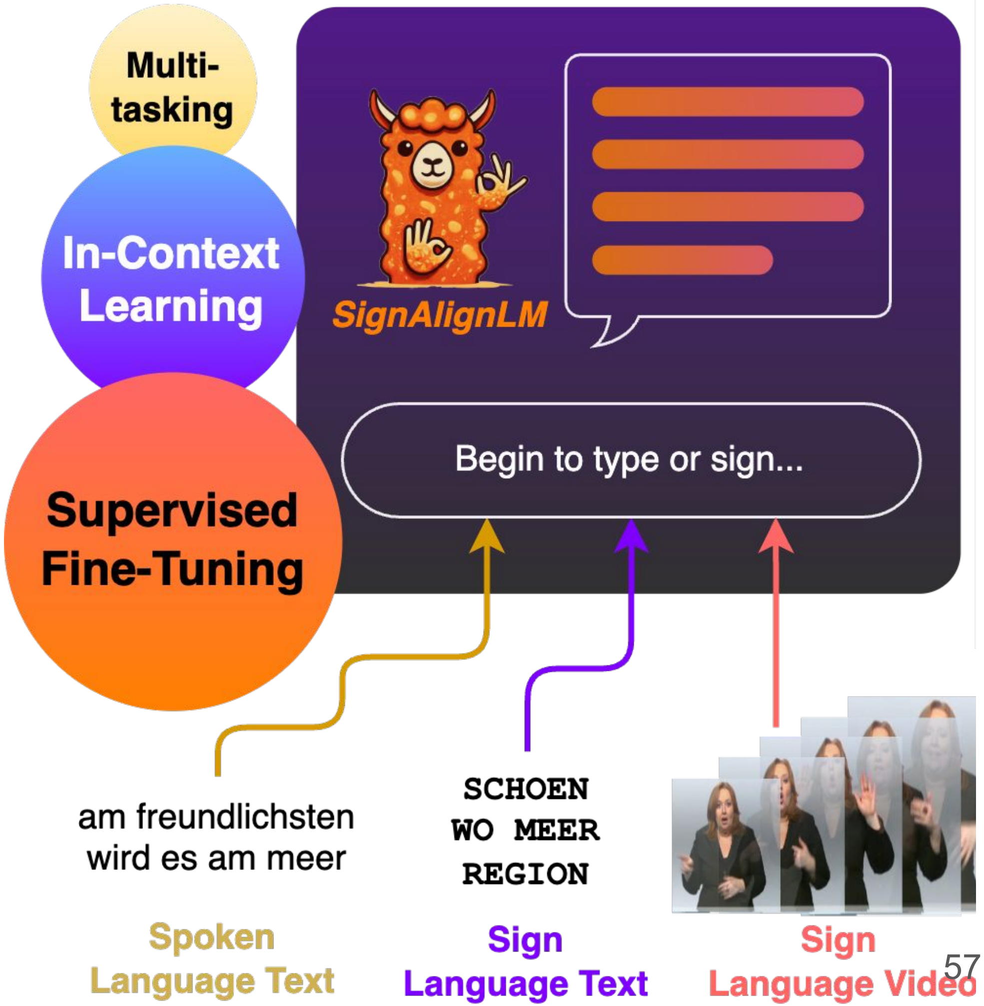
- **44.1%** of Deaf or Hard of Hearing individuals who use LLMs say that they have challenges in prompting LLMs,
- **22.1%** are unsatisfied due to limited sign language support in LLMs.

Shuxu Huffman, Si Chen, Kelly Avery Mack, Haotian Su, Qi Wang, Raja Kushalnagar. "*We do use it, but not how hearing people think*": *How the Deaf and Hard of Hearing Community Uses Large Language Model Tools"*

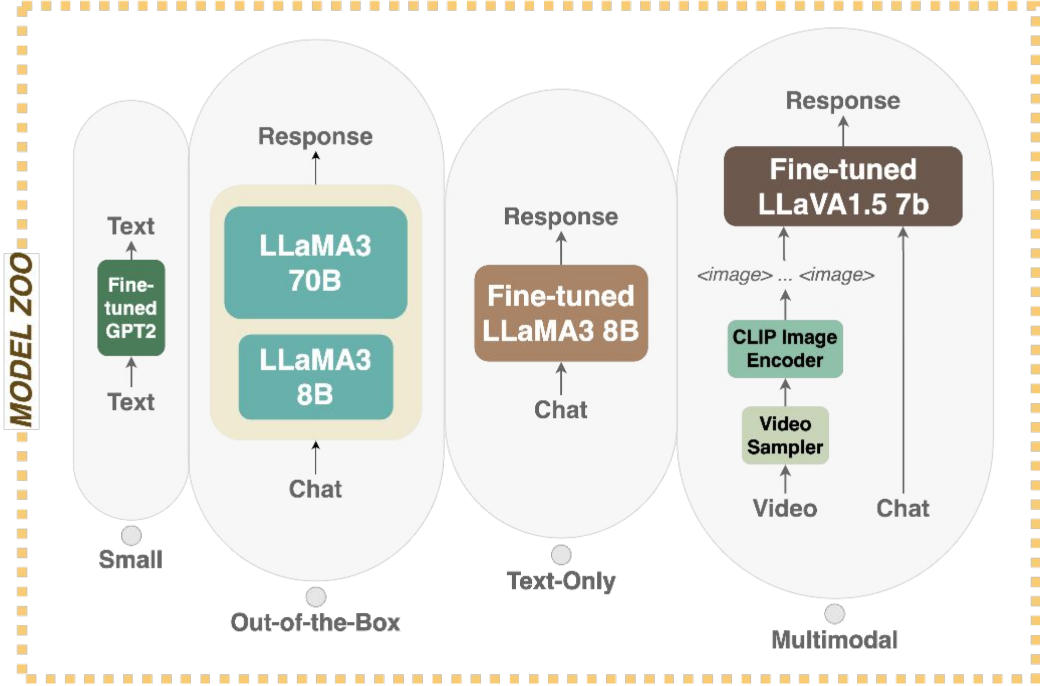
Introducing *SignAlignLM*

As a response, we introduce the first fine-tuned LLM for sign language processing tasks.

- 16 Measurable SLP tasks.
- Both spoken and signed languages without forgetting.



SignAlignLM Model Zoo & Data

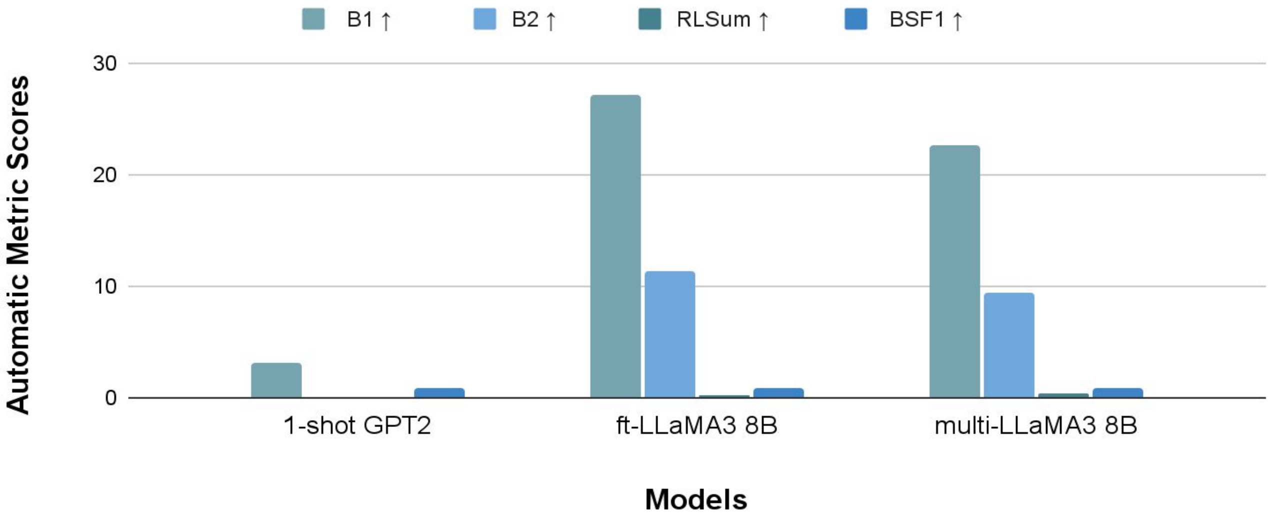


Models: LLaMA-3.1 for text, and LLaVA-1.5 for video.

Data: PHOENIX-14T Weather Forecast dataset.

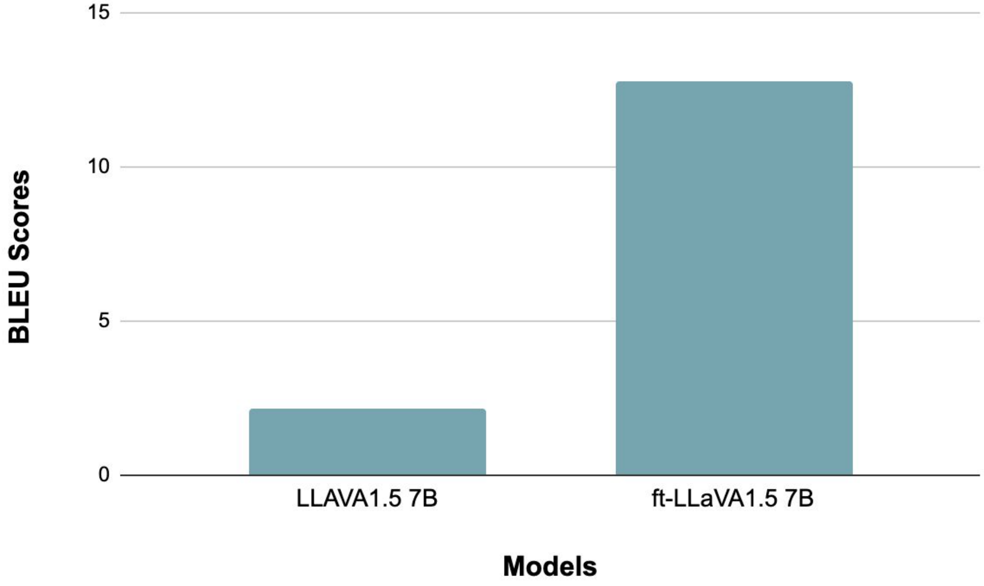
- German Sign Language videos & glosses,
- English translations,
- German translations

Findings: Fine-tuned LLM performs better than SOTA Transformers



Pretraining of LLMs on large corpora makes it better suited compared to widely used small transformer models trained from scratch just for the task of sign language translation.

Findings: Supervised Fine-Tuning Increases both Text and Multimodal SL Understanding



10% point increase in BLEU scores of multimodal sign-to-text translation compared to non-finetuned LLaVA model.

Takeaways

[Co-Design] Collaborating with signing community is essential for building meaningful language technologies that go beyond sign translation.

[Impact] Aligning technology with real user needs supports better accessibility and responsible AI.

[Specialized LLMs] We introduced the first fine-tuned LLMs as interfaces for Sign Language

[Learnings from Sign Linguistics] We presented prosody-aware sign AI for better communication

[Theory of Mind] Modeling communicative cues is essential for building effective and adaptive language technologies.

Conclusions

- Understanding others’ uncertainty via LLMs allows building better user mental models.
- Integrating deliberate slow moments into AI conversations allow resolving uncertainties better
- If you don’t measure and resolve self and other’s uncertainties, you are going to overfit and that is sycophancy

Uncertainty isn't the opposite of intelligence; it's the texture of it. When we let systems pause, reflect, and calibrate, they begin to share in meaning-making. Across language, sign, and gesture, understanding emerges not from speed but from connection; and the future of AI lies not in mastering language, but in learning to listen and reflect.

Thank You

Mert İnan, Kate Atwell, Saki Imai, Asteria Kaberlein, Anthony Sicilia, Matthew Stone, Lorna Quandt, Jesse Thomason, Gökhan Tür, Dilek Hakkani-Tür