COLORADO STATE UNIVERSITY

Faster Isn’t Always Better: Building Reliable and Accountable AI Collaborators in the Age of LLMs

Nikhil Krishnaswamy, Colorado State University

CLASP Seminar, Gothenburg, Sweden

June 25, 2025

Background

The Promise(s) of AI

"[AI could] do anything you'd be happy with a remote coworker doing"

2022, predicting massive job losses

"[Digital superintelligence is] much less weird than it seems like it should be"

2025, predicting the "gentle singularity"

Large Language Models, Small Labor Market Effects

University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2025-56

117 Pages • Posted: 17 Apr 2025 • Last revised: 13 May 2025

Anders Humlum
University of Chicago - Booth School of Business

Emilie Vestergaard
University of Copenhagen, Department of Economics, Students

There are 2 versions of this paper

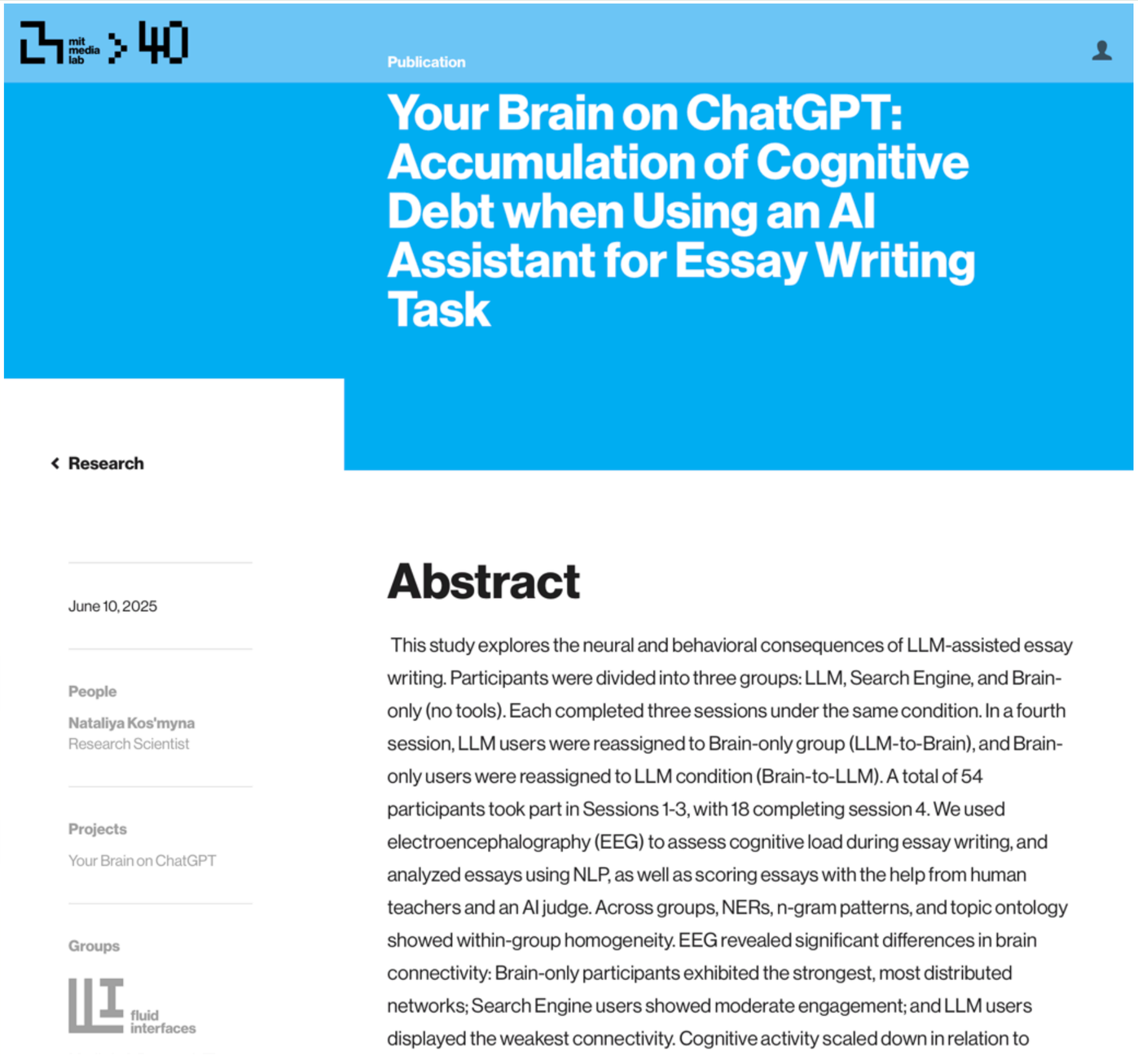
Date Written: April 15, 2025

Abstract

We examine the labor market effects of AI chatbots using two large-scale adoption surveys (late 2023 and 2024) covering 11 exposed occupations (25,000 workers, 7,000 workplaces), linked to matched employer-employee data in Denmark. AI chatbots are now widespread—most employers encourage their use, many deploy in-house models, and training initiatives are common. These firm-led investments boost adoption, narrow demographic gaps in take-up, enhance workplace utility, and create new job tasks. Yet, despite substantial investments, economic impacts remain minimal. Using difference-in-differences and employer policies as quasi-experimental variation, we estimate precise zeros: AI chatbots have had no significant impact on earnings or recorded hours in any occupation, with confidence intervals ruling out effects larger than 1%. Modest productivity gains (average time savings of 3%), combined with weak wage pass-through, help explain these limited labor market effects. Our findings challenge narratives of imminent labor market transformation due to Generative AI.

What changed?

- **Rapid adoption of generative AI in the "real world"**
- Sold as instantaneous "upskilling" of labor
- Incentivized for seamless use and integration



mit media lab > 40

Publication

Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task

‹ Research

June 10, 2025

People

Nataliya Kos'myna
Research Scientist

Projects

Your Brain on ChatGPT

Groups

fluid interfaces

Abstract

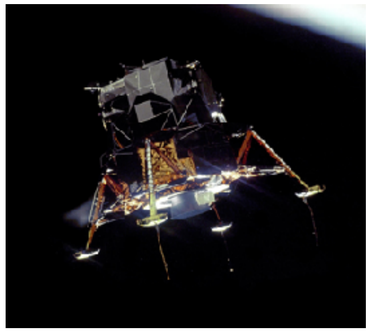
This study explores the neural and behavioral consequences of LLM-assisted essay writing. Participants were divided into three groups: LLM, Search Engine, and Brain-only (no tools). Each completed three sessions under the same condition. In a fourth session, LLM users were reassigned to Brain-only group (LLM-to-Brain), and Brain-only users were reassigned to LLM condition (Brain-to-LLM). A total of 54 participants took part in Sessions 1-3, with 18 completing session 4. We used electroencephalography (EEG) to assess cognitive load during essay writing, and analyzed essays using NLP, as well as scoring essays with the help from human teachers and an AI judge. Across groups, NERs, n-gram patterns, and topic ontology showed within-group homogeneity. EEG revealed significant differences in brain connectivity: Brain-only participants exhibited the strongest, most distributed networks; Search Engine users showed moderate engagement; and LLM users displayed the weakest connectivity. Cognitive activity scaled down in relation to

What changed?

- **MIT Study on “Cognitive atrophy”**
- Brain connectivity data suggests “overreliance” affects retention and production abilities
- If chatbots aren’t taking people’s jobs, what are they doing?

To Collaborate Is Human

- **Humans are a fundamentally collaborative species**
- The human capacity to resolve ambiguity and conflicting assumptions is key to our ability to work together toward shared goals

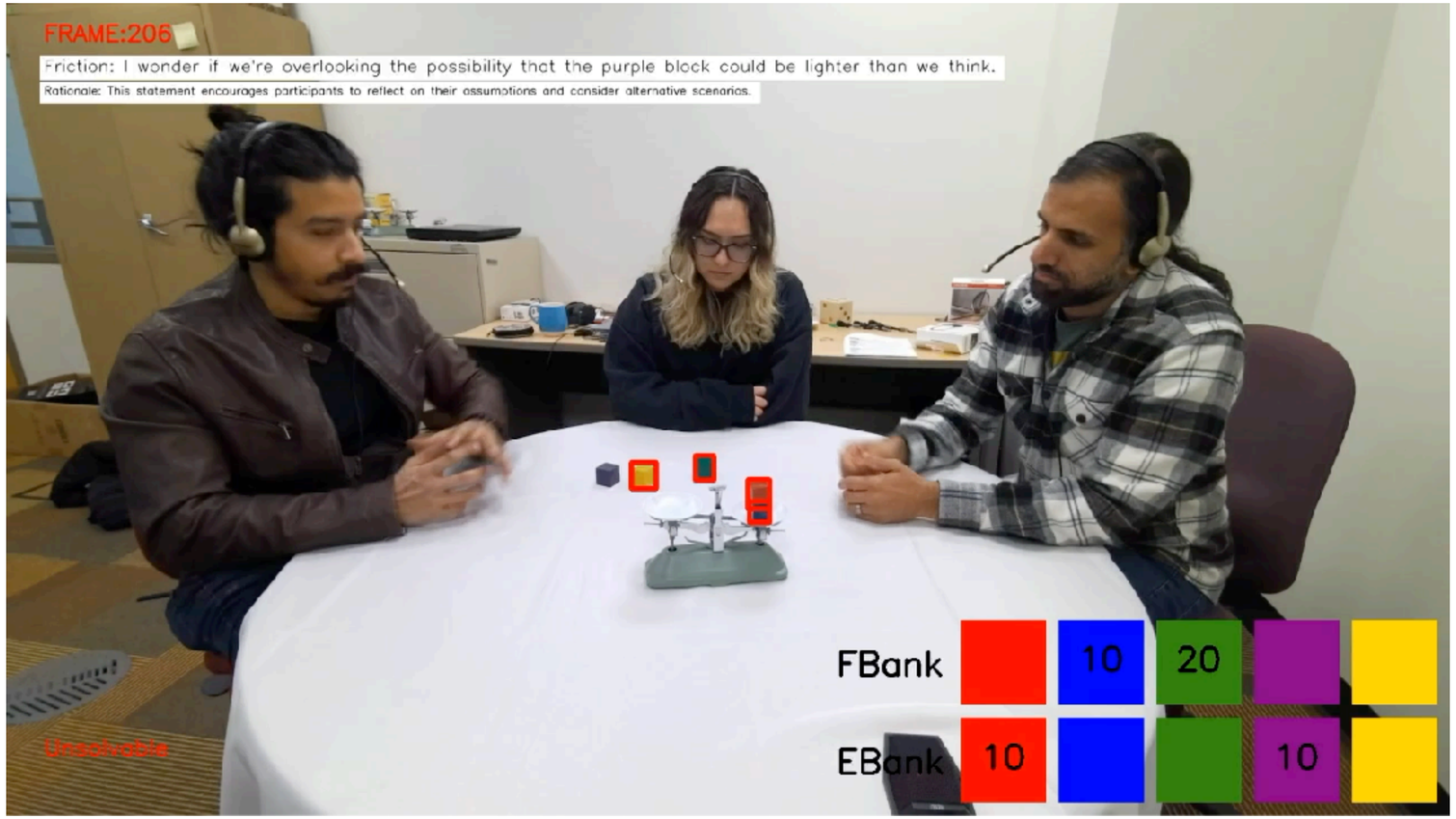


- People imbue language generation systems with other cognitive characteristics
- People are now using LLMs as “collaborators”

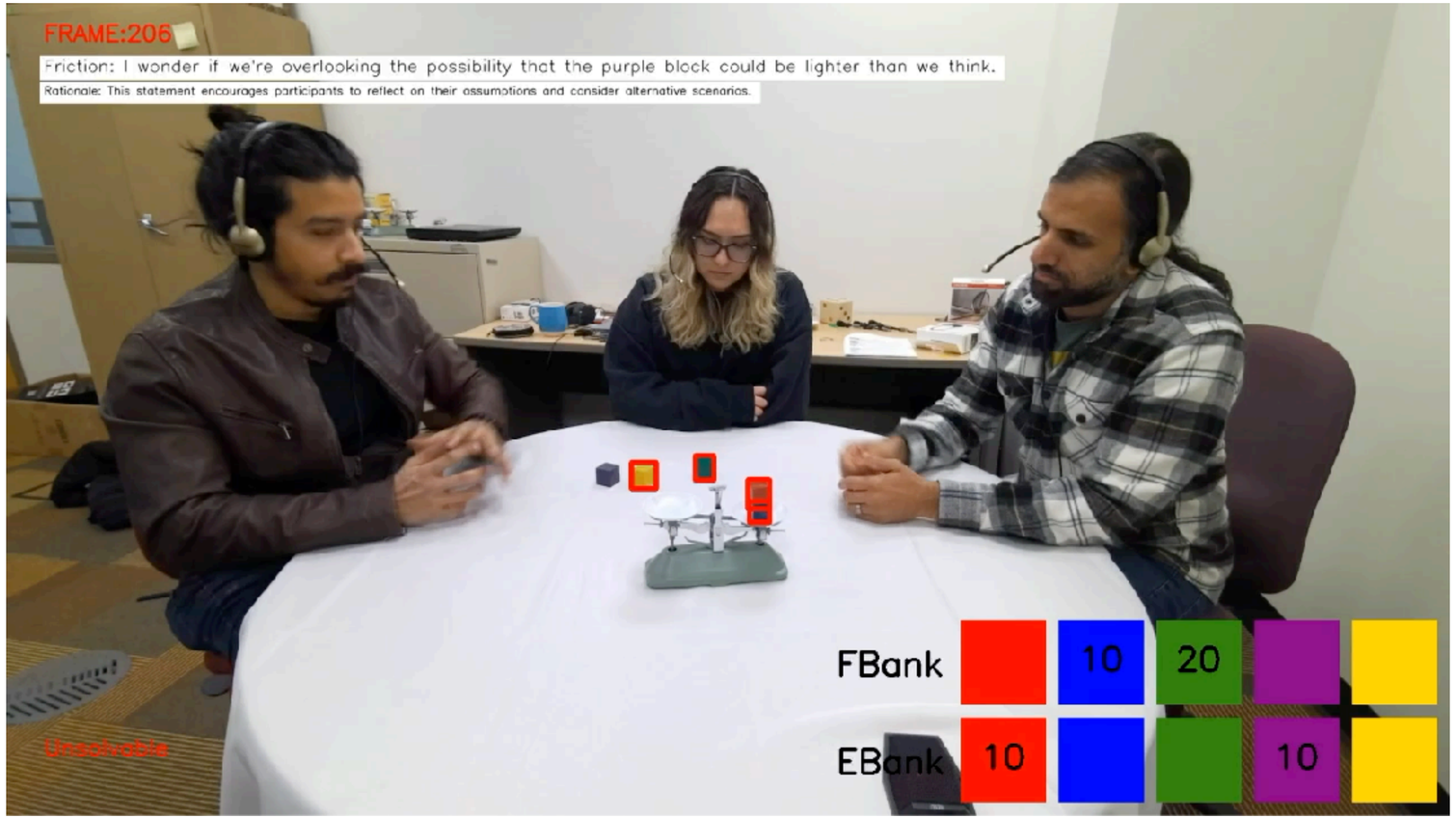
What Makes a Good Collaborator?

- Negotiating intents toward a shared goal is a hallmark of human intelligence predicated upon *theory of mind*
  - The attribution of mental states to others, to predict and explain behavior
- Predict behavior → anticipate needs → **reliable collaboration**
- Explain behavior → assign responsibility → **accountable collaboration**
- AI has been sold as as way to reduce workload and increase speed and efficiency
- Non-experts may **overrely** on AI outputs
  - Consequences may be inconvenient to catastrophic

Collaborative Tasks

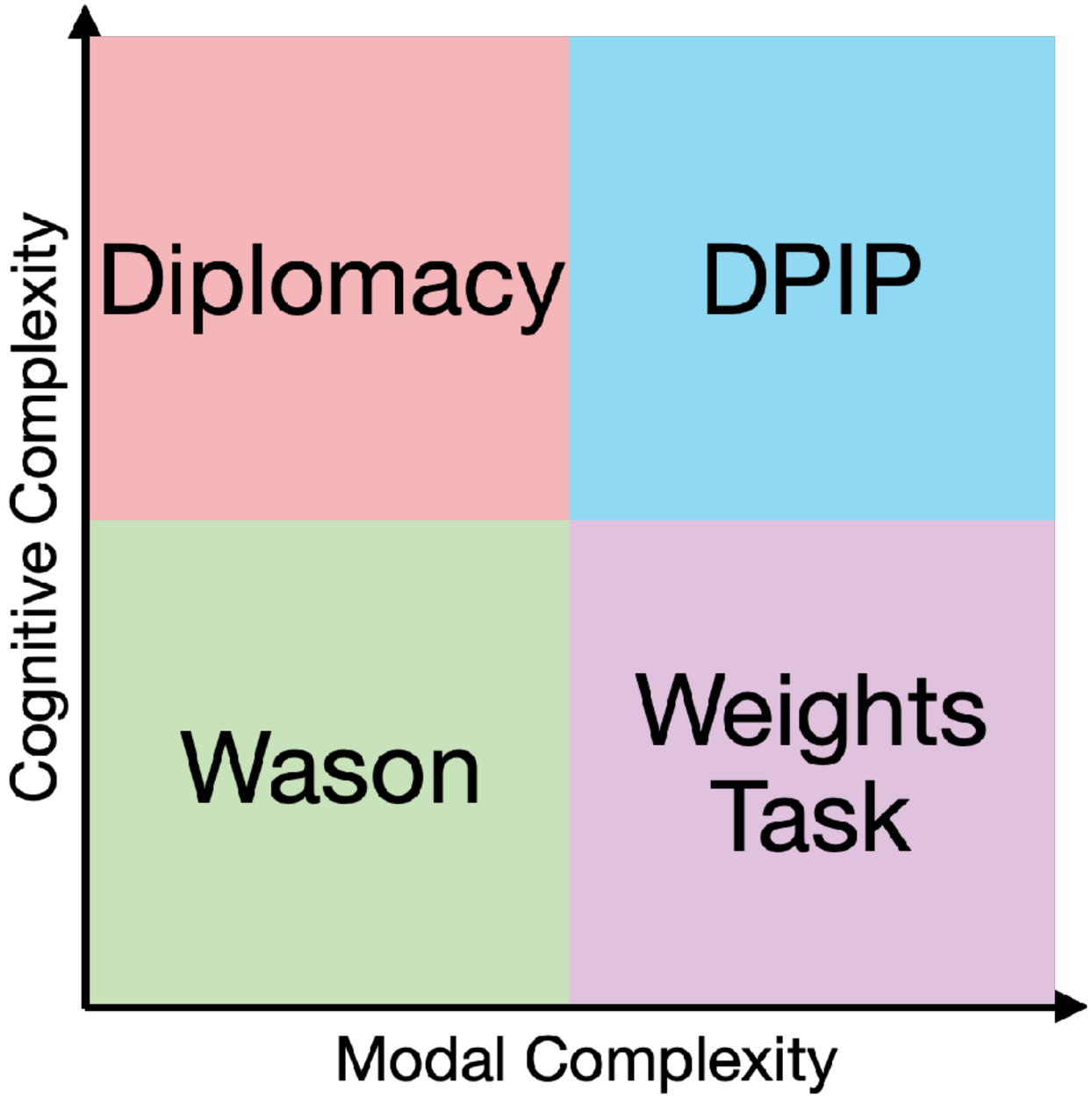


Collaborative Tasks



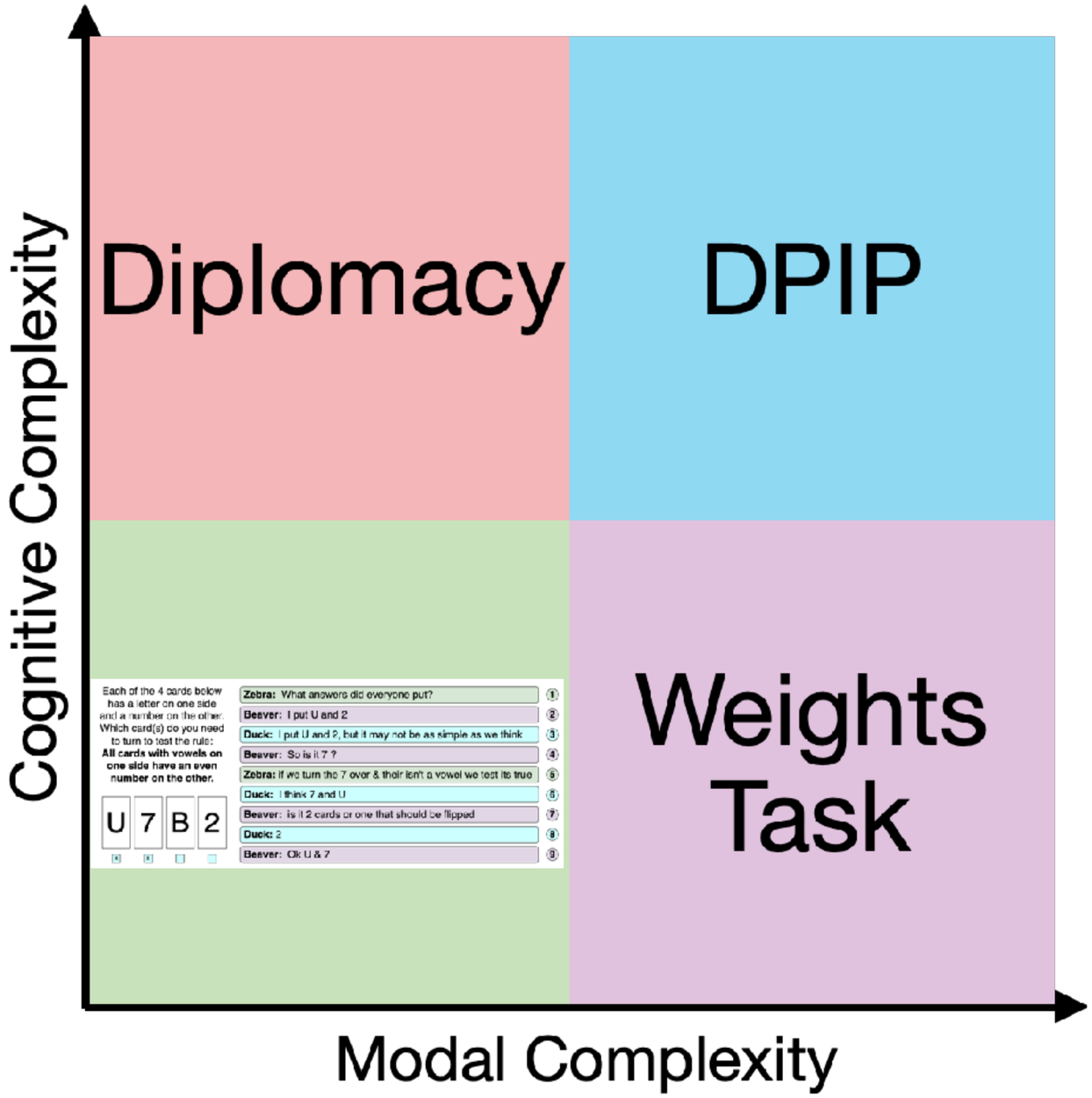
Collaborative Tasks

- **Collaboration:**
  - Helps with learning
  - Leads to generally better outcomes
  - May be fully- or partially-observed
  - May include elements of competition



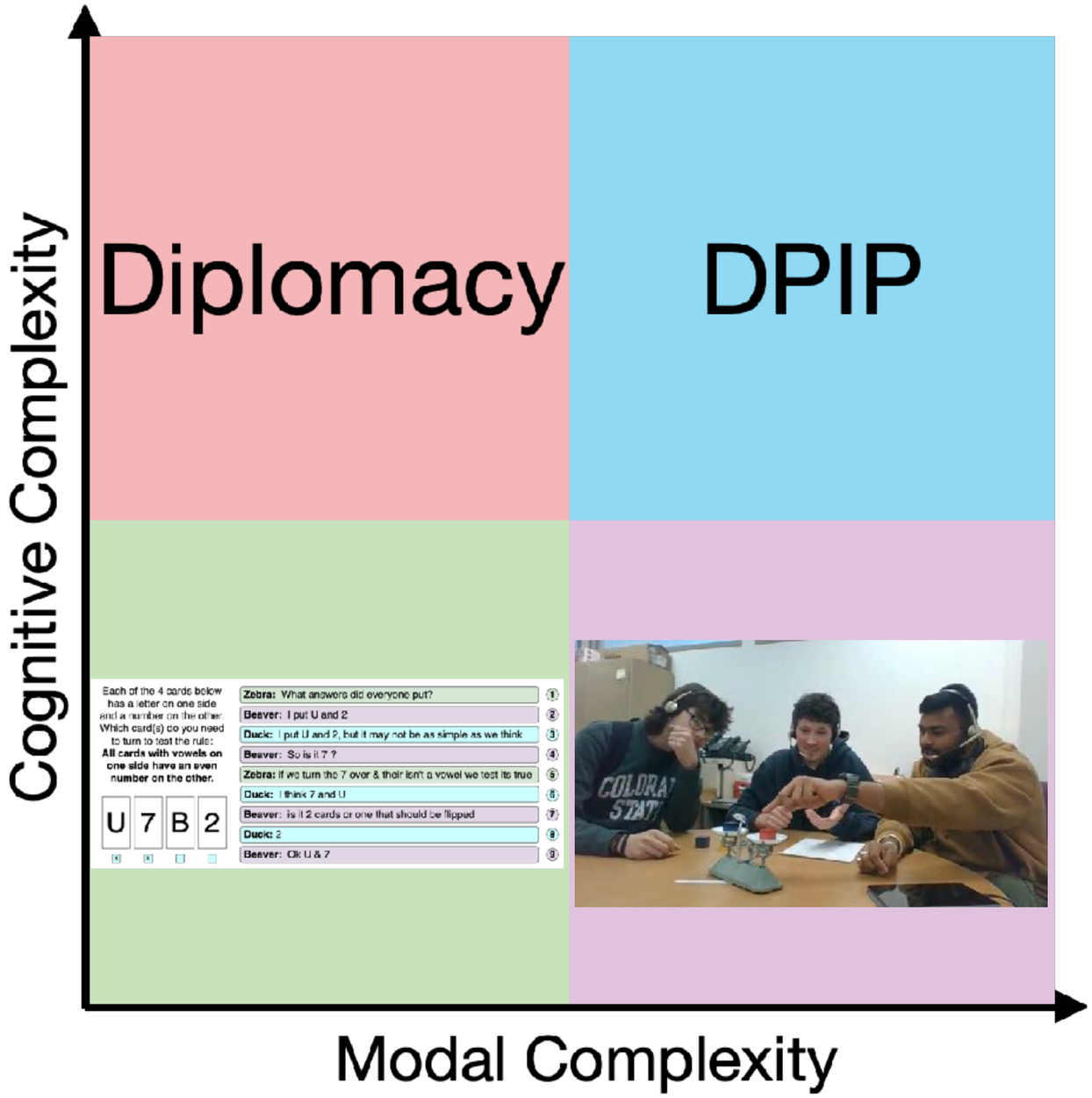
Collaborative Tasks

- **Collaboration:**
  - Helps with learning
  - Leads to generally better outcomes
  - May be fully- or partially-observed
  - May include elements of competition



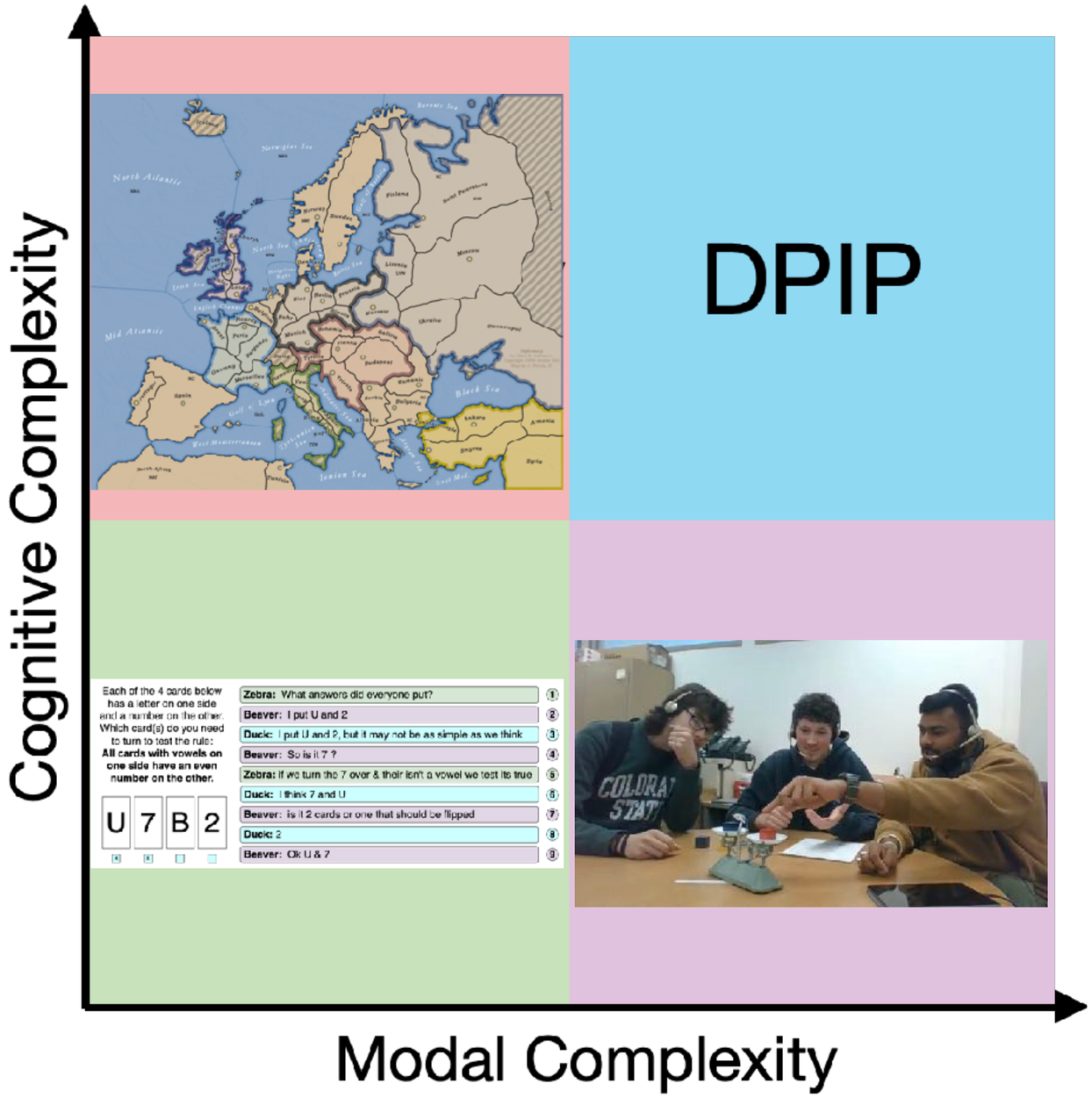
Collaborative Tasks

- **Collaboration:**
  - Helps with learning
  - Leads to generally better outcomes
  - May be fully- or partially-observed
  - May include elements of competition



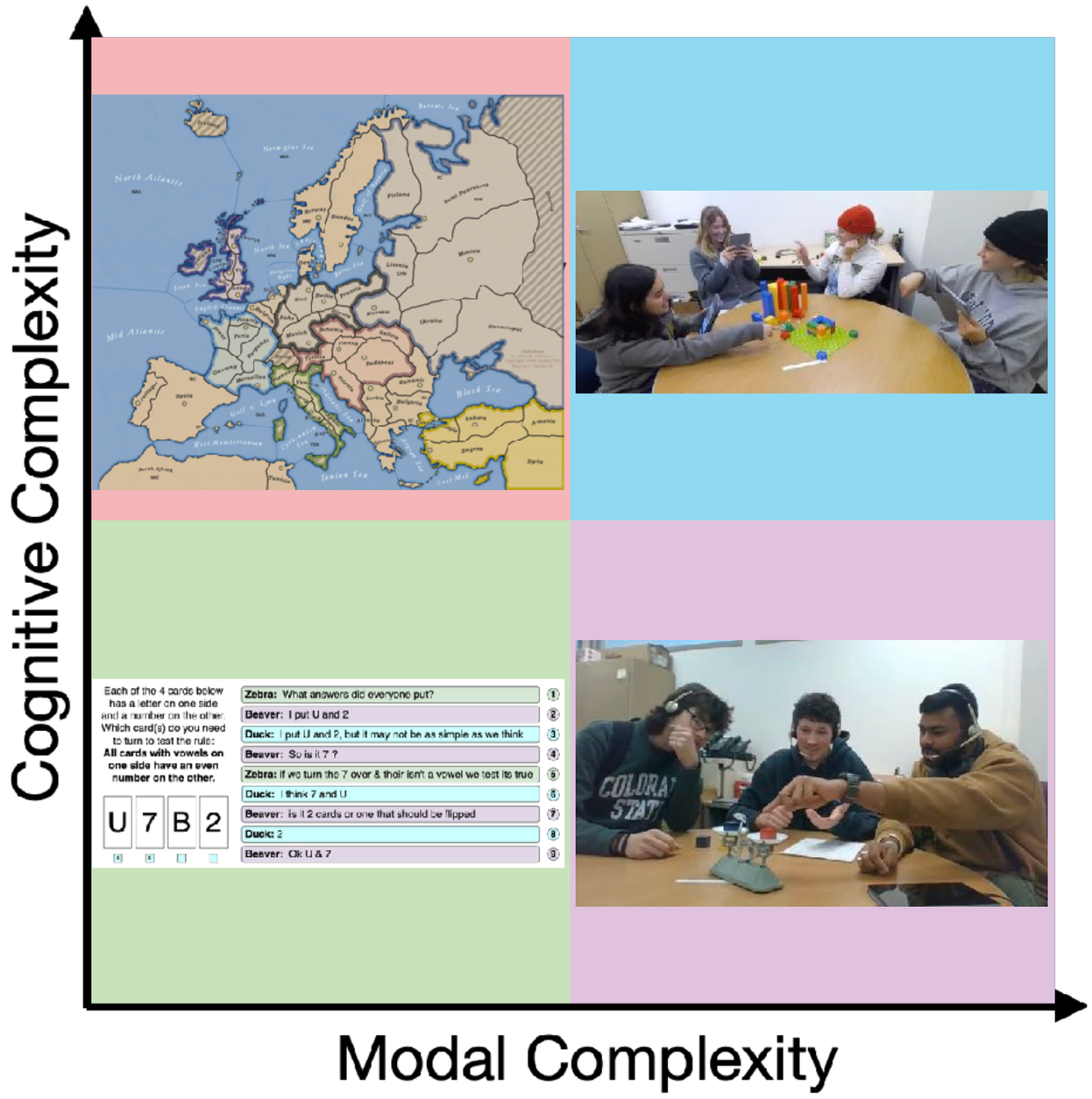
Collaborative Tasks

- **Collaboration:**
  - Helps with learning
  - Leads to generally better outcomes
  - May be fully- or partially-observed
  - May include elements of competition



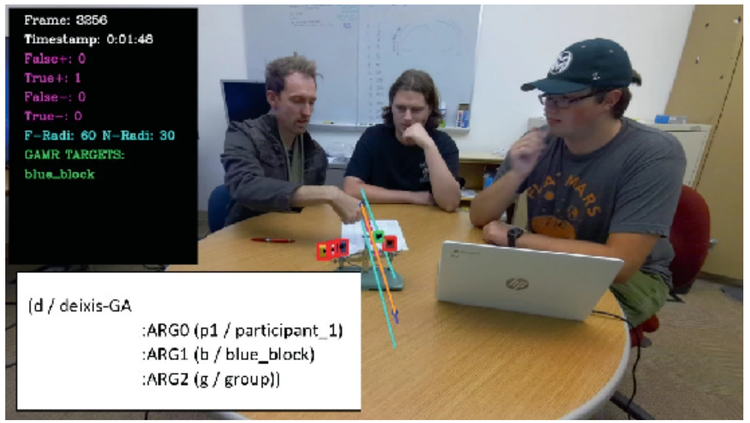
Collaborative Tasks

- **Collaboration:**
  - Helps with learning
  - Leads to generally better outcomes
  - May be fully- or partially-observed
  - May include elements of competition

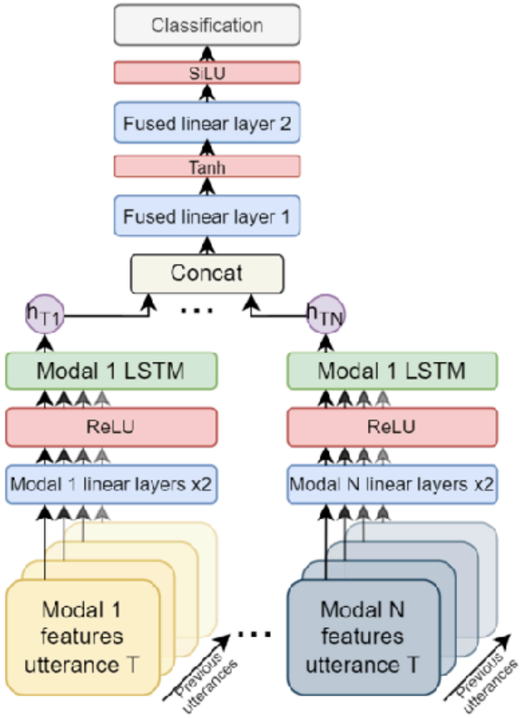


Common Ground Tracking in Multimodal Dialogue

- Encode information from multimodal channels:
  - Communicative expressions (speech, gesture)
  - Jointly-perceived actions
  - Nonverbal behaviors (gaze, pose facial expressions)
- Identify intentions, goals, and attitudes of team members
- Track shared knowledge about tasks and goals
- Update evidence and beliefs from actions in context
- "Banks" of questions under discussion (QUDs), evidence, and agreed-upon facts
- Closure rules move propositions between banks
- Different modalities contribute different information to both proposition and epistemic positioning



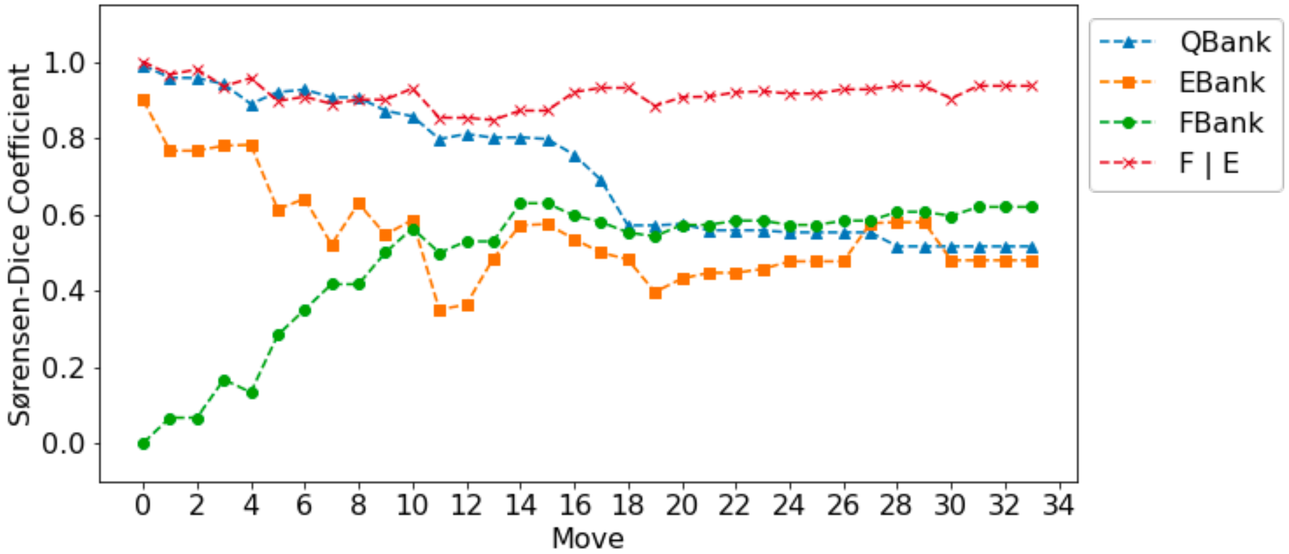
Deictic gesture and equivalent GAMR



Epistemic move classifier architecture

Common Ground Tracking in Multimodal Dialogue

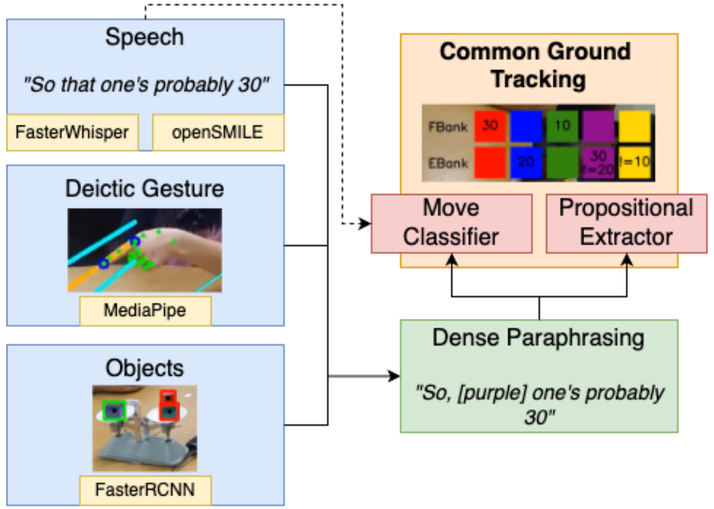
- Evidence may be contributed through multiple modalities
  - "This one's 30" (STATEMENT) vs. "This one's 30?" (DOUBT)
  - Classify speaker's epistemic positioning based on:
    - Language (BERT)
    - Prosody (openSMILE)
    - Collaborative Problem Solving facets (Learning Sciences framework)
    - Action (VoxML)
    - Gesture (Gesture AMR)



- I. Khebour, K. Lai, M. Bradford, Y. Zhu, R. Brutti, C. Tam, J. Tu, B. Ibarra, N. Blanchard, N. Krishnaswamy, and J. Pustejovsky. 2024. *Common Ground Tracking in Multimodal Dialogue*. In Proceedings of LREC-COLING 2024.

Common Ground Tracking in Multimodal Dialogue

- Previous work performed in an offline condition (standard train/test setting)
- Key challenge: to build a real-time system that enables tracking shared beliefs
- Speech transcriptions, deictic gesture, and detected objects are aligned for real-time multimodal dense paraphrasing
- Signals within each utterance span input to move classifier and propositional extractor
- Closure rules populate common ground
- Extensible, dependency graph-based architecture facilitates additional modules



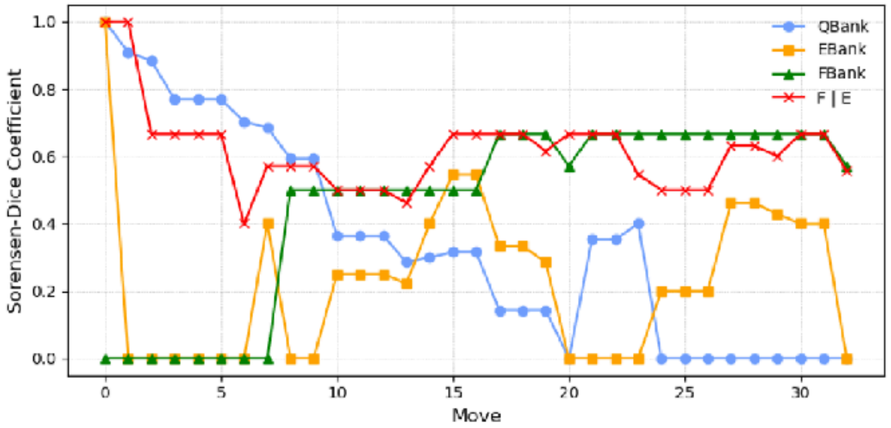
- H. VanderHoeven, B. Bhalla, A. Youngren, V. Venkatesha, I. Khebour, M. Bradford, J. Fitzgerald, C. Mabrey, J. Tu, Y. Zhu, K. Lai, J. Pustejovsky and N. Krishnaswamy. *Real-Time Multimodal Common Ground Tracking in Situated Collaborative Dialogues.* In Proceedings of NAACL 2025: System Demonstrations.

Common Ground Tracking in Multimodal Dialogue

- Performance degradation is expected in a live condition, but how much?
- Live performance displays characteristic patterns in common ground
  - QBank empties, jump in FBank score toward end, bump in EBank in the middle
- Ablation tests reveal where model errors lead to downstream failures
- Small improvements in ASR, gesture detection, object tracking improve overall performance

	Group 1	Group 2	Group 4	Group 5
		TRACE		
QBank	0.349	0.656	0.741	0.546
EBank	0.063	0.135	0.231	0.214
FBank	0.000	0.205	0.000	0.000
F ∪ E	0.246	0.377	0.231	0.464
		Khebour et al. (2024b)		
QBank	0.767	0.911	0.817	0.514
EBank	0.344	0.713	0.812	0.335
FBank	0.000	0.528	0.045	0.165
F ∪ E	1.000	0.922	0.832	0.959

Average DSC of 4 test groups compared to original CGT paper

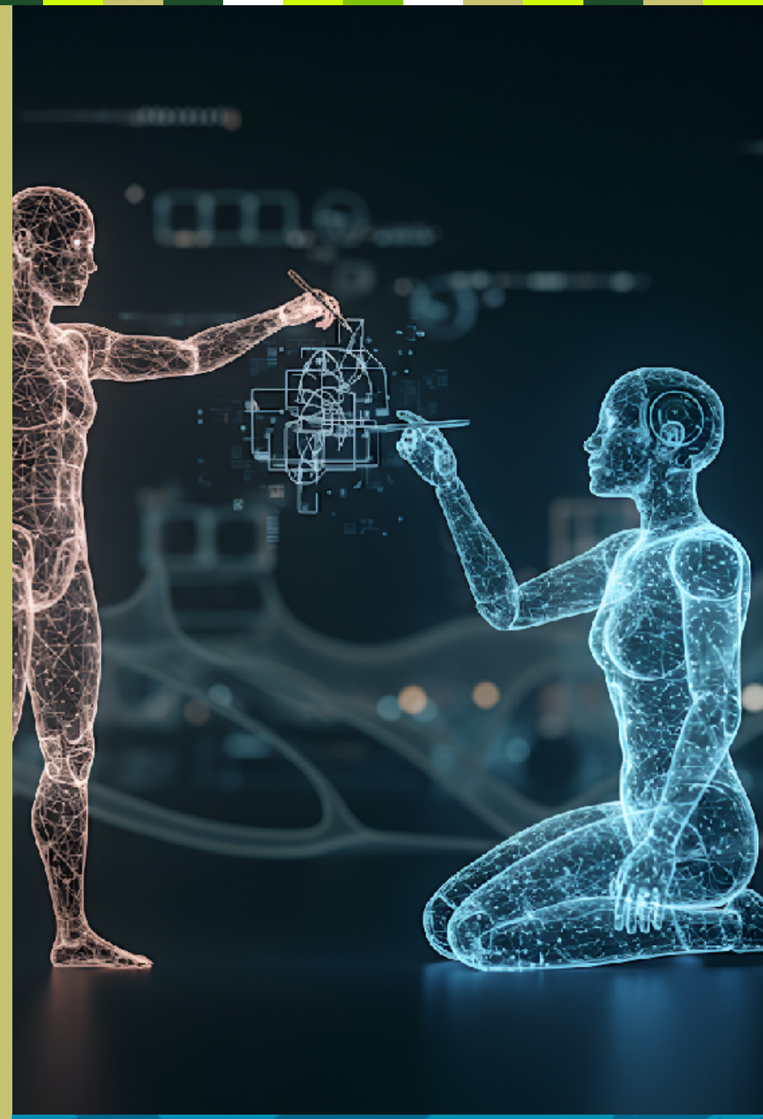


DSC over time displays characteristic pattern

	Ground truth utterances				Ground truth gestures				Ground truth objects			
	Group 1	**Group 2**	**Group 4**	**Group 5**	**Group 1**	**Group 2**	**Group 4**	**Group 5**	**Group 1**	**Group 2**	**Group 4**	**Group 5**
QBank	0.423	0.498	0.714	0.549	0.343	0.634	0.783	0.570	0.351	0.657	0.762	0.554
EBank	0.031	0.042	0.248	0.263	0.050	0.147	0.280	0.290	0.067	0.135	0.231	0.247
FBank	0.054	0.183	0.247	0.000	0.053	0.202	0.000	0.000	0.204	0.228	0.000	0.000
F ∪ E	0.383	0.324	0.419	0.555	0.384	0.377	0.368	0.608	0.220	0.405	0.255	0.508

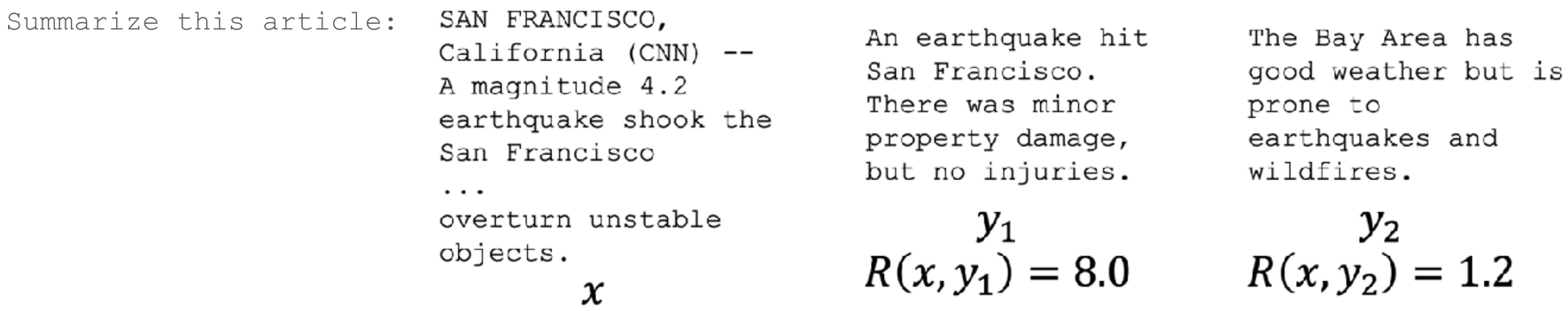
Ablation testing using ground truth utterances, gestures, and object detections

LLMs and Theory of Mind



How LLMs Are Trained

- **Generative LLMs are trained to simultaneously become iteratively better predictors of both next tokens and overall responses**
- Applies Bellman optimality guarantees of token MDP to full responses $\hat{y}$
- Or, maximize distance between "winning" label y_w and "losing" label(s) y_ℓ



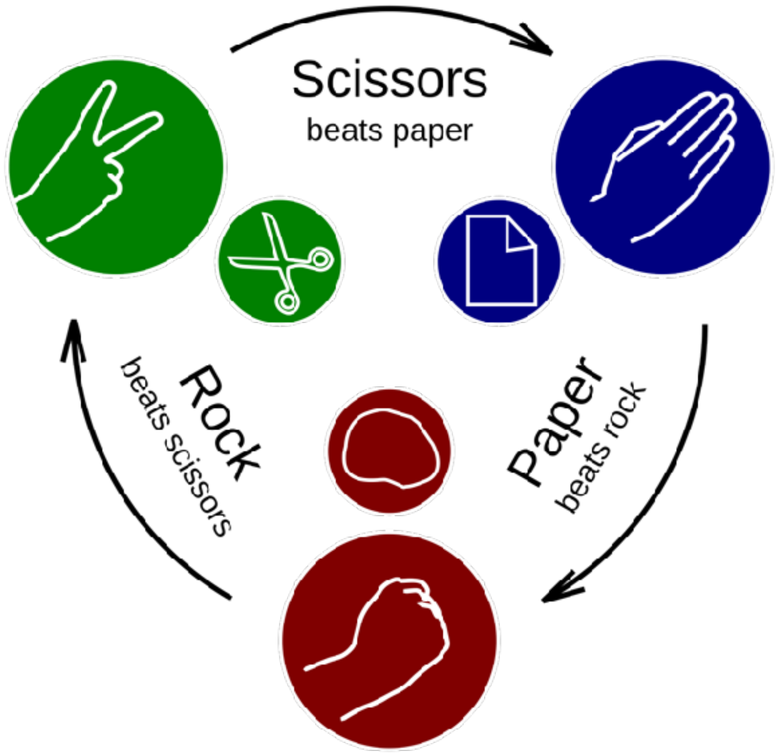
- **If y_A is better than y_B, and y_B is better than y_C, then y_A must be better than y_C**

LLMs and Theory of Mind

- If $y_A \succ y_B$ and $y_B \succ y_C$, then necessarily $y_A \succ y_C$

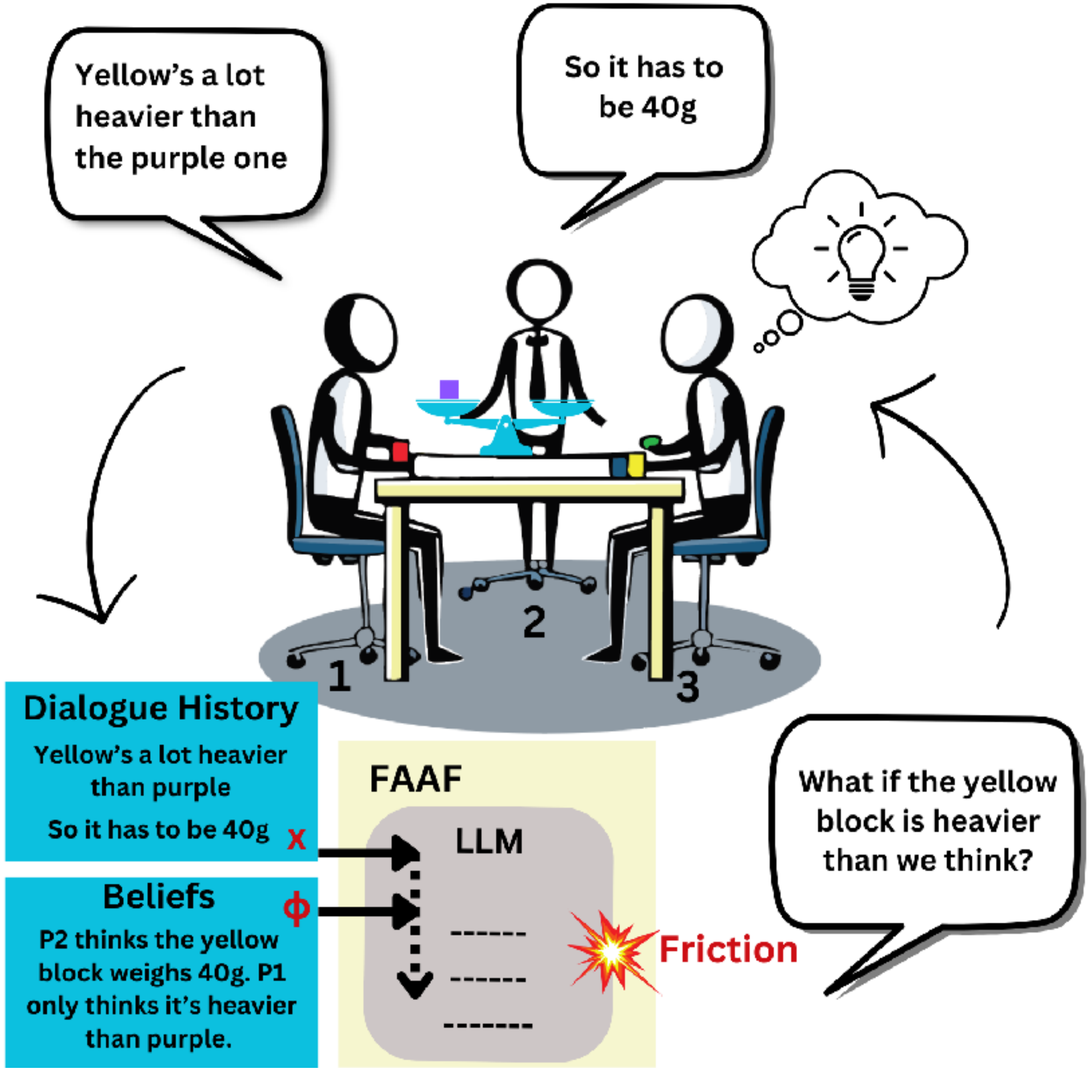
LLMs and Theory of Mind

- If $y_A \succ y_B$ and $y_B \succ y_C$, then necessarily $y_A \succ y_C$



LLMs and Theory of Mind

- Collaborators maintain **evidence models** consisting of their own beliefs and models of others'
  - Evidence models change over time, for reasons not apparent through surface level text
  - e.g., I believe something, you do something to change my belief
- Most LLM alignment methods are trained over pairwise "preference" comparisons
- Evidence model becomes **obscured variable**
- Obscured evidence model leads to non-deterministic preferences: $y_A \succ y_B \succ y_C \succ y_A$

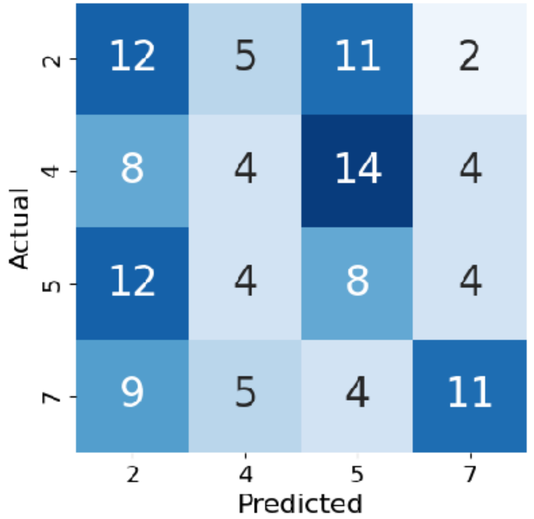


The Weights Task: triads collaborate to deduce the weights of blocks

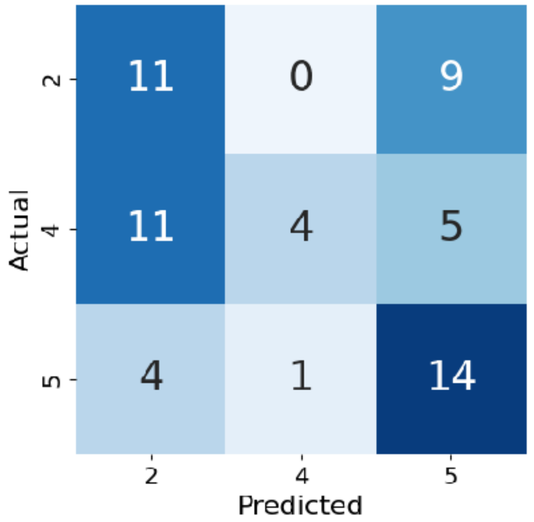
LLMs and Theory of Mind

- Generation as **action taking**
- Training fits a reward model (RM) of underlying true reward r^* using *reward advantage* of winner y_w over loser y_ℓ
- If $y_A \succ y_B$ and $y_B \succ y_C$ in data, RM implicitly scores $y_A \succ y_C$
- Given samples where $y_C \succ y_A$, $\sigma\big(RM_\theta(x, y_C) - RM_\theta(x, y_A)\big) \to 0 \Rightarrow$ $-\mathbb{E}[\log\sigma\big(RM_\theta(x, y_C) - RM_\theta(x, y_A)\big)] \to \infty$, leading to unstable updates
- RM underfits to at least one of $y_A \succ y_B$, $y_B \succ y_C$, or $y_C \succ y_A$ samples, stochastically flip preferences
- Preference-aligned LLMs do not capture ToM

When asked to choose the next utterance in a dialogue, LLMs produce noise



GPT-4o next utterance predictions (10x) in collaborative dialogue sequence: Weights Task



GPT-4o next utterance predictions (10x) in collaborative dialogue sequence: Wason (DeliData)

Solution: “Friction”

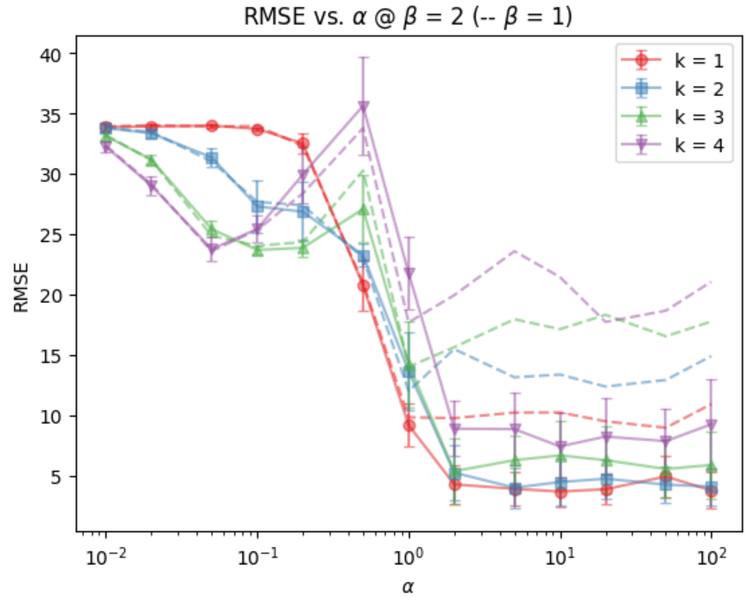
- Let evidence model $\mathscr{M}_a$ be a Kripke model $\langle A, W, E, V \rangle$
  - Agents A, worlds W, accessibility (evidencing) relation E, valuation function V
- If an event is public, each agent’s belief set typically refines to those worlds consistent with the event’s precondition
  - Usually, we assume that all agents smoothly integrate the new proposition
  - But if the proposition conflicts strongly with the agent’s prior beliefs, friction ensues
- Friction occurs when an agent’s newly updated beliefs cannot be derived by simple monotonic restriction of the old ones
  - Formally, consider an agent a with old beliefs B_a^{old}, updated by ϕ_j to B_a^{new}
  - If $B_a^{\text{new}} \not\subseteq B_a^{\text{old}} \cup \{\psi \text{ easily entailed}\}$, we interpret this as friction (a “**frictive state**”)
  - In simpler terms, friction is a necessity of nontrivial belief revision rather than a smooth refinement

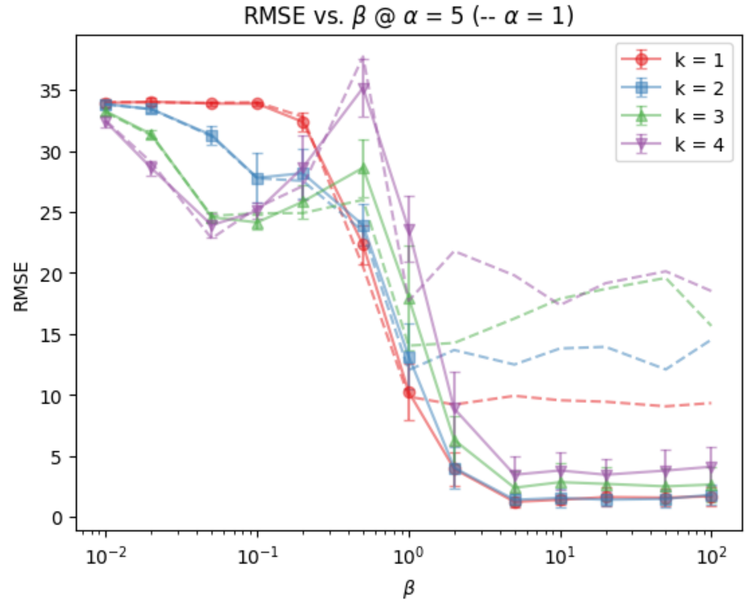
Dynamic Epistemic Friction

- Friction may both indicate an impasse (frictive state) and be used to resolve it
  - When impeded from moving forward along its present course, a fluid system redirects to the path of least resistance
  - When impeded in its present direction, a dialogue redirects in order to proceed
- Define an update function $B_a^{k+1} = B_a^k + \Delta B_a$, where $\Delta B_a = -\nabla F(\phi_j^*, B_a, E_j')$, and either adds evidence E_j' to existing beliefs B_a or **modifies** $\phi_j \rightarrow \phi_j^*$ to make it align better with B_a
- When ϕ_j^* is non-contradictory to B_a, this may redirect the dialogue
- Examples: intervening to reconsider assumptions, prompting to consider alternative options, maintaining appropriate levels of uncertainty as evidence accumulates

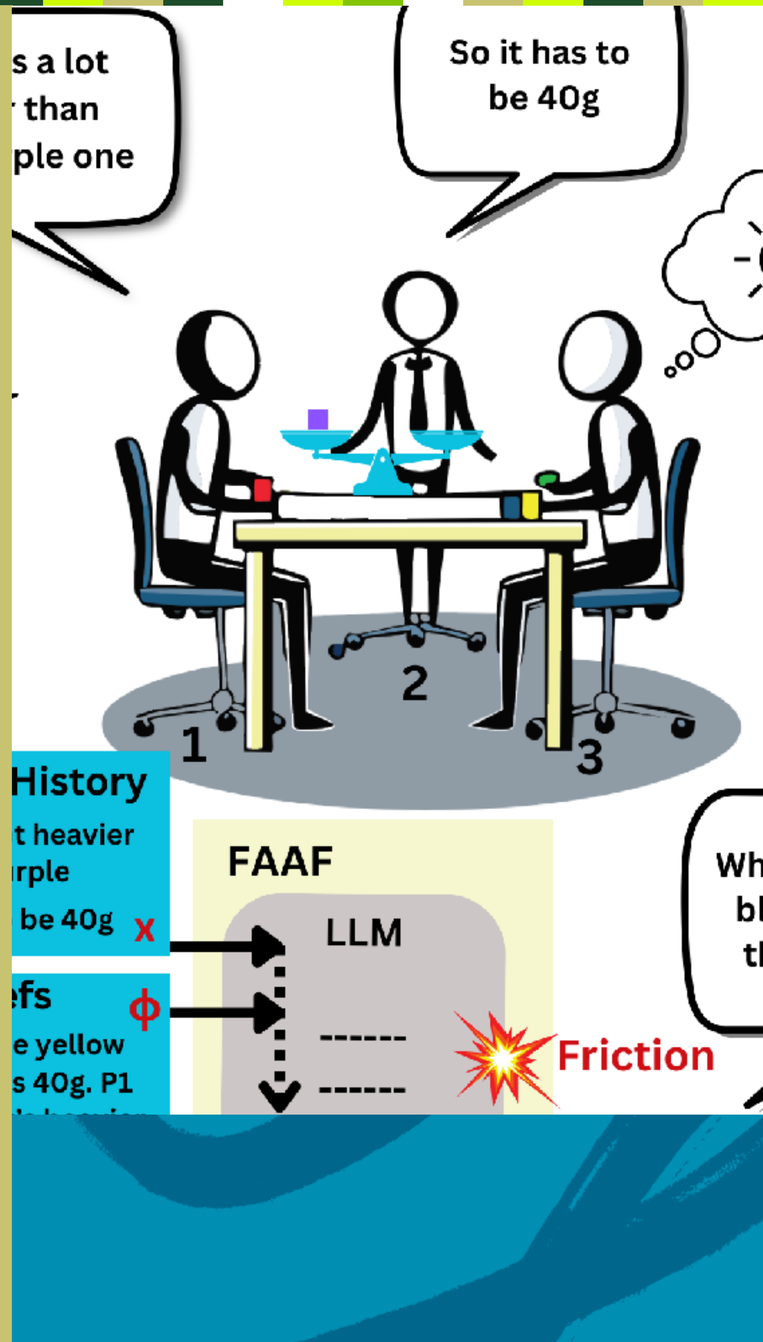
Dynamic Epistemic Friction

- Dynamic Epistemic Friction (DEF) predicts task state
- Weights Task: “red and blue are both 10” $\rightarrow [10,10,0,0,0]$, “blue isn’t 10” $\rightarrow [0,-10,0,0,0]$, correct final task state $\rightarrow [10,10,20,30,50]$
- Update function:
  - $\overrightarrow{\varphi}'_a = \overrightarrow{\varphi}_a + \min(\beta, \alpha \times \text{sgn}(\varphi_a \cdot \overrightarrow{\varphi}_b)) \times \text{CosSim}(\overrightarrow{\varphi}_a, \overrightarrow{\varphi}_b) \odot \overrightarrow{\varphi}_b$, where coefficients α and β modulate “force” of updates
- Fit linear ridge regressor over final computed task state, apply to held-out test group
- Appropriate friction coefficients act as strong regularizers
- High accuracy in predicting final task state (common ground)
- T. Obiso, K. Lai, A., Nath, N. Krishnaswamy, and J. Pustejovsky. *Dynamic Epistemic Friction in Dialogue*, to appear at CoNLL 2025





Frictional Agent Alignment



Frictional Agent Alignment

- DEF: Empirical demonstration of friction in numerical simulation, not applied to real LLMs

Frictional Agent Alignment

- DEF: Empirical demonstration of friction in numerical simulation, not applied to real LLMs
- **How can a preference-aligned LLM determine what to say when:**
  - **Given a frictive state, multiple interventions may be equally grounded to the available evidence?**
  - **Preferences may be changing, non-deterministic, and/or intransitive?**

Frictional Agent Alignment

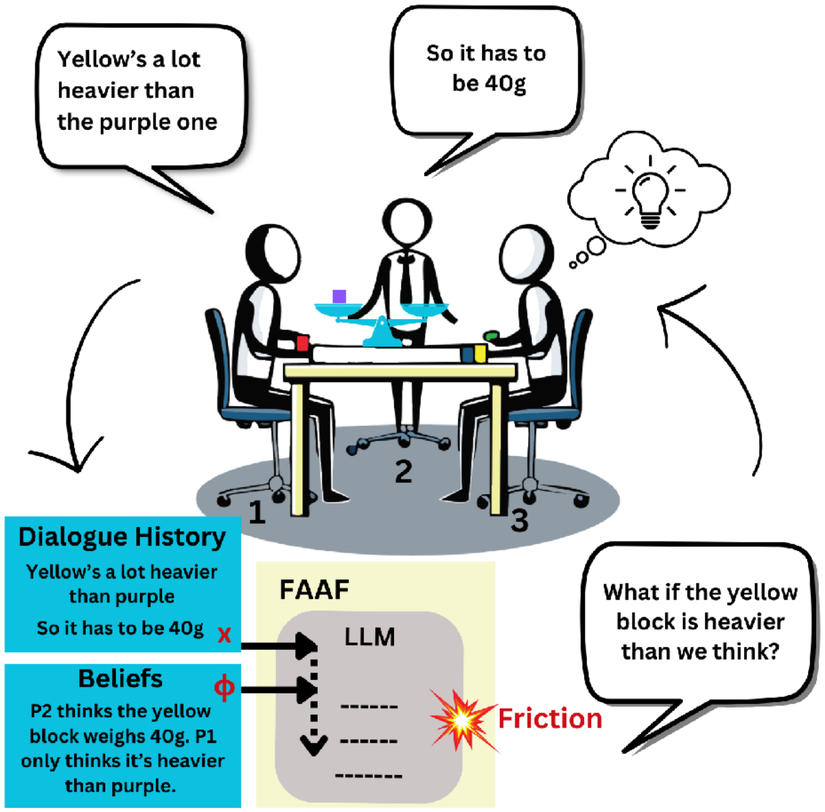
- DEF: Empirical demonstration of friction in numerical simulation, not applied to real LLMs
- **How can a preference-aligned LLM determine what to say when:**
  - **Given a frictive state, multiple interventions may be equally grounded to the available evidence?**
  - **Preferences may be changing, non-deterministic, and/or intransitive?**
- Intuition: *any* agent intervention satisfies “slow down” requirement of friction, therefore we can model advantage of one intervention over another

Frictional Agent Alignment

- DEF: Empirical demonstration of friction in numerical simulation, not applied to real LLMs
- **How can a preference-aligned LLM determine what to say when:**
  - **Given a frictive state, multiple interventions may be equally grounded to the available evidence?**
  - **Preferences may be changing, non-deterministic, and/or intransitive?**
- Intuition: *any* agent intervention satisfies “slow down” requirement of friction, therefore we can model advantage of one intervention over another
- Weighted pair objective: $w = \frac{\Delta F}{\Delta F + c}, \Delta F = R(x, y^+) - R(x, y^-)$, where y^+ and y^- represent helpful vs. non-helpful interventions, respectively

Frictional Agent Alignment

- **Problem 1: SOTA LLMs struggle to track and reason about evidence models**
  - So what? With data, we could train a model to do this!
- **Problem 2: Naturally-occurring friction contributes to task success but is sparse in data**
  - 3.46 frictive (probing) utterances per group in DeliData, 4 per group in Weights Task
  - Current practice uses high-capacity LLMs to generate training data, but sparseness in the source data induces skewness in the generated distribution!
- **Proposed solution: Frictional Agent Alignment Framework (FAAF)**
  - Models reward advantage of intervention f over frictive state ϕ



A. Nath, C. Graff, A. Bachinin, and N. Krishnaswamy.
Frictional Agent Alignment Framework: Slow Down and Don't Break Things,
to appear at ACL 2025.

Frictional Agent Alignment: Theory

- **Two-player min-max objective that learns two interdependent collaborative policies**
  - "Frictive state policy" π_ϕ^* generates the most semantically rich frictive states
  - "Friction intervention policy" π_f^* generates constructive interventions conditioned on the frictive state
- Optimal combined policy should not generate arbitrary interventions in the dialogue
  - Should surface presuppositions that gave rise to the salient frictive state
  - Should make interventions precise and interpretable

$$J^*_{\text{FAAF}} = \min_{\pi_\phi} \max_{\pi_f} \mathbb{E}_{x\sim\rho,\phi\sim\pi_\phi(\cdot|x),f\sim\pi_f(\cdot|\phi,x)} \Big[\mathscr{P}(f \succ \phi \mid x) - \beta D_{\text{KL}}(\pi_f \,\|\, \pi_{\text{ref}} \mid \phi, x) + \beta D_{\text{KL}}(\pi_\phi \,\|\, \pi_{\text{ref}} \mid x) \Big]$$

Frictional Agent Alignment: Theory

- Derivation of the frictive state policy as a valid probability distribution allows expression of combined objective in terms of a single policy π_θ
  - Introduce a Lagrange multiplier and define the corresponding function L to derive the optimality conditions
- Two terms:
  - ΔR conditions likelihood ratio on context *and* frictive state
  - $\Delta R'$ conditions likelihood ratio on *only* context
- Allows an IPO-like empirical loss $\mathscr{L} = \mathbb{E}_{\mathscr{D}_\mu}[(1 - \beta(\Delta R + \Delta R'))^2]$
- Quadratic loss mitigates effect of negative reward differences (e.g., in intransitive preference cycles) while ΔR constrains likelihood ratio difference due to frictive state conditioning

Algorithm 1 Frictional Agent Alignment Framework

Require: Training data $\mathcal{D}_\mu$ containing tuples (x, ϕ, f_w, f_l), where x: prompt, ϕ: frictive state, f_w: preferred response, f_l: non-preferred response.

1: Define likelihood ratios:

2: $\Delta R = \log\left(\frac{\pi_\theta(f_w|\phi,x)}{\pi_{\text{ref}}(f_w|\phi,x)}\right) - \log\left(\frac{\pi_\theta(f_l|\phi,x)}{\pi_{\text{ref}}(f_l|\phi,x)}\right)$

3: $\Delta R' = \log\left(\frac{\pi_\theta(f_w|x)}{\pi_{\text{ref}}(f_w|x)}\right) - \log\left(\frac{\pi_\theta(f_l|x)}{\pi_{\text{ref}}(f_l|x)}\right)$

4: Loss function: $\mathcal{L} = \mathbb{E}_{\mathcal{D}_\mu}[(1 - \beta(\Delta R + \Delta R'))^2]$

5: Gradient update: $\nabla_\theta \mathcal{L} = \mathbb{E}_{\mathcal{D}_\mu}[-2\beta\delta\nabla_\theta \log(\Delta R \cdot \Delta R')]$, where $\delta = 1 - \beta(\log \Delta R + \log \Delta R')$

6: Update policy parameters θ using gradient descent

Frictional Agent Alignment: Experiments

- Experimental datasets: GPT-4o-augmented DeliData and Weights Task Dataset, fully-simulated WTD dialogues
- Self-rewarding strategy to label frictive states, generate interventions and score them 1-10
  - DeliData: **68,618** preference training samples; mean preferred reward **8.03**, dispreferred **3.96**
    - 50 randomly-sampled dialogues held out for testing
  - Simulated WTD: 56,698 preference training samples; mean preferred reward **8.48**, dispreferred **6.01**
    - 54 dialogues held out for testing
  - Original WTD: 4,299 preference samples; mean preferred reward **8.36**, dispreferred **6.35**
    - All held out for OOD testing
- Compare FAAF to PPO, DPO, IPO in terms of LLM-judge-assigned win-rate against SFT model

Policy	Overall	Ac	Ga	Im	Rf	Re	Sp	Th
				DELIDATA				
PPO	$68.9_{\pm 1.5}$	$59.9_{\pm 1.5}$	$65.4_{\pm 1.5}$	$68.6_{\pm 1.5}$	$64.9_{\pm 1.5}$	$65.1_{\pm 1.5}$	$71.1_{\pm 1.4}$	$64.0_{\pm 1.5}$
IPO	$70.1_{\pm 1.4}$	$61.2_{\pm 1.5}$	$65.7_{\pm 1.5}$	$69.3_{\pm 1.5}$	$65.3_{\pm 1.5}$	$65.5_{\pm 1.5}$	$72.1_{\pm 1.4}$	$64.1_{\pm 1.5}$
DPO	$70.8_{\pm 1.4}$	$61.0_{\pm 1.5}$	$66.8_{\pm 1.5}$	$69.6_{\pm 1.5}$	$66.1_{\pm 1.5}$	$67.5_{\pm 1.5}$	$72.2_{\pm 1.4}$	$66.2_{\pm 1.5}$
FAAF	$75.7_{\pm 1.4}$	$65.6_{\pm 1.5}$	$69.5_{\pm 1.5}$	$75.0_{\pm 1.4}$	$72.0_{\pm 1.4}$	$71.1_{\pm 1.4}$	$75.3_{\pm 1.4}$	$70.4_{\pm 1.4}$
				WTD ORIGINAL				
PPO	$76.0_{\pm 4.3}$	$74.0_{\pm 4.4}$	$75.0_{\pm 4.3}$	$75.0_{\pm 4.3}$	$67.0_{\pm 4.7}$	$70.0_{\pm 4.6}$	$73.0_{\pm 4.4}$	$74.0_{\pm 4.4}$
IPO	$82.0_{\pm 3.8}$	$87.0_{\pm 3.4}$	$75.0_{\pm 4.3}$	$84.0_{\pm 3.7}$	$75.0_{\pm 4.3}$	$80.0_{\pm 4.0}$	$88.0_{\pm 3.2}$	$78.0_{\pm 4.1}$
DPO	$89.0_{\pm 3.1}$	$92.0_{\pm 2.7}$	$82.0_{\pm 3.8}$	$89.0_{\pm 3.1}$	$84.0_{\pm 3.7}$	$87.0_{\pm 3.4}$	$89.0_{\pm 3.1}$	$79.0_{\pm 4.1}$
FAAF	$90.9_{\pm 2.9}$	$81.8_{\pm 3.9}$	$84.8_{\pm 3.6}$	$90.9_{\pm 2.9}$	$86.9_{\pm 3.4}$	$89.9_{\pm 3.0}$	$88.9_{\pm 3.1}$	$90.9_{\pm 2.9}$
				WTD SIMULATED				
PPO	$73.6_{\pm 1.5}$	$69.7_{\pm 1.5}$	$64.9_{\pm 1.6}$	$74.2_{\pm 1.5}$	$67.6_{\pm 1.6}$	$71.9_{\pm 1.5}$	$78.1_{\pm 1.4}$	$78.3_{\pm 1.4}$
IPO	$83.0_{\pm 1.3}$	$74.8_{\pm 1.4}$	$78.4_{\pm 1.4}$	$82.9_{\pm 1.3}$	$76.9_{\pm 1.4}$	$81.4_{\pm 1.3}$	$82.5_{\pm 1.3}$	$83.2_{\pm 1.2}$
DPO	$82.9_{\pm 1.3}$	$80.4_{\pm 1.3}$	$75.8_{\pm 1.4}$	$81.3_{\pm 1.3}$	$72.9_{\pm 1.5}$	$76.3_{\pm 1.4}$	$80.2_{\pm 1.3}$	$79.2_{\pm 1.4}$
FAAF	$91.5_{+0.9}$	$87.5_{+1.1}$	$87.1_{+1.1}$	$90.1_{+1.0}$	$82.0_{+1.3}$	$85.1_{+1.2}$	$90.3_{+1.0}$	$90.1_{+1.0}$

Table 1: Win-rates (%) against the SFT model (π_{ref}) for all alignment methods on sampled interventions (temperature of 0.7, top-p of 0.9) from 500 randomly-sampled prompts from DeliData and WTD evaluation sets, according to GPT-4o. Metrics: **Ac** (*Actionability*), **Ga** (*Gold-alignment*), **Im** (*Impact*), **Rf** (*Rationale-fit*), **Re** (*Relevance*), **Sp** (*Specificity*), and **Th** (*Thought-provoking*). The LLM-as-a-judge evaluation follows Cui et al. (2024). Average win rates are reported over two runs, with positional swapping to mitigate position bias.

Frictional Agent Alignment: Experiments

- Experimental datasets: GPT-4o-augmented DeliData and Weights Task Dataset, fully-simulated WTD dialogues
- Self-rewarding strategy to label frictive states, generate interventions and score them 1-10
  - DeliData: **68,618** preference training samples; mean preferred reward **8.03**, dispreferred **3.96**
    - 50 randomly-sampled dialogues held out for testing
  - Simulated WTD: 56,698 preference training samples; mean preferred reward **8.48**, dispreferred **6.01**
    - 54 dialogues held out for testing
  - Original WTD: 4,299 preference samples; mean preferred reward **8.36**, dispreferred **6.35**
    - All held out for OOD testing
- Compare FAAF to PPO, DPO, IPO in terms of LLM-judge-assigned win-rate against SFT model

Policy	Overall	Ac	Ga	Im	Rf	Re	Sp	Th
				DELIDATA				
PPO	$68.9_{\pm 1.5}$	$59.9_{\pm 1.5}$	$65.4_{\pm 1.5}$	$68.6_{\pm 1.5}$	$64.9_{\pm 1.5}$	$65.1_{\pm 1.5}$	$71.1_{\pm 1.4}$	$64.0_{\pm 1.5}$
IPO	$70.1_{\pm 1.4}$	$61.2_{\pm 1.5}$	$65.7_{\pm 1.5}$	$69.3_{\pm 1.5}$	$65.3_{\pm 1.5}$	$65.5_{\pm 1.5}$	$72.1_{\pm 1.4}$	$64.1_{\pm 1.5}$
DPO	$70.8_{\pm 1.4}$	$61.0_{\pm 1.5}$	$66.8_{\pm 1.5}$	$69.6_{\pm 1.5}$	$66.1_{\pm 1.5}$	$67.5_{\pm 1.5}$	$72.2_{\pm 1.4}$	$66.2_{\pm 1.5}$
FAAF	$75.7_{\pm 1.4}$	$65.6_{\pm 1.5}$	$69.5_{\pm 1.5}$	$75.0_{\pm 1.4}$	$72.0_{\pm 1.4}$	$71.1_{\pm 1.4}$	$75.3_{\pm 1.4}$	$70.4_{\pm 1.4}$
				WTD ORIGINAL				
PPO	$76.0_{\pm 4.3}$	$74.0_{\pm 4.4}$	$75.0_{\pm 4.3}$	$75.0_{\pm 4.3}$	$67.0_{\pm 4.7}$	$70.0_{\pm 4.6}$	$73.0_{\pm 4.4}$	$74.0_{\pm 4.4}$
IPO	$82.0_{\pm 3.8}$	$87.0_{\pm 3.4}$	$75.0_{\pm 4.3}$	$84.0_{\pm 3.7}$	$75.0_{\pm 4.3}$	$80.0_{\pm 4.0}$	$88.0_{\pm 3.2}$	$78.0_{\pm 4.1}$
DPO	$89.0_{\pm 3.1}$	$92.0_{\pm 2.7}$	$82.0_{\pm 3.8}$	$89.0_{\pm 3.1}$	$84.0_{\pm 3.7}$	$87.0_{\pm 3.4}$	$89.0_{\pm 3.1}$	$79.0_{\pm 4.1}$
FAAF	$90.9_{\pm 2.9}$	$81.8_{\pm 3.9}$	$84.8_{\pm 3.6}$	$90.9_{\pm 2.9}$	$86.9_{\pm 3.4}$	$89.9_{\pm 3.0}$	$88.9_{\pm 3.1}$	$90.9_{\pm 2.9}$
				WTD SIMULATED				
PPO	$73.6_{\pm 1.5}$	$69.7_{\pm 1.5}$	$64.9_{\pm 1.6}$	$74.2_{\pm 1.5}$	$67.6_{\pm 1.6}$	$71.9_{\pm 1.5}$	$78.1_{\pm 1.4}$	$78.3_{\pm 1.4}$
IPO	$83.0_{\pm 1.3}$	$74.8_{\pm 1.4}$	$78.4_{\pm 1.4}$	$82.9_{\pm 1.3}$	$76.9_{\pm 1.4}$	$81.4_{\pm 1.3}$	$82.5_{\pm 1.3}$	$83.2_{\pm 1.2}$
DPO	$82.9_{\pm 1.3}$	$80.4_{\pm 1.3}$	$75.8_{\pm 1.4}$	$81.3_{\pm 1.3}$	$72.9_{\pm 1.5}$	$76.3_{\pm 1.4}$	$80.2_{\pm 1.3}$	$79.2_{\pm 1.4}$
FAAF	$91.5_{\pm 0.9}$	$87.5_{\pm 1.1}$	$87.1_{\pm 1.1}$	$90.1_{\pm 1.0}$	$82.0_{\pm 1.3}$	$85.1_{\pm 1.2}$	$90.3_{\pm 1.0}$	$90.1_{\pm 1.0}$

Table 1: Win-rates (%) against the SFT model (π_{ref}) for all alignment methods on sampled interventions (temperature of 0.7, top-p of 0.9) from 500 randomly-sampled prompts from DeliData and WTD evaluation sets, according to GPT-4o. Metrics: **Ac** (*Actionability*), **Ga** (*Gold-alignment*), **Im** (*Impact*), **Rf** (*Rationale-fit*), **Re** (*Relevance*), **Sp** (*Specificity*), and **Th** (*Thought-provoking*). The LLM-as-a-judge evaluation follows Cui et al. (2024). Average win rates are reported over two runs, with positional swapping to mitigate position bias.

Frictional Agent Alignment: Experiments

- Does ϕ-conditioning actually help?
- $\text{FAAF}_{\Delta R}$ (ϕ-conditioned) vs. $\text{FAAF}_{\Delta R'}$ (nonconditioned) vs. full-objective against SFT, PPO, DPO, IPO

Dataset	Policy	Win-rate vs. Base	Win-rate vs. SFT	Win-rate vs. DPO	Win-rate vs. IPO	Win-rate vs. PPO
DeliData	$\text{FAAF}_{\Delta R'}$	$82.2_{\pm 1.7}$	$78.8_{\pm 1.8}$	$74.0_{\pm 1.9}$	$53.6_{\pm 2.2}$	$79.2_{\pm 1.8}$
	$\text{FAAF}_{\Delta R}$	$85.8_{\pm 1.5}$	$81.4_{\pm 1.7}$	$73.2_{\pm 1.9}$	$54.2_{\pm 2.2}$	$73.4_{\pm 1.9}$
	$\text{FAAF}_{\Delta(R+R')}$	$86.2_{\pm 1.5}$	$84.0_{\pm 1.6}$	$75.6_{\pm 1.9}$	$79.6_{\pm 1.8}$	$76.0_{\pm 1.9}$
WTD Orig.	$\text{FAAF}_{\Delta R'}$	$78.0_{\pm 5.8}$	$78.0_{\pm 5.8}$	$76.0_{\pm 6.0}$	$58.0_{\pm 6.9}$	$58.0_{\pm 6.9}$
	$\text{FAAF}_{\Delta R}$	$68.0_{\pm 6.5}$	$74.0_{\pm 6.2}$	$72.0_{\pm 6.3}$	$62.0_{\pm 6.8}$	$70.0_{\pm 6.4}$
	$\text{FAAF}_{\Delta(R+R')}$	$84.0_{\pm 5.1}$	$76.0_{\pm 6.0}$	$74.0_{\pm 6.2}$	$74.0_{\pm 6.2}$	$82.0_{\pm 5.4}$
WTD Sim.	$\text{FAAF}_{\Delta R'}$	$79.1_{\pm 1.9}$	$80.2_{\pm 1.8}$	$70.4_{\pm 2.1}$	$68.6_{\pm 2.1}$	$60.8_{\pm 2.3}$
	$\text{FAAF}_{\Delta R}$	$85.7_{\pm 1.6}$	$80.8_{\pm 1.8}$	$70.8_{\pm 2.1}$	$72.2_{\pm 2.1}$	$74.8_{\pm 2.0}$
	$\text{FAAF}_{\Delta(R+R')}$	$88.0_{\pm 1.5}$	$83.7_{\pm 1.7}$	$72.8_{\pm 2.0}$	$73.7_{\pm 2.0}$	$75.1_{\pm 2.0}$

Table 2: Win rates of of FAAF variants—$\text{FAAF}_{\Delta R'}$ (not ϕ-conditioned), $\text{FAAF}_{\Delta R}$ (ϕ-conditioned), and $\text{FAAF}_{\Delta(R+R')}$ (full objective)—against competing methods in pairwise comparisons (temperature of 0.7, top-p of 0.9). All alignment baselines are SFT-initialized and `Meta-Llama-3-8B-Instruct` is used as Base.

Frictional Agent Alignment: Experiments

- Does ϕ-conditioning actually help?
- $\text{FAAF}_{\Delta R}$ (ϕ-conditioned) vs. $\text{FAAF}_{\Delta R'}$ (nonconditioned) vs. full-objective against SFT, PPO, DPO, IPO

Dataset	Policy	Win-rate vs. Base	Win-rate vs. SFT	Win-rate vs. DPO	Win-rate vs. IPO	Win-rate vs. PPO
DeliData	$\text{FAAF}_{\Delta R'}$	$82.2_{\pm 1.7}$	$78.8_{\pm 1.8}$	$74.0_{\pm 1.9}$	$53.6_{\pm 2.2}$	$79.2_{\pm 1.8}$
	$\text{FAAF}_{\Delta R}$	$85.8_{\pm 1.5}$	$81.4_{\pm 1.7}$	$73.2_{\pm 1.9}$	$54.2_{\pm 2.2}$	$73.4_{\pm 1.9}$
	$\text{FAAF}_{\Delta(R+R')}$	$86.2_{\pm 1.5}$	$84.0_{\pm 1.6}$	$75.6_{\pm 1.9}$	$79.6_{\pm 1.8}$	$76.0_{\pm 1.9}$
WTD Orig.	$\text{FAAF}_{\Delta R'}$	$78.0_{\pm 5.8}$	$78.0_{\pm 5.8}$	$76.0_{\pm 6.0}$	$58.0_{\pm 6.9}$	$58.0_{\pm 6.9}$
	$\text{FAAF}_{\Delta R}$	$68.0_{\pm 6.5}$	$74.0_{\pm 6.2}$	$72.0_{\pm 6.3}$	$62.0_{\pm 6.8}$	$70.0_{\pm 6.4}$
	$\text{FAAF}_{\Delta(R+R')}$	$84.0_{\pm 5.1}$	$76.0_{\pm 6.0}$	$74.0_{\pm 6.2}$	$74.0_{\pm 6.2}$	$82.0_{\pm 5.4}$
WTD Sim.	$\text{FAAF}_{\Delta R'}$	$79.1_{\pm 1.9}$	$80.2_{\pm 1.8}$	$70.4_{\pm 2.1}$	$68.6_{\pm 2.1}$	$60.8_{\pm 2.3}$
	$\text{FAAF}_{\Delta R}$	$85.7_{\pm 1.6}$	$80.8_{\pm 1.8}$	$70.8_{\pm 2.1}$	$72.2_{\pm 2.1}$	$74.8_{\pm 2.0}$
	$\text{FAAF}_{\Delta(R+R')}$	$88.0_{\pm 1.5}$	$83.7_{\pm 1.7}$	$72.8_{\pm 2.0}$	$73.7_{\pm 2.0}$	$75.1_{\pm 2.0}$

Table 2: Win rates of of FAAF variants—$\text{FAAF}_{\Delta R'}$ (not ϕ-conditioned), $\text{FAAF}_{\Delta R}$ (ϕ-conditioned), and $\text{FAAF}_{\Delta(R+R')}$ (full objective)—against competing methods in pairwise comparisons (temperature of 0.7, top-p of 0.9). All alignment baselines are SFT-initialized and `Meta-Llama-3-8B-Instruct` is used as Base.

Evaluating Friction

Counterfactual Evaluation

- Problems with offline evaluation conditions
  - Multiturn benchmarks (e.g., MTBench) evaluate against fixed set of user responses
  - Rated impact on future dialogue is never fully verified
- Counterfactual evaluation using LLM “role-play”
  - In two tasks (Weights Task and Wason Card Task), we generate 15-turn dialogue trajectories between collaborators (π^C, explicitly roleplaying humans) and a friction agent ($\mathcal{O}$) with GPT-4o
  - “Gold” data used to train multiple friction agent policies π^F according to various alignment frameworks

Counterfactual Evaluation

- Use *untrained* instruction-tuned π^{base} to replace $\pi^{F\prime}$s interventions $\mathcal{F}$ with "no friction" (NF) interventions $\mathcal{F}^{\text{base}}$
- Re-sample π^{C} responses to $\mathcal{F}^{\text{base}}$ instead of $\mathcal{F}$, creating dialogue where π^{C} received interventions from untrained agent
- Assess size of common ground and task solution after final turn

Counterfactual Evaluation

- Use *untrained* instruction-tuned π^{base} to replace $\pi^{F\prime}$s interventions $\mathcal{F}$ with "no friction" (NF) interventions $\mathcal{F}^{\text{base}}$
- Re-sample π^{C} responses to $\mathcal{F}^{\text{base}}$ instead of $\mathcal{F}$, creating dialogue where π^{C} received interventions from untrained agent
- Assess size of common ground and task solution after final turn
  - How agreed is the group?

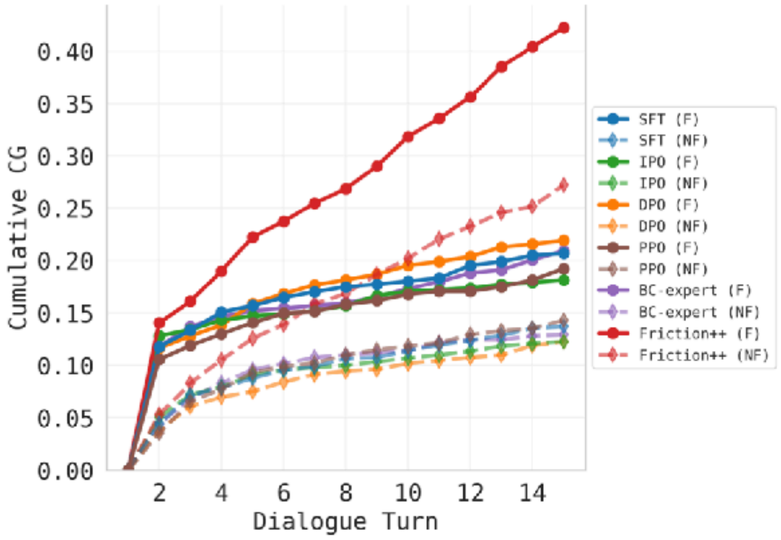
Counterfactual Evaluation

- Use *untrained* instruction-tuned π^{base} to replace $\pi^{F\prime}$s interventions $\mathcal{F}$ with “no friction” (NF) interventions $\mathcal{F}^{\text{base}}$
- Re-sample π^{C} responses to $\mathcal{F}^{\text{base}}$ instead of $\mathcal{F}$, creating dialogue where π^{C} received interventions from untrained agent
- Assess size of common ground and task solution after final turn
  - How agreed is the group?
  - How correct is their answer?

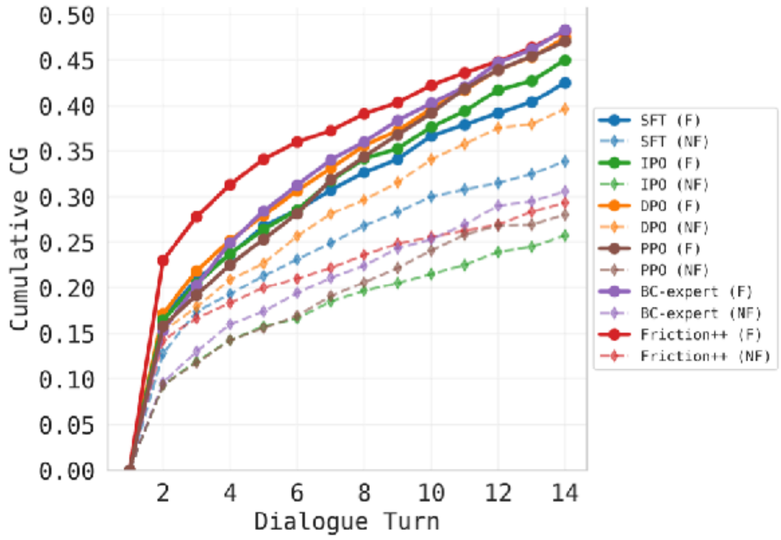
Counterfactual Evaluation

- Use *untrained* instruction-tuned π^{base} to replace $\pi^{F\prime}$s interventions $\mathcal{F}$ with “no friction” (NF) interventions $\mathcal{F}^{\text{base}}$
- Re-sample π^{C} responses to $\mathcal{F}^{\text{base}}$ instead of $\mathcal{F}$, creating dialogue where π^{C} received interventions from untrained agent
- Assess size of common ground and task solution after final turn
  - How agreed is the group?
  - How correct is their answer?
  - How does performance compare to counterfactual dialogue where agent is not optimized for friction?

Counterfactual Evaluation



Weights Task



DeliData

Slow down to speed up: groups with optimized friction agent converge to common ground more quickly

Friction++ (FAAF) optimization is best friction agent

Counterfactual Evaluation

Model	WTD		DeliData			
	Acc.	Acc. (MA)	Acc.	FG Acc.	Acc. (MA)	FG Acc. (MA)
SFT	$7.45_{\pm 0.10}$	$6.28_{\pm 0.05}$	$0.29_{\pm 0.05}$	$0.75_{\pm 0.02}$	$0.18_{\pm 0.04}$	$0.48_{\pm 0.02}$
IPO	$12.57_{\pm 0.13}$	$9.73_{\pm 0.09}$	$0.44_{\pm 0.05}$	$0.82_{\pm 0.02}$	$0.31_{\pm 0.05}$	$0.69_{\pm 0.02}$
DPO	$11.76_{\pm 0.13}$	$8.58_{\pm 0.08}$	$0.48_{\pm 0.05}$	$0.81_{\pm 0.02}$	$0.27_{\pm 0.04}$	$0.70_{\pm 0.02}$
PPO	$8.70_{\pm 0.09}$	$9.93_{\pm 0.10}$	$0.36_{\pm 0.05}$	$0.75_{\pm 0.02}$	$0.36_{\pm 0.04}$	$0.67_{\pm 0.02}$
BC-EXPERT	$14.82_{\pm 0.13}$	$10.10_{\pm 0.11}$	$0.54_{\pm 0.05}$	$0.80_{\pm 0.02}$	$0.37_{\pm 0.04}$	$0.72_{\pm 0.02}$
FRICTION$_{\Delta R'}$	$9.03_{\pm 0.10}$	$7.56_{\pm 0.08}$	$0.39_{\pm 0.05}$	$0.79_{\pm 0.02}$	$0.30_{\pm 0.05}$	$0.62_{\pm 0.02}$
FRICTION++	$\mathbf{14.91}_{\pm 0.14}$	$\mathbf{14.16}_{\pm 0.13}$	$\mathbf{0.60}_{\pm 0.05}$	$\mathbf{0.87}_{\pm 0.02}$	$\mathbf{0.45}_{\pm 0.05}$	$\mathbf{0.80}_{\pm 0.02}$

Mean task solution accuracy by friction agent type (FG = “fine-grained” accuracy with partial credit)

Slow down to speed up: groups with optimized friction agent arrive at more correct solutions

Counterfactual Evaluation

Model	WTD		DeliData			
	Acc.	Acc. (MA)	Acc.	FG Acc.	Acc. (MA)	FG Acc. (MA)
SFT	$7.45_{\pm 0.10}$	$6.28_{\pm 0.05}$	$0.29_{\pm 0.05}$	$0.75_{\pm 0.02}$	$0.18_{\pm 0.04}$	$0.48_{\pm 0.02}$
IPO	$12.57_{\pm 0.13}$	$9.73_{\pm 0.09}$	$0.44_{\pm 0.05}$	$0.82_{\pm 0.02}$	$0.31_{\pm 0.05}$	$0.69_{\pm 0.02}$
DPO	$11.76_{\pm 0.13}$	$8.58_{\pm 0.08}$	$0.48_{\pm 0.05}$	$0.81_{\pm 0.02}$	$0.27_{\pm 0.04}$	$0.70_{\pm 0.02}$
PPO	$8.70_{\pm 0.09}$	$9.93_{\pm 0.10}$	$0.36_{\pm 0.05}$	$0.75_{\pm 0.02}$	$0.36_{\pm 0.04}$	$0.67_{\pm 0.02}$
BC-EXPERT	$14.82_{\pm 0.13}$	$10.10_{\pm 0.11}$	$0.54_{\pm 0.05}$	$0.80_{\pm 0.02}$	$0.37_{\pm 0.04}$	$0.72_{\pm 0.02}$
FRICTION$_{\Delta R'}$	$9.03_{\pm 0.10}$	$7.56_{\pm 0.08}$	$0.39_{\pm 0.05}$	$0.79_{\pm 0.02}$	$0.30_{\pm 0.05}$	$0.62_{\pm 0.02}$
FRICTION++	$\mathbf{14.91}_{\pm 0.14}$	$\mathbf{14.16}_{\pm 0.13}$	$\mathbf{0.60}_{\pm 0.05}$	$\mathbf{0.87}_{\pm 0.02}$	$\mathbf{0.45}_{\pm 0.05}$	$\mathbf{0.80}_{\pm 0.02}$

Mean task solution accuracy by friction agent type (FG = “fine-grained” accuracy with partial credit)

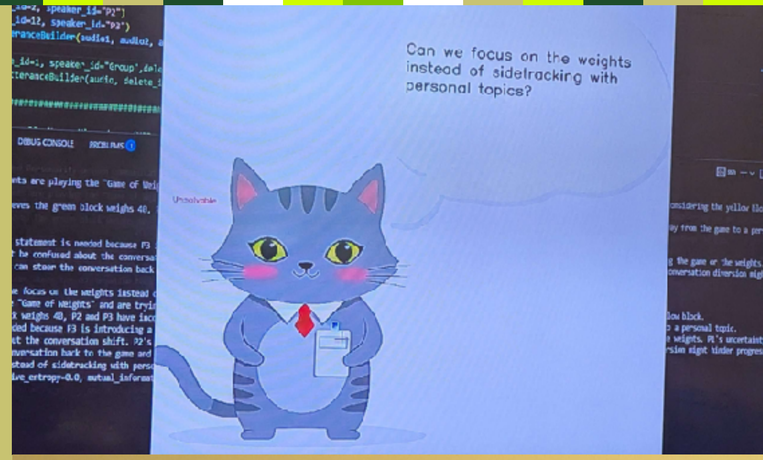
Slow down to speed up: groups with optimized friction agent arrive at more correct solutions

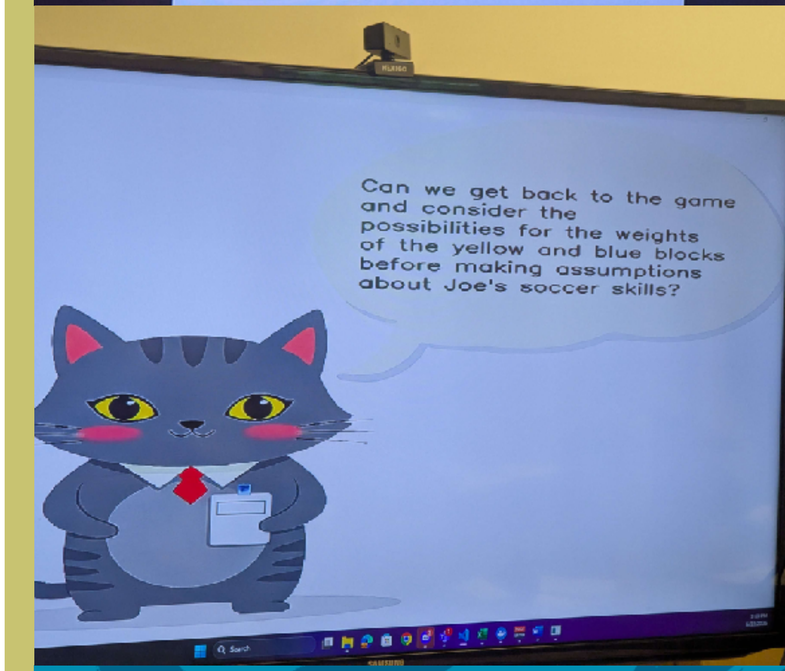
Counterfactual Evaluation

Model	WTD		DeliData			
	Acc.	Acc. (MA)	Acc.	FG Acc.	Acc. (MA)	FG Acc. (MA)
SFT	$7.45_{\pm 0.10}$	$6.28_{\pm 0.05}$	$0.29_{\pm 0.05}$	$0.75_{\pm 0.02}$	$0.18_{\pm 0.04}$	$0.48_{\pm 0.02}$
IPO	$12.57_{\pm 0.13}$	$9.73_{\pm 0.09}$	$0.44_{\pm 0.05}$	$0.82_{\pm 0.02}$	$0.31_{\pm 0.05}$	$0.69_{\pm 0.02}$
DPO	$11.76_{\pm 0.13}$	$8.58_{\pm 0.08}$	$0.48_{\pm 0.05}$	$0.81_{\pm 0.02}$	$0.27_{\pm 0.04}$	$0.70_{\pm 0.02}$
PPO	$8.70_{\pm 0.09}$	$9.93_{\pm 0.10}$	$0.36_{\pm 0.05}$	$0.75_{\pm 0.02}$	$0.36_{\pm 0.04}$	$0.67_{\pm 0.02}$
BC-EXPERT	$14.82_{\pm 0.13}$	$10.10_{\pm 0.11}$	$0.54_{\pm 0.05}$	$0.80_{\pm 0.02}$	$0.37_{\pm 0.04}$	$0.72_{\pm 0.02}$
FRICTION$_{\Delta R'}$	$9.03_{\pm 0.10}$	$7.56_{\pm 0.08}$	$0.39_{\pm 0.05}$	$0.79_{\pm 0.02}$	$0.30_{\pm 0.05}$	$0.62_{\pm 0.02}$
FRICTION++	$\mathbf{14.91}_{\pm 0.14}$	$\mathbf{14.16}_{\pm 0.13}$	$\mathbf{0.60}_{\pm 0.05}$	$\mathbf{0.87}_{\pm 0.02}$	$\mathbf{0.45}_{\pm 0.05}$	$\mathbf{0.80}_{\pm 0.02}$

- Remember this? Friction update function may **modify** intervention to better align with beliefs
- Under “modified action” condition, collaborator will ignore or interpret intervention in way that most aligns with existing beliefs
    - In this setting, Friction++ (FAAF) is most robust to perturbation

Modality Matters





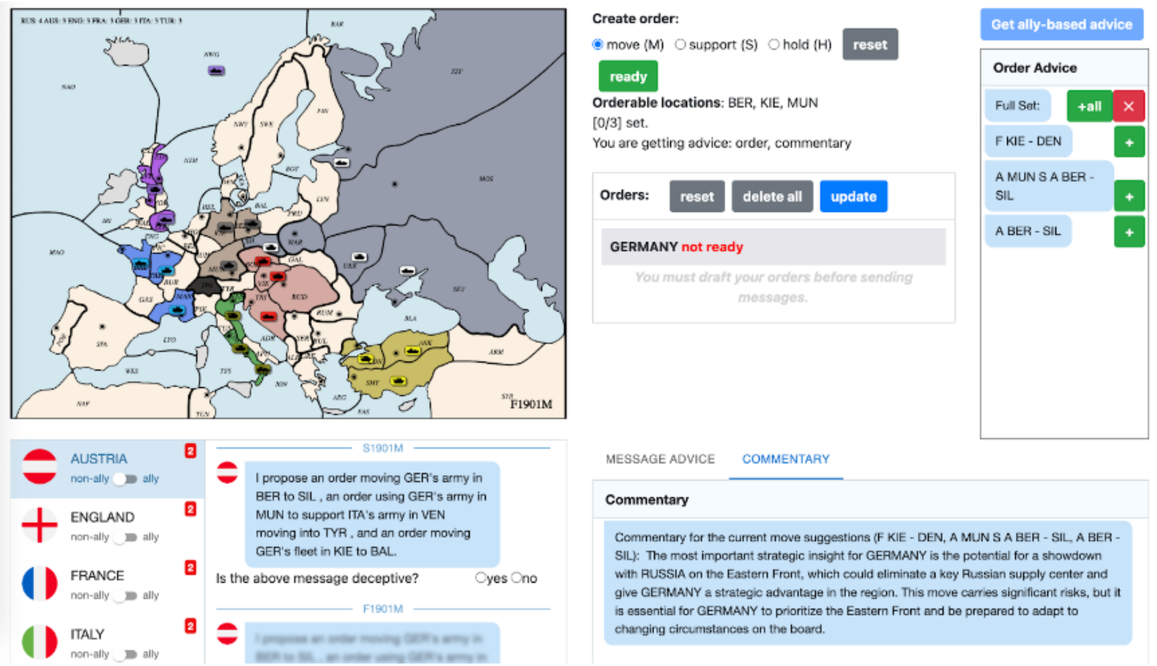
Human-User Evaluation

- Humans are famous for flummoxing the most theoretically-rigorous and best-evaluated AI systems

Imprecise, hard to interpret FAAF intervention

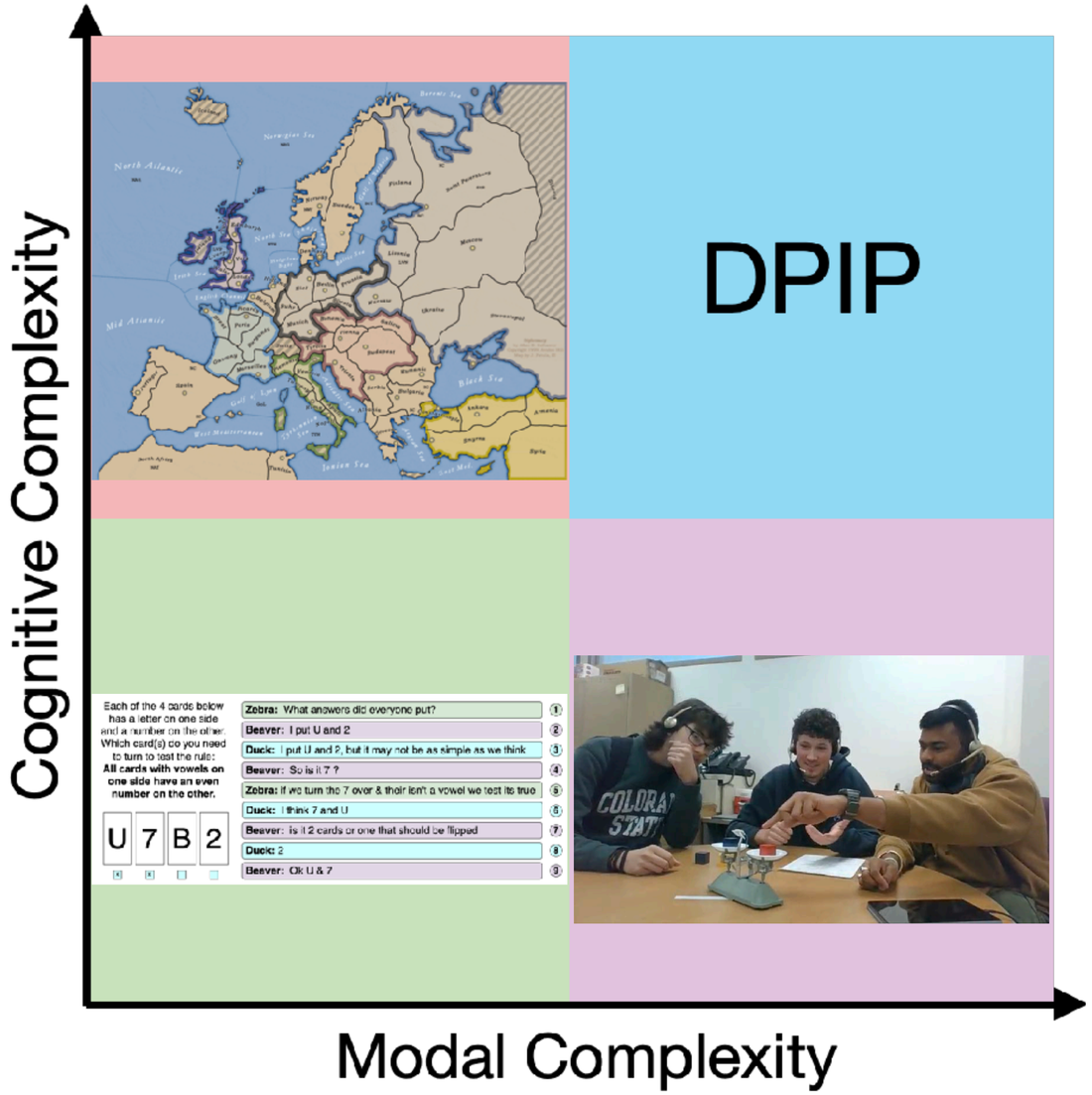
Complexity Axes

- FAAF aligned on generated Wason data deployed in Diplomacy games
- Out of the box, excelled at picking up novel task context
- Advice on topic, coherent, specific
- Even helped some players win games

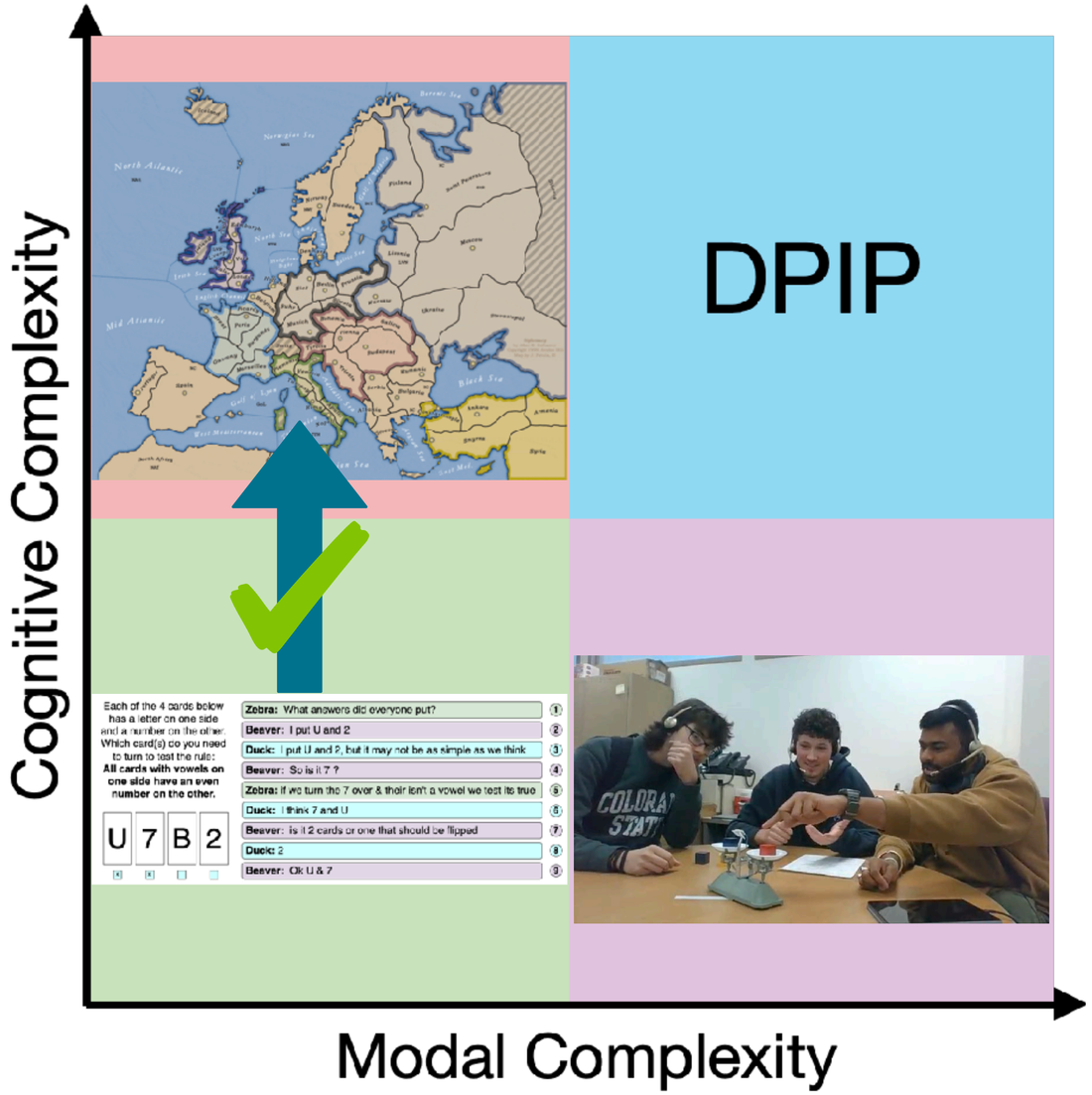


USC-ISI/UMD Diplomacy Interface (cf. Wongkamjan et al., ACL 2024)

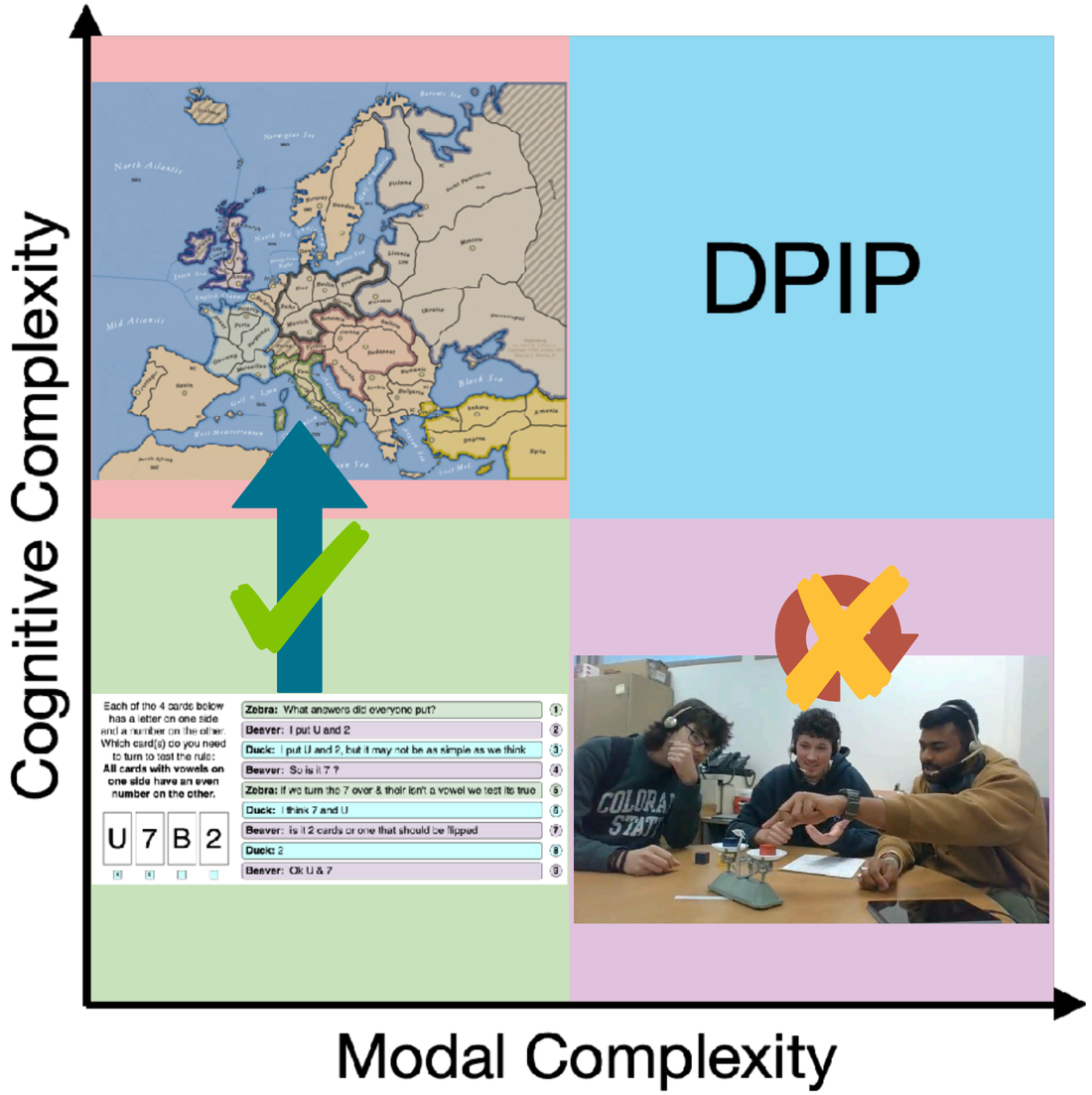
Complexity Axes



Complexity Axes



Complexity Axes



Modality Matters

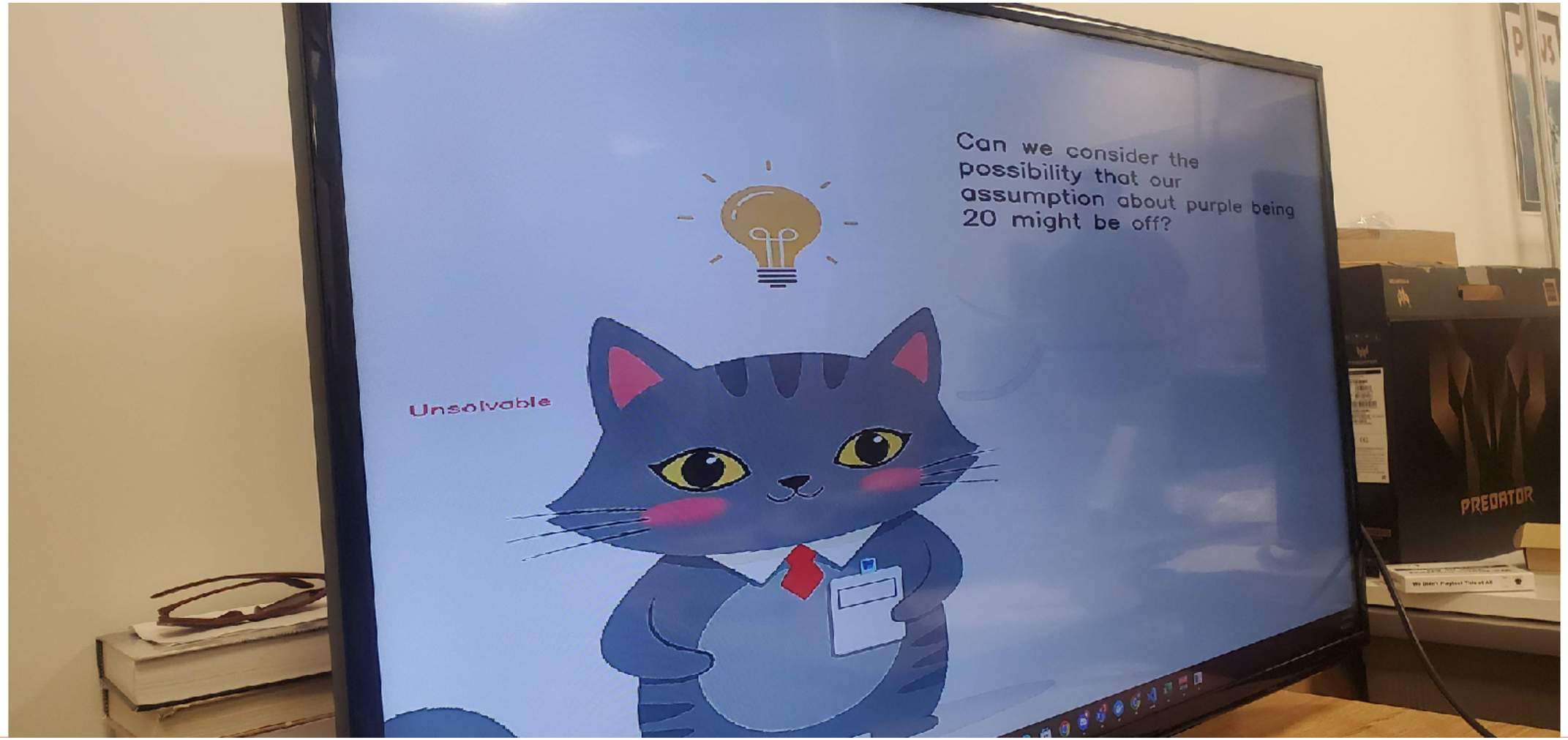
- Segmentation and transcription errors turned FAAF's contextual strength into a weakness
- Role of reference model in LLM alignment: KL-divergence term $\beta D_{\text{KL}}(\pi^* \parallel \pi_{\text{ref}} \mid x)$ ensures that π^* generates desirable outputs without destroying general language processing ability

FAAF Outputs
Can we assume the zero is the only reference point for comparison?
Can we assume the weight difference is always exactly 10 grams when the slider isn't at zero?
Can we assume that yellow always stays at zero if not moved without considering other weighing possibilities?
Can we assume assuming green is 50 assumes the same weighing scenario as before?

Modality Matters

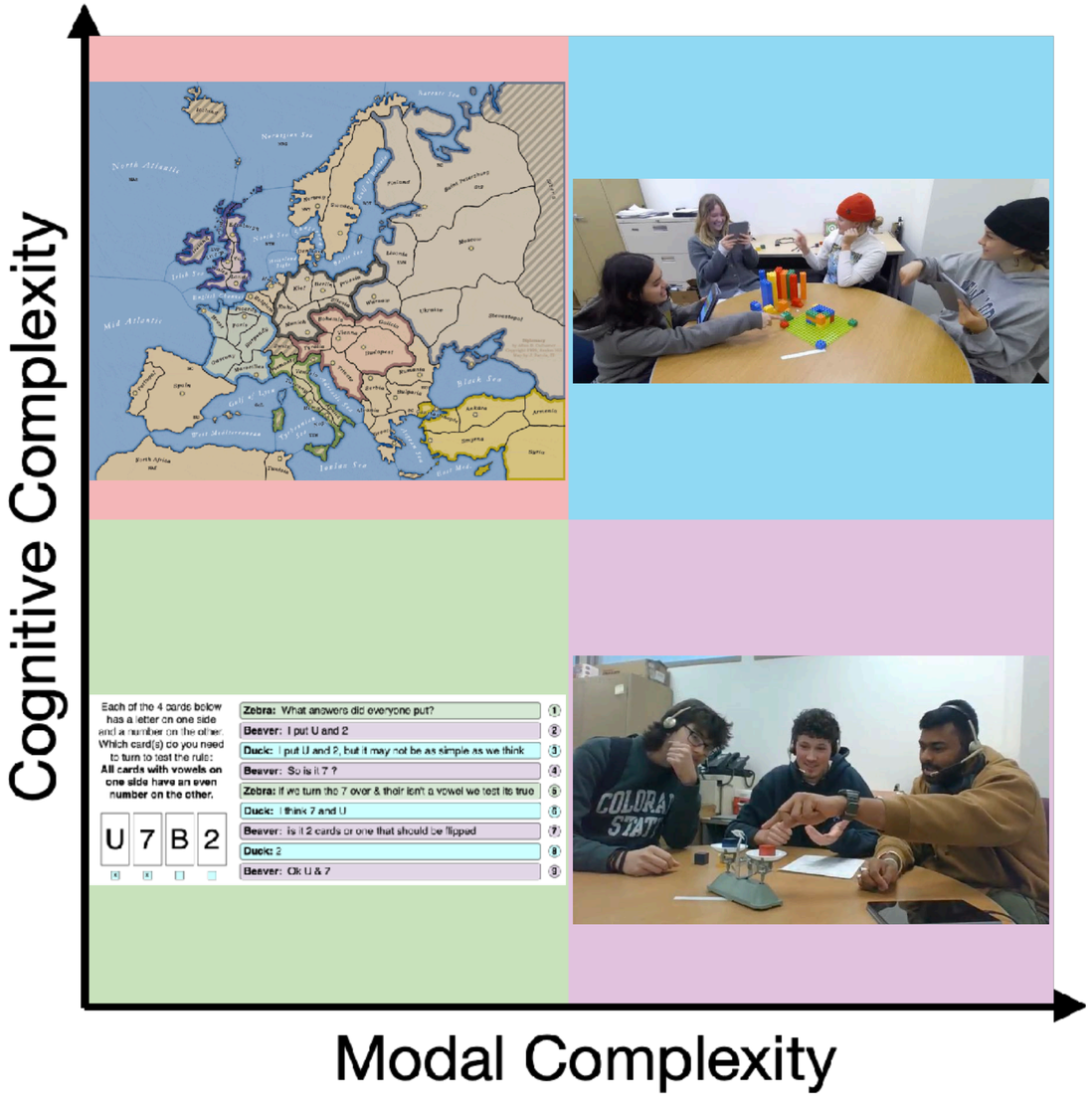
- FAAF reference model fine-tuned on synthetic generated (clean) data
- Solution: SFT the reference model on noisy original dialogues
    - Small amount of data, complete with automatic transcription and segmentation errors

Old FAAF	New FAAF
Can we assume the zero is the only reference point for comparison?	Can we consider the possibility that the slider doesn't need to be at zero for the scale to provide accurate readings?
Can we assume the weight difference is always exactly 10 grams when the slider isn't at zero?	Can we reconsider our assumptions about yellow's weight before making a final decision?
Can we assume that yellow always stays at zero if not moved without considering other weighing possibilities?	Can we re-examine our assumptions about the yellow block's weight and its role in the game?
Can we assume assuming green is 50 assumes the same weighing scenario as before?	Can we revisit our assumptions about green and blue before making any further conclusions?

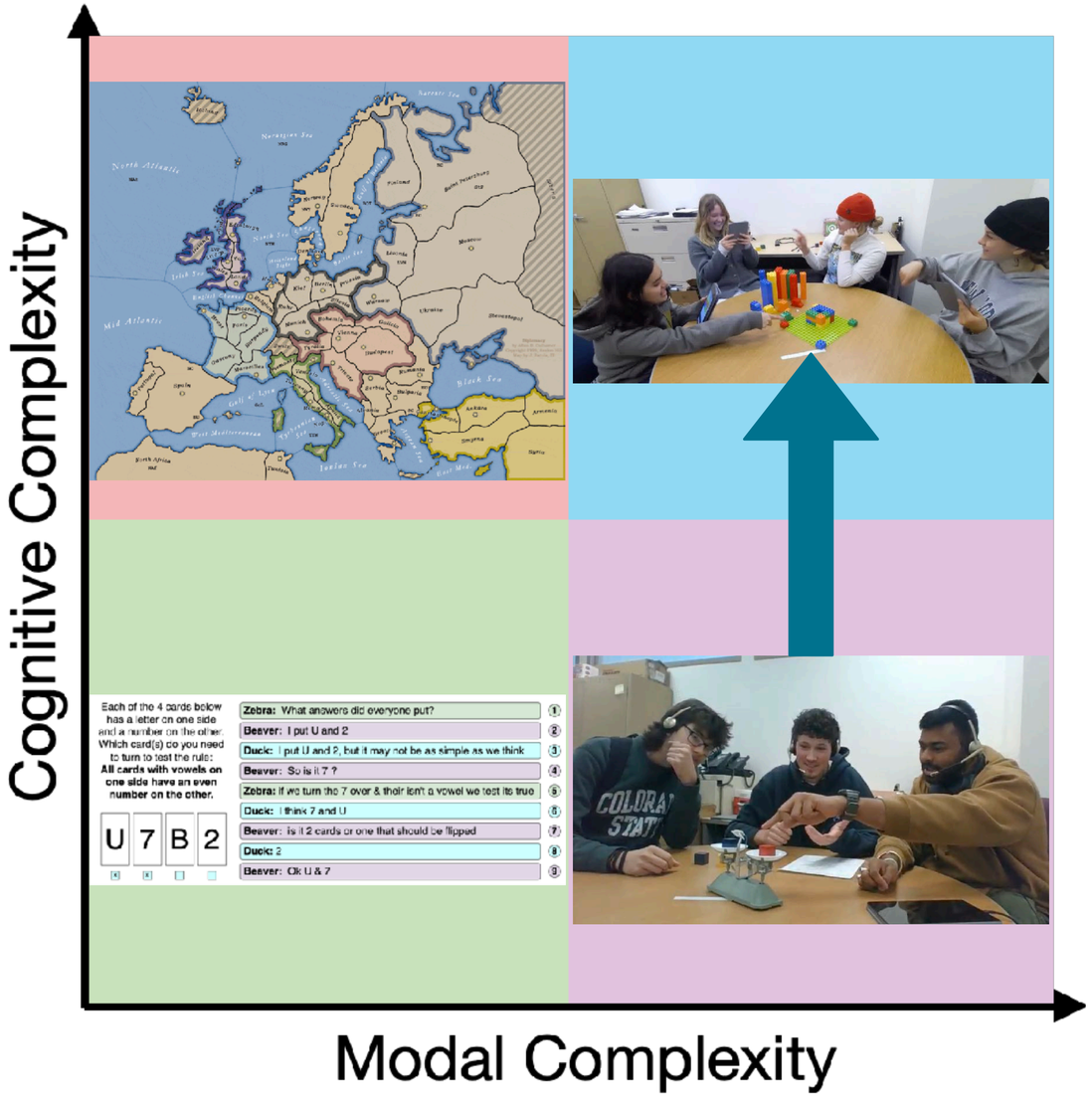

Can we consider the possibility that our assumption about purple being 20 might be off?
Unsolvable
PREDATOR

Future Directions

Distributed Partial Information Puzzles



Distributed Partial Information Puzzles



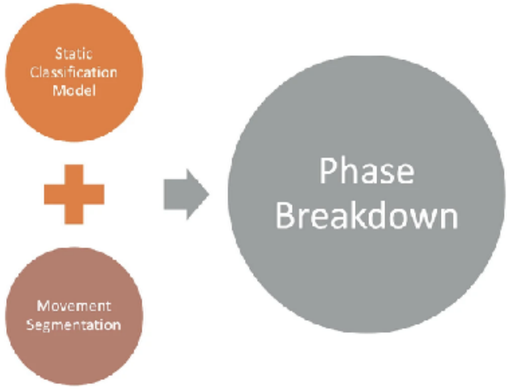
Distributed Partial Information Puzzles

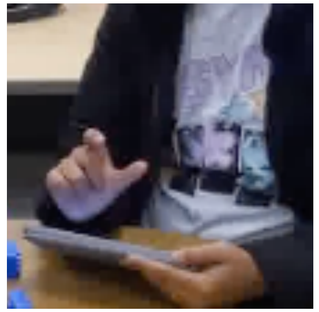
- **Problems with the Weights Task**
  - **Agreement/Disagreement:** Fewer opportunities for disagreement; task is well-structured, with clear solutions
  - **Complexity:** Lower cognitive and interpretive complexity; disagreements typically procedural or computational
  - **Reusability**: One a group knows the solution, they cannot organically reuse the task
- **DPIP Lego Task**: 3 *directors* with partial information guide one *builder* to construct a goal structure
  - Each director has a 2D view of one-side of the goal structure
  - Simulates team with different background knowledge/ expertise

Distributed Partial Information Puzzles

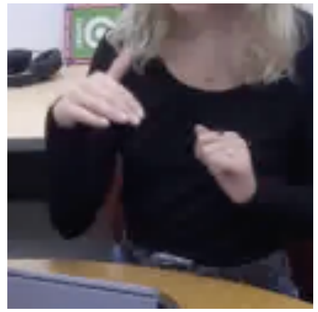
- **Identified gesture types**
- Characterized by hand pose/arm motion
- Representable in existing GAMR semantics
- Extractable using adaptation of existing gesture recognition pipeline

Side-by-side

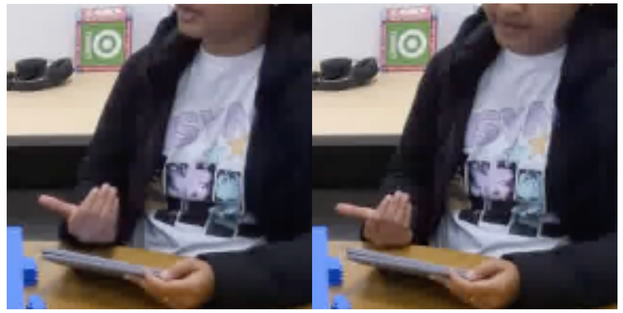




Square



Slope

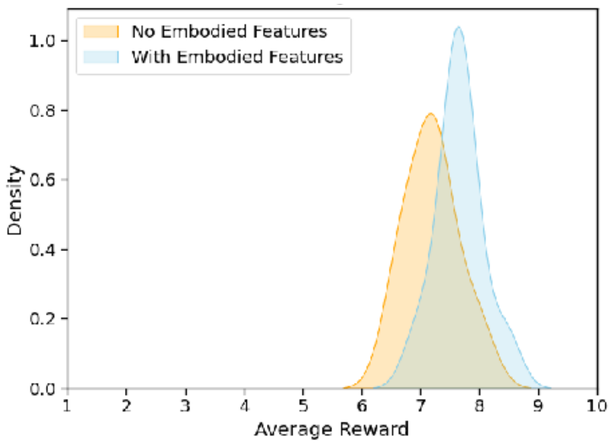


Bring forward/backward

Rotate

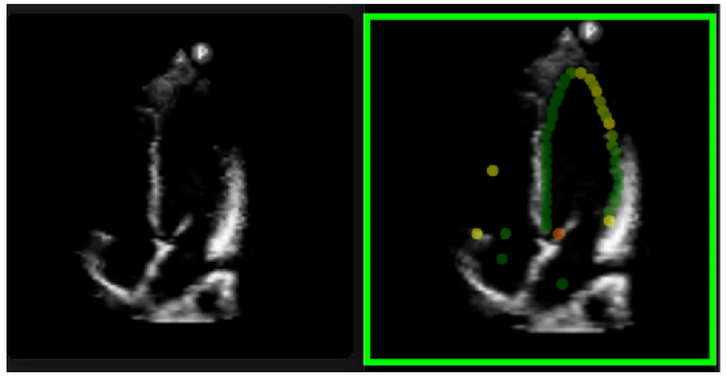
Intelligent Task Guidance

- Explicit guidance can also be viewed as friction
  - Intervention to catch mistakes before they are made
  - Human-AI collaboration in medical domain
- An LLM or VLM is not a guidance model
  - However, including embodied information in the context provides better quality interventions (higher rewards)

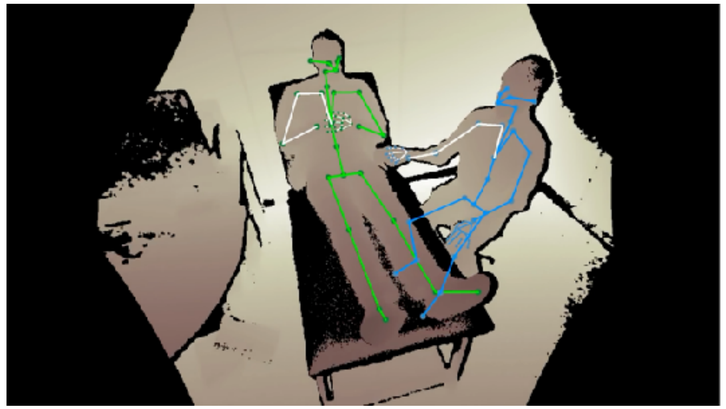


Intelligent Task Guidance

- Application: ultrasound scanning - how to acquire an image suitable for diagnosis
- How human experts do it: “[M]oving from novice to expert is a matter of practice and feedback. […] [C]ombination of the ultrasound technician’s and the radiologist's expertise.”
- Wide variability in preference and technique, even among experts
- Given a current image x, what action y should I take to improve it?
  - Maximize: $P(y_w \succ y_\ell \mid x)$
- **Your Reward Model is Secretly a Guidance Model**



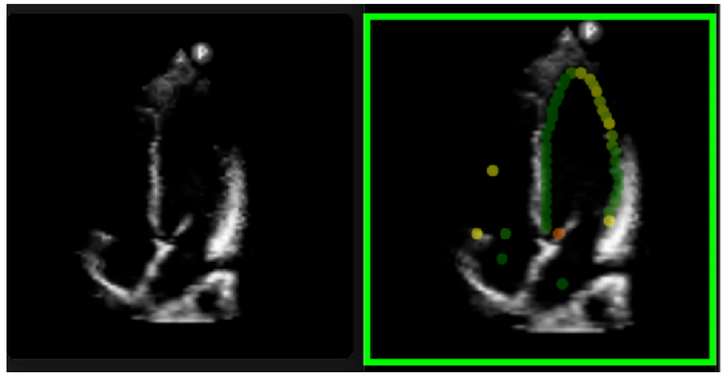
Heart scan



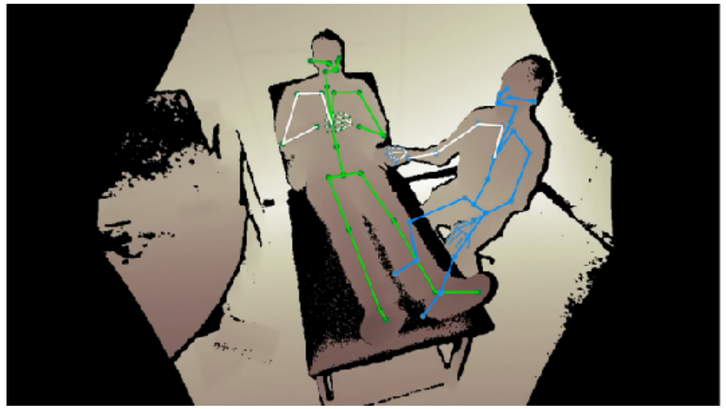
Lower limb scan

Intelligent Task Guidance

- Application: ultrasound scanning - how to acquire an image suitable for diagnosis
- How human experts do it: “[M]oving from novice to expert is a matter of practice and feedback. […] [C]ombination of the ultrasound technician’s and the radiologist's expertise.”
- Wide variability in preference and technique, even among experts
- Given a current image x, what action y should I take to improve it?
  - Maximize: $P(y_w \succ y_\ell \mid x)$
- **Your Reward Model is Secretly a Guidance Model**



Heart scan

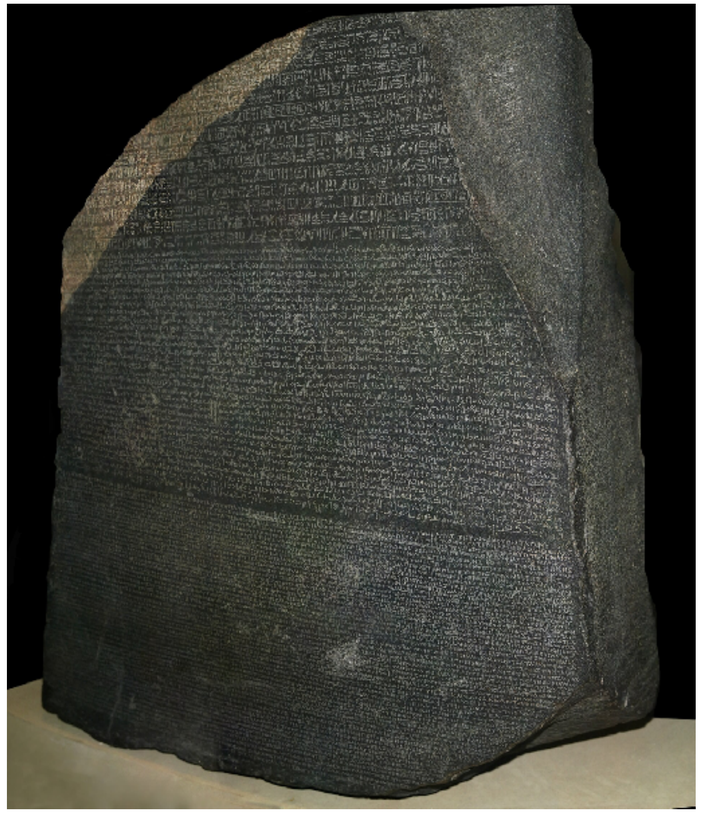


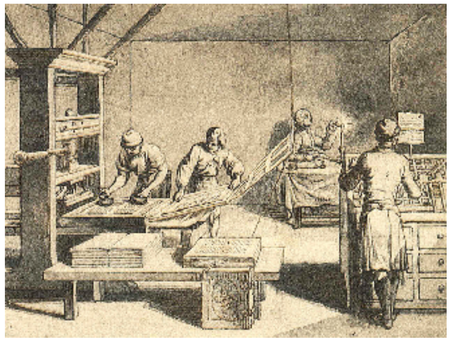
Lower limb scan

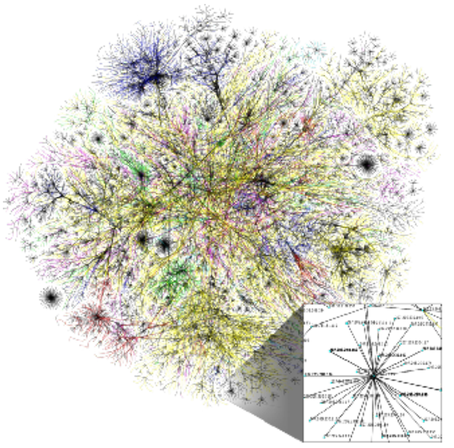
Conclusion

The Metaphor(s) of AI

- AI is like: 3 previous technologies that changed our relationship with truth

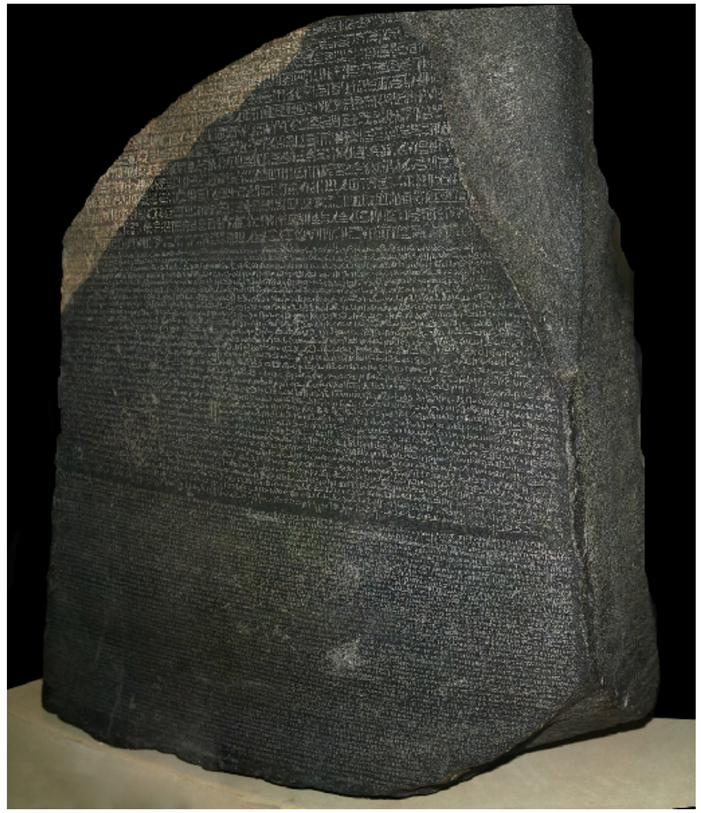


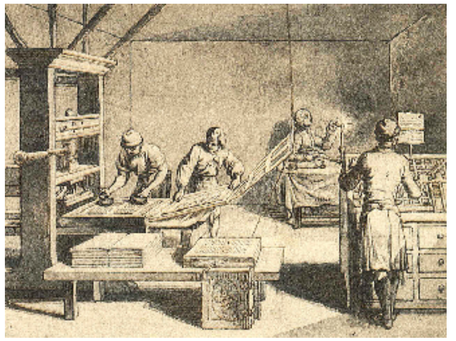


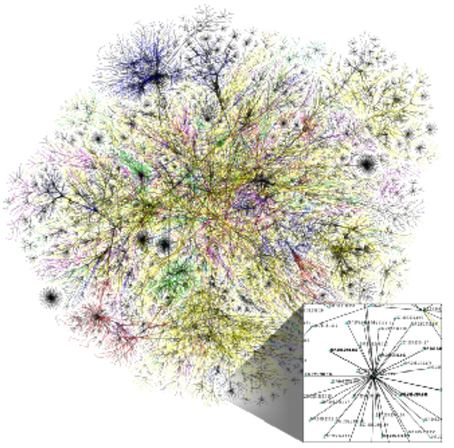


The Metaphor(s) of AI

- AI is like: 3 previous technologies that changed our relationship with truth
  - Writing

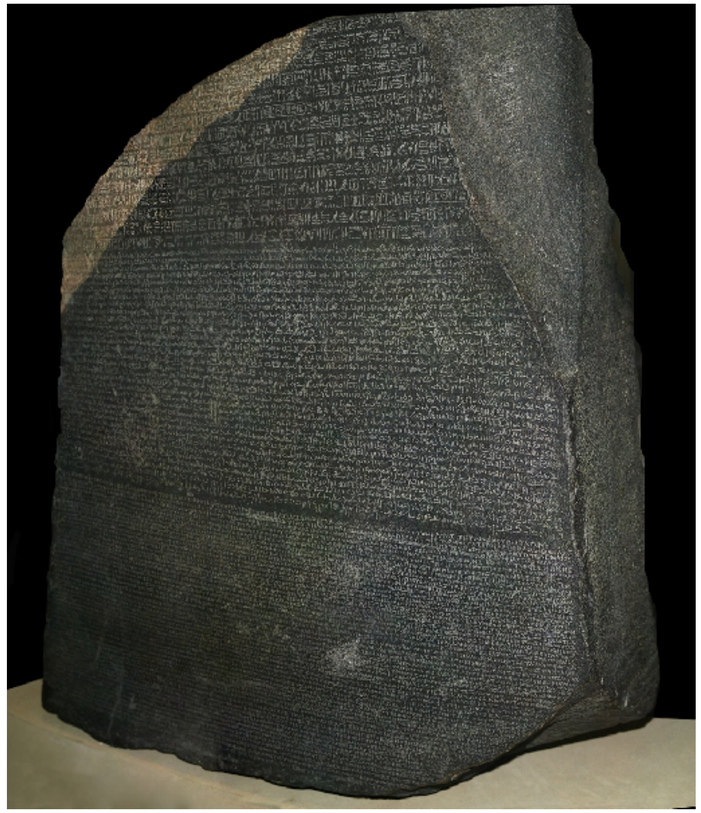


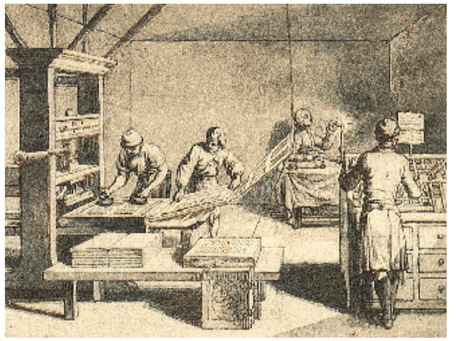


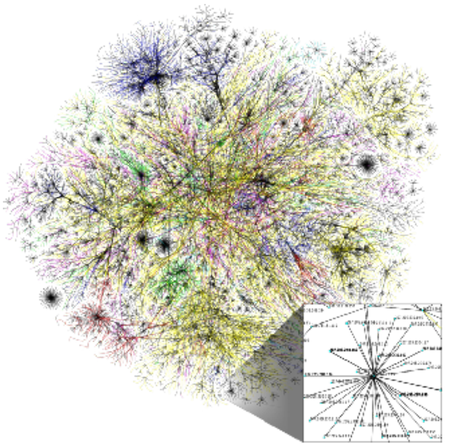


The Metaphor(s) of AI

- AI is like: 3 previous technologies that changed our relationship with truth
  - Writing
  - Printing press

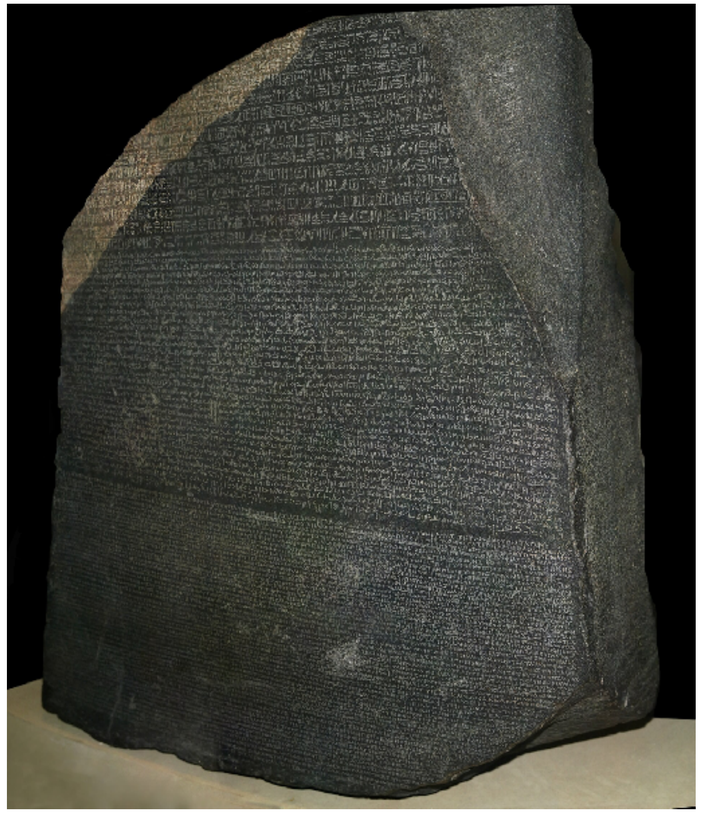


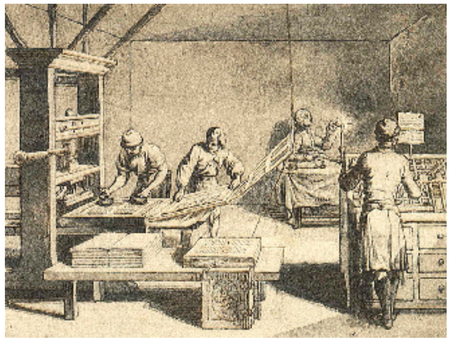


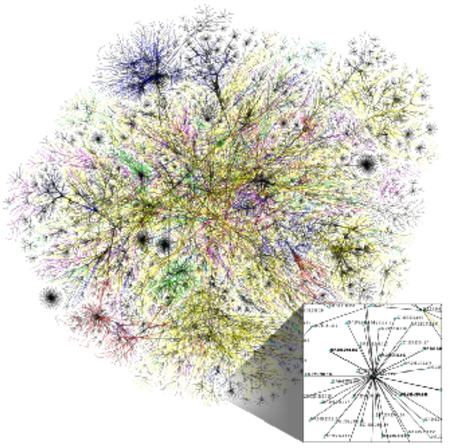


The Metaphor(s) of AI

- AI is like: 3 previous technologies that changed our relationship with truth
  - Writing
  - Printing press
  - Internet

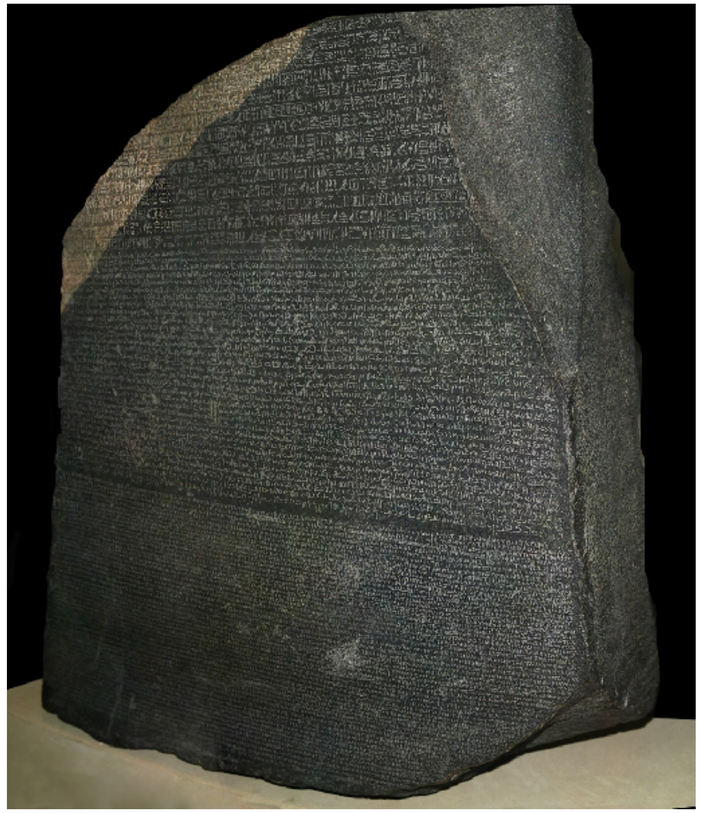


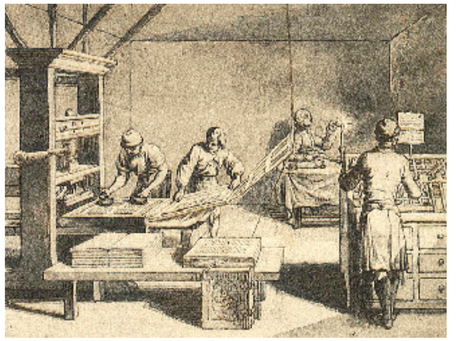


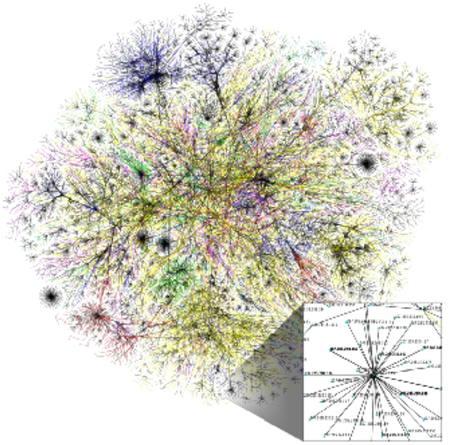


The Metaphor(s) of AI

- AI is like: 3 previous technologies that changed our relationship with truth
  - Writing
  - Printing press
  - Internet
- All 3 lowered barriers to entry, and changed how humans processed information

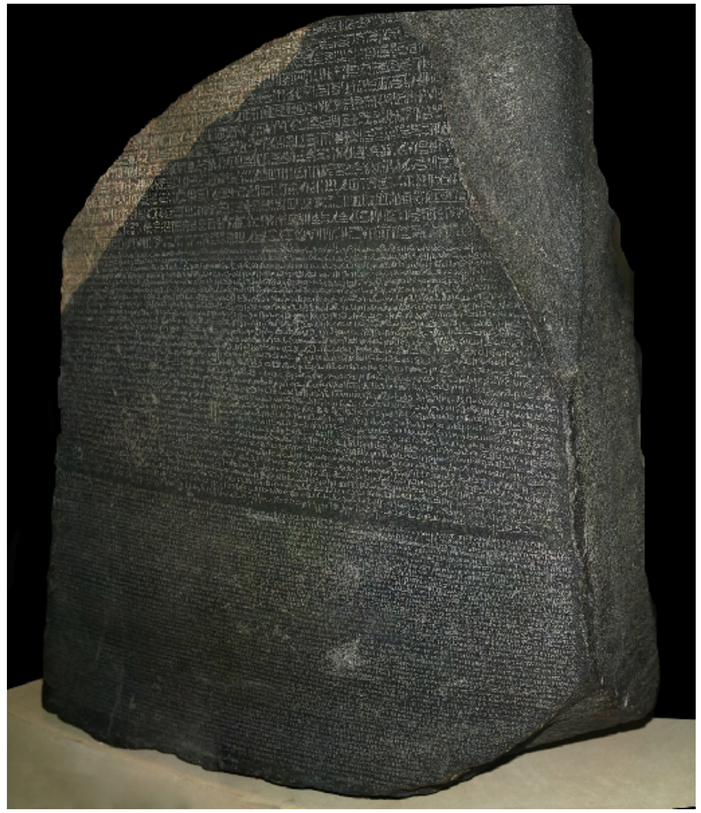


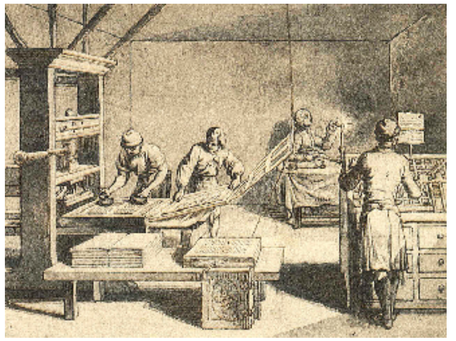


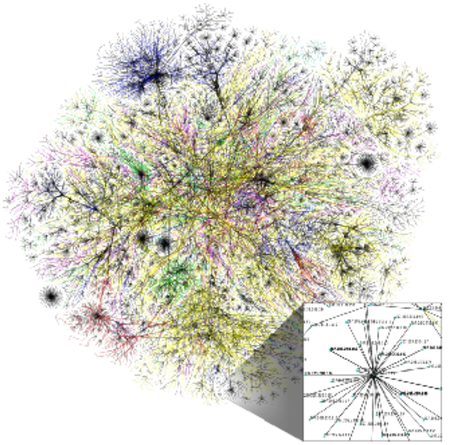


The Metaphor(s) of AI

- AI is like: 3 previous technologies that changed our relationship with truth
  - Writing
  - Printing press
  - Internet
- All 3 lowered barriers to entry, and changed how humans processed information
- Democratized access to information and audience

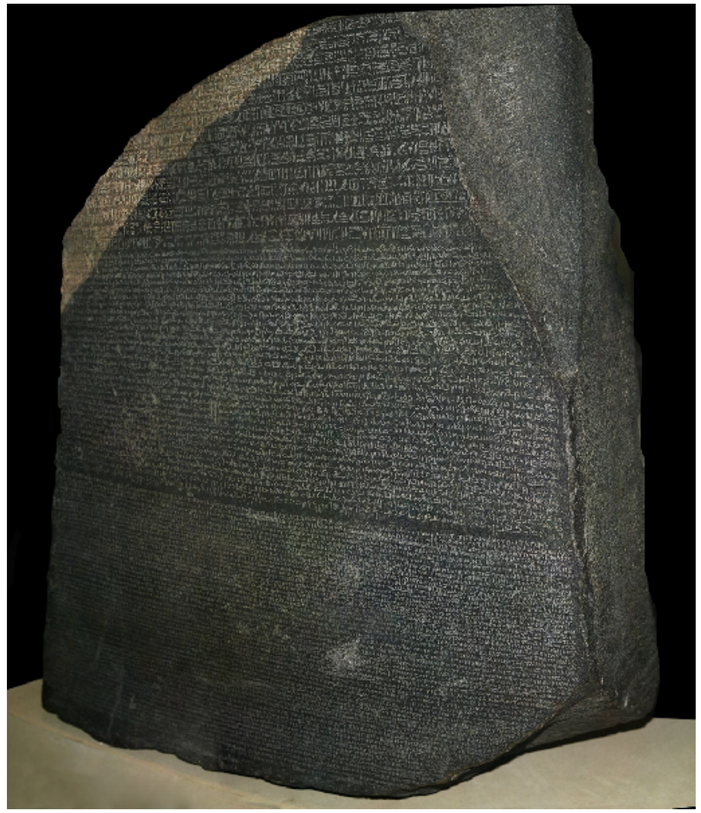


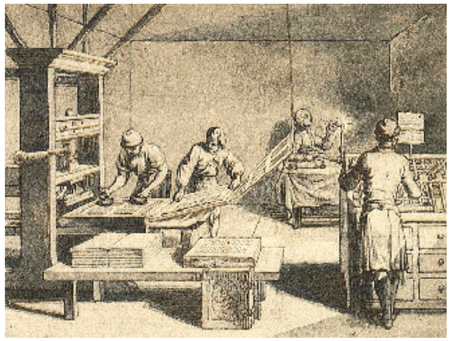


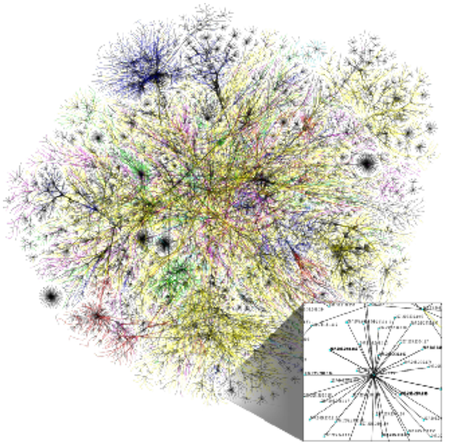


The Metaphor(s) of AI

- AI is like: 3 previous technologies that changed our relationship with truth
  - Writing
  - Printing press
  - Internet
- All 3 lowered barriers to entry, and changed how humans processed information
- Democratized access to information and audience
- LLMs: all this, and incentivized for instant access, responsiveness, prompting overreliance







Conclusion

Conclusion

- Many people are rapidly coming to rely on LLMs/GenAI technologies

Conclusion

- Many people are rapidly coming to rely on LLMs/GenAI technologies
- Overreliance on LLMs is beginning to show negative effects on cognition and outcomes

Conclusion

- Many people are rapidly coming to rely on LLMs/GenAI technologies
- Overreliance on LLMs is beginning to show negative effects on cognition and outcomes
- “Friction” mechanism repositions LLMs/GenAI as collaborative “thought partners”

Conclusion

- Many people are rapidly coming to rely on LLMs/GenAI technologies
- Overreliance on LLMs is beginning to show negative effects on cognition and outcomes
- “Friction” mechanism repositions LLMs/GenAI as collaborative “thought partners”
    - Creates opportunity for negotiation of intents toward shared goals, space for accountability, collaborative reasoning

Conclusion

- Many people are rapidly coming to rely on LLMs/GenAI technologies
- Overreliance on LLMs is beginning to show negative effects on cognition and outcomes
- “Friction” mechanism repositions LLMs/GenAI as collaborative “thought partners”
  - Creates opportunity for negotiation of intents toward shared goals, space for accountability, collaborative reasoning
  - May result in net slower interaction, but are critical to task success

Conclusion

- Many people are rapidly coming to rely on LLMs/GenAI technologies
- Overreliance on LLMs is beginning to show negative effects on cognition and outcomes
- “Friction” mechanism repositions LLMs/GenAI as collaborative “thought partners”
  - Creates opportunity for negotiation of intents toward shared goals, space for accountability, collaborative reasoning
  - May result in net slower interaction, but are critical to task success
- From friction to guidance: wield similar alignment techniques to reformulate action generation as step prompting, then render guidance appropriately

Conclusion

- Many people are rapidly coming to rely on LLMs/GenAI technologies
- Overreliance on LLMs is beginning to show negative effects on cognition and outcomes
- “Friction” mechanism repositions LLMs/GenAI as collaborative “thought partners”
  - Creates opportunity for negotiation of intents toward shared goals, space for accountability, collaborative reasoning
  - May result in net slower interaction, but are critical to task success
- From friction to guidance: wield similar alignment techniques to reformulate action generation as step prompting, then render guidance appropriately
- Research should proceed with an understanding that collaborative *process* is as important as outcome

Workshop on Overreliance and Accountability

https://origen-workshop.github.io

Workshop at COLM in October!

Submissions due **June 27** (but stay tuned)

COLORADO STATE UNIVERSITY

Tack så mycket!

nkrishna@colostate.edu

https://www.signallab.ai