(Image credit: ChatGPT)

Beyond Logic and Math: Probabilistic Reasoning for Real-World Decision Making

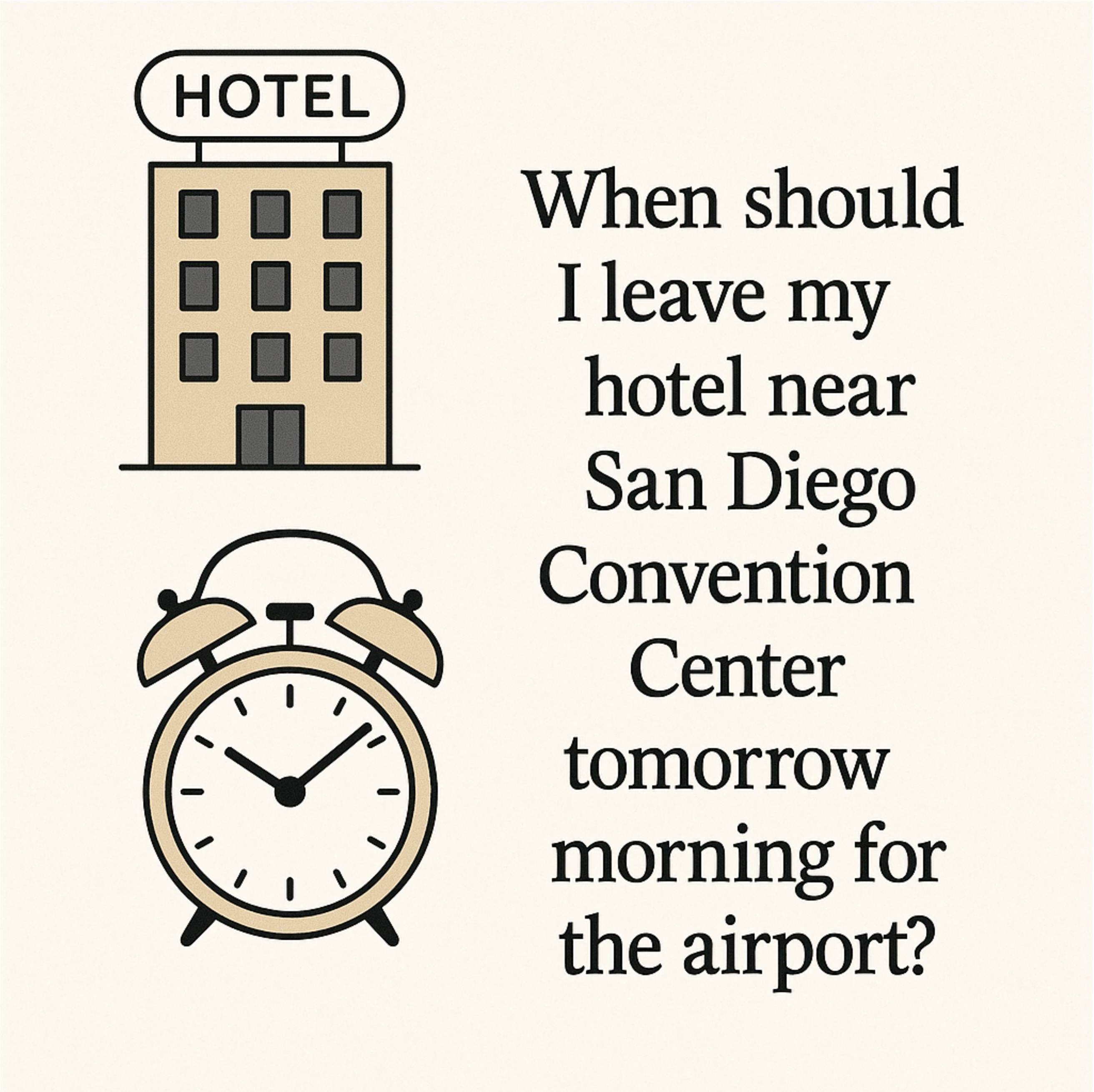
Wei Xu (associate professor)
College of Computing
Georgia Institute of Technology
Twitter/X @cocoweixu

Georgia Tech

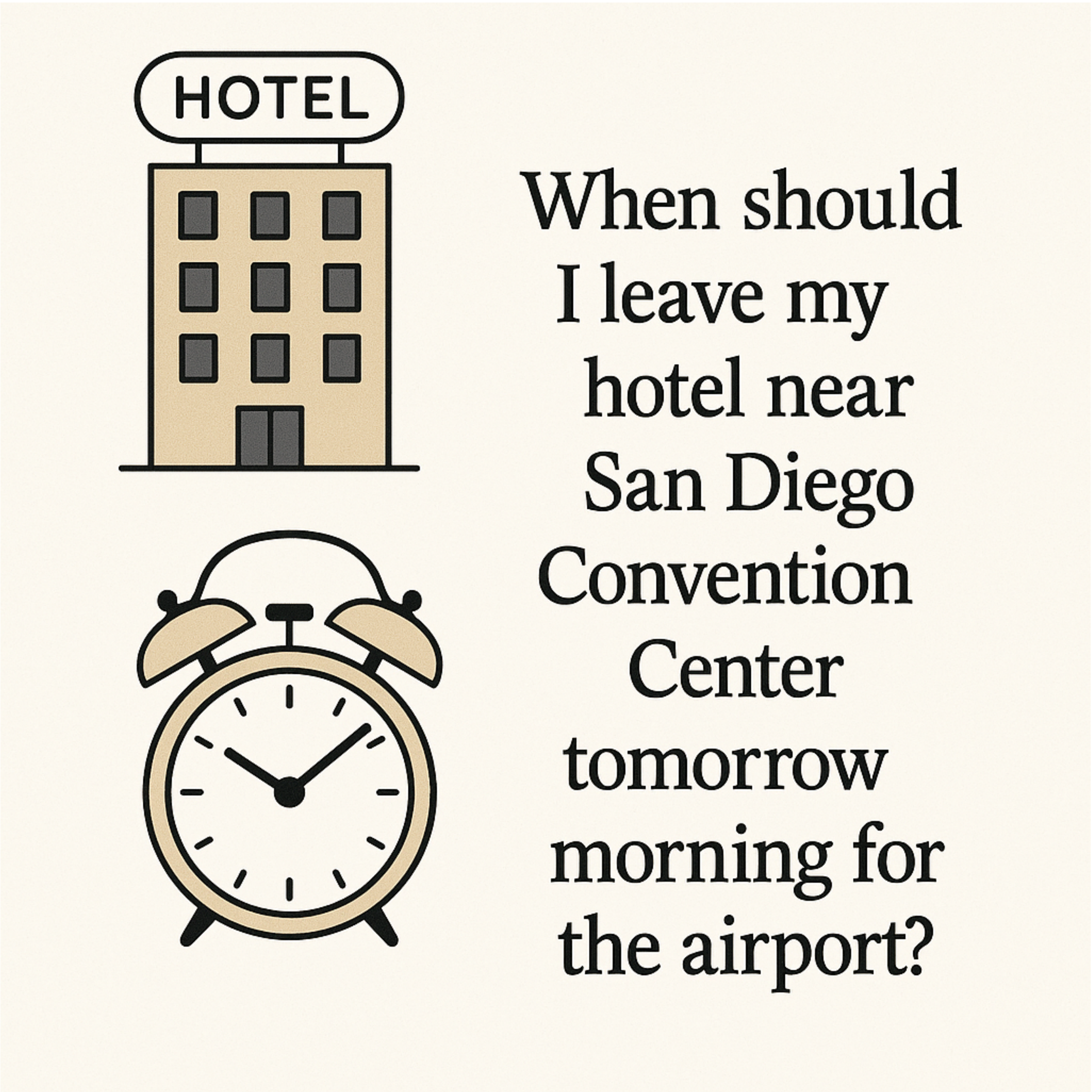
Efficient Reasoning Workshop @ NeurIPS (Dec 6th, 2025)

So … what did you reason about today?

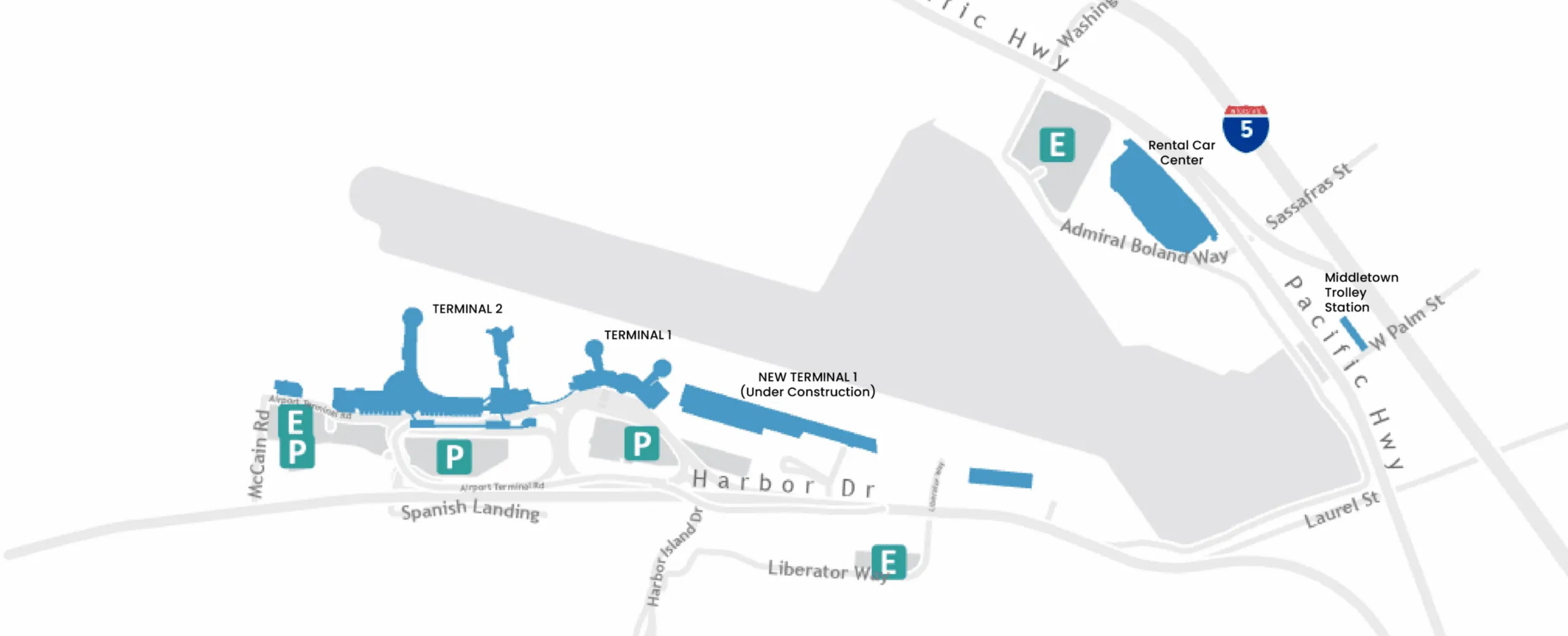
So … what did you reason about today?



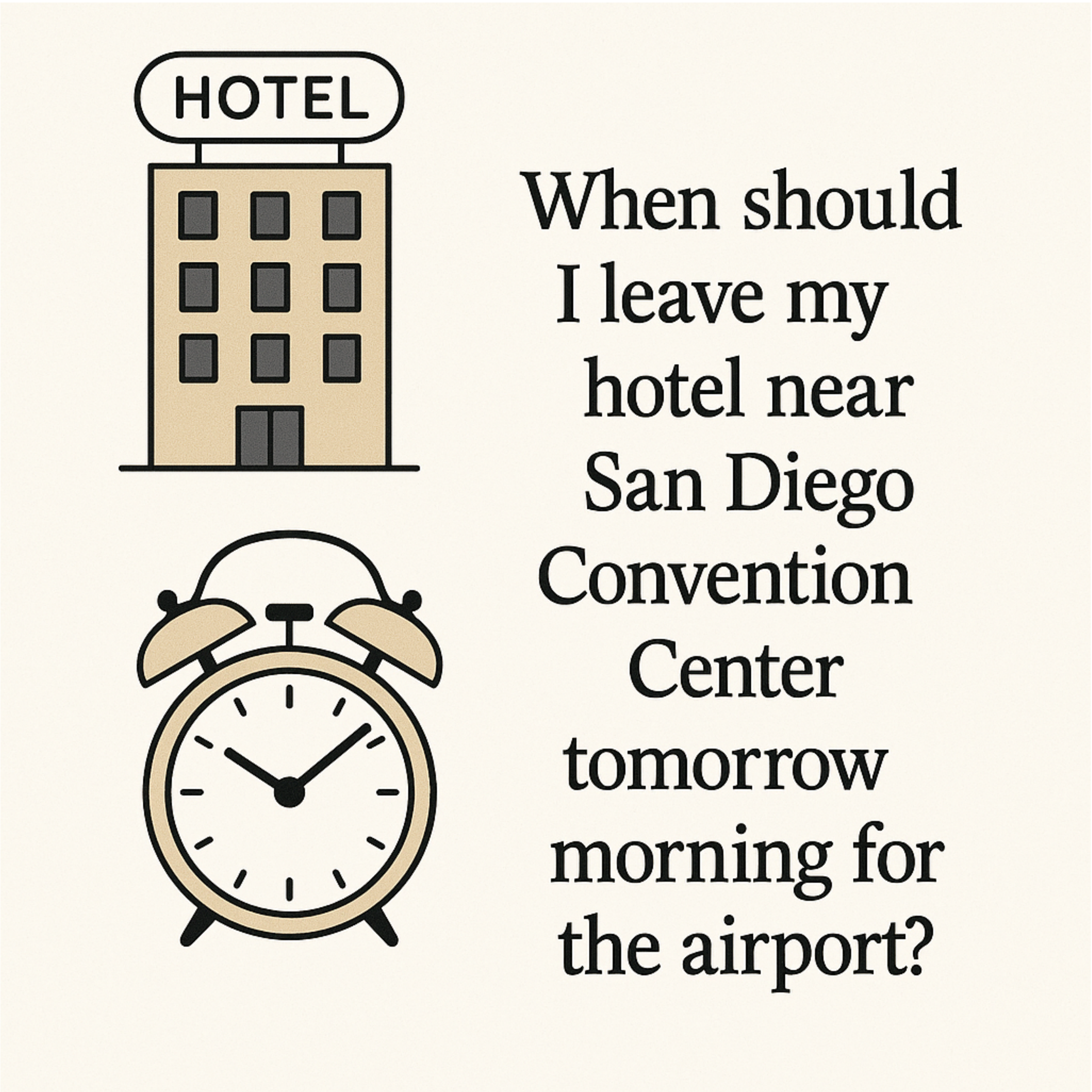
So … what did you reason about today?



- 15 minutes drive without traffic
- not a major holiday, but 20k+ attendants from NeurIPS
- will it take long to get a ride share?
- will it take long to check out the hotel?
- not too big an airport (500-600 flights daily)
  in comparison to Atlanta (2000+ flights daily)

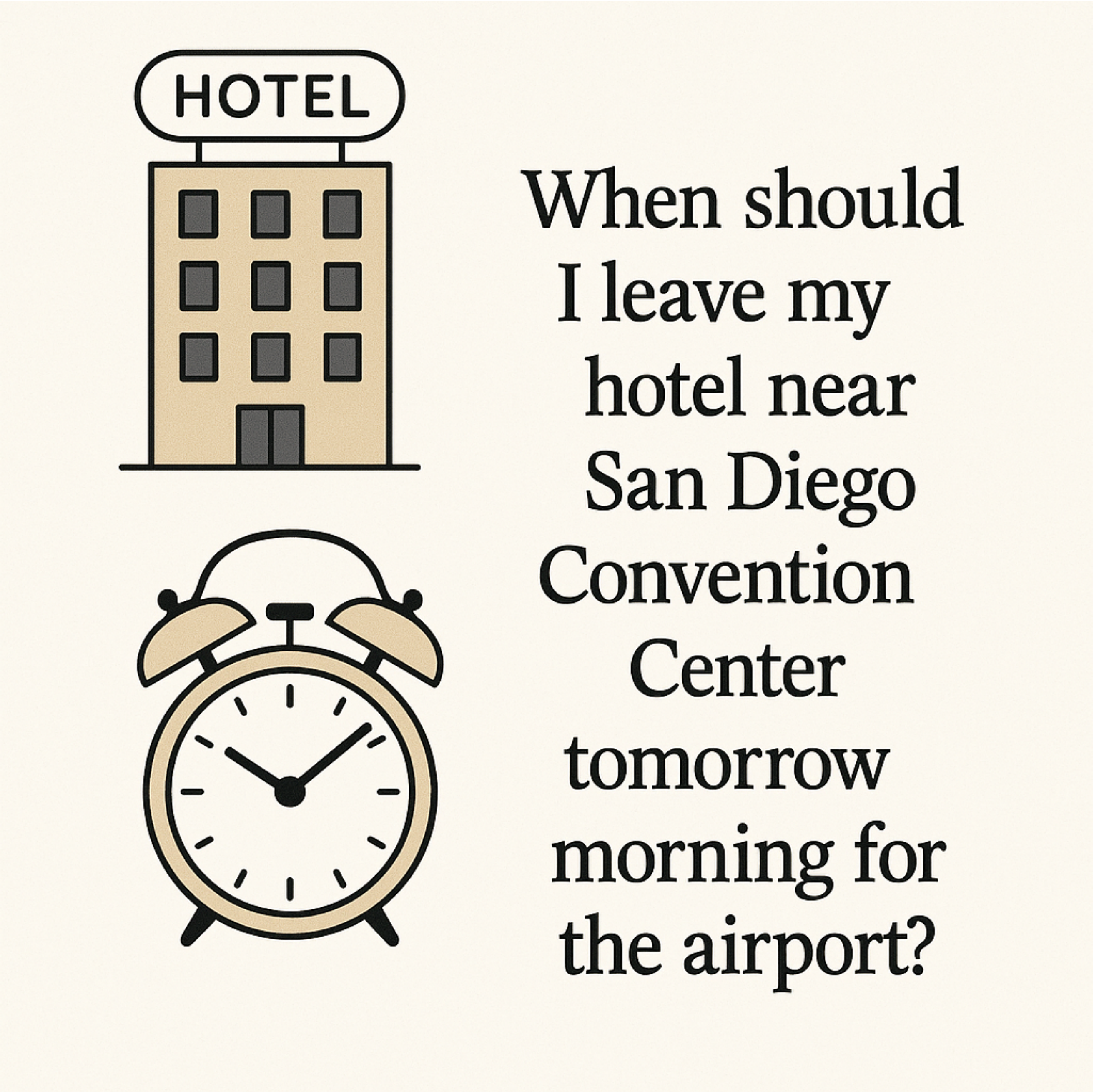


This vs. Current LLM reasoning benchmarks

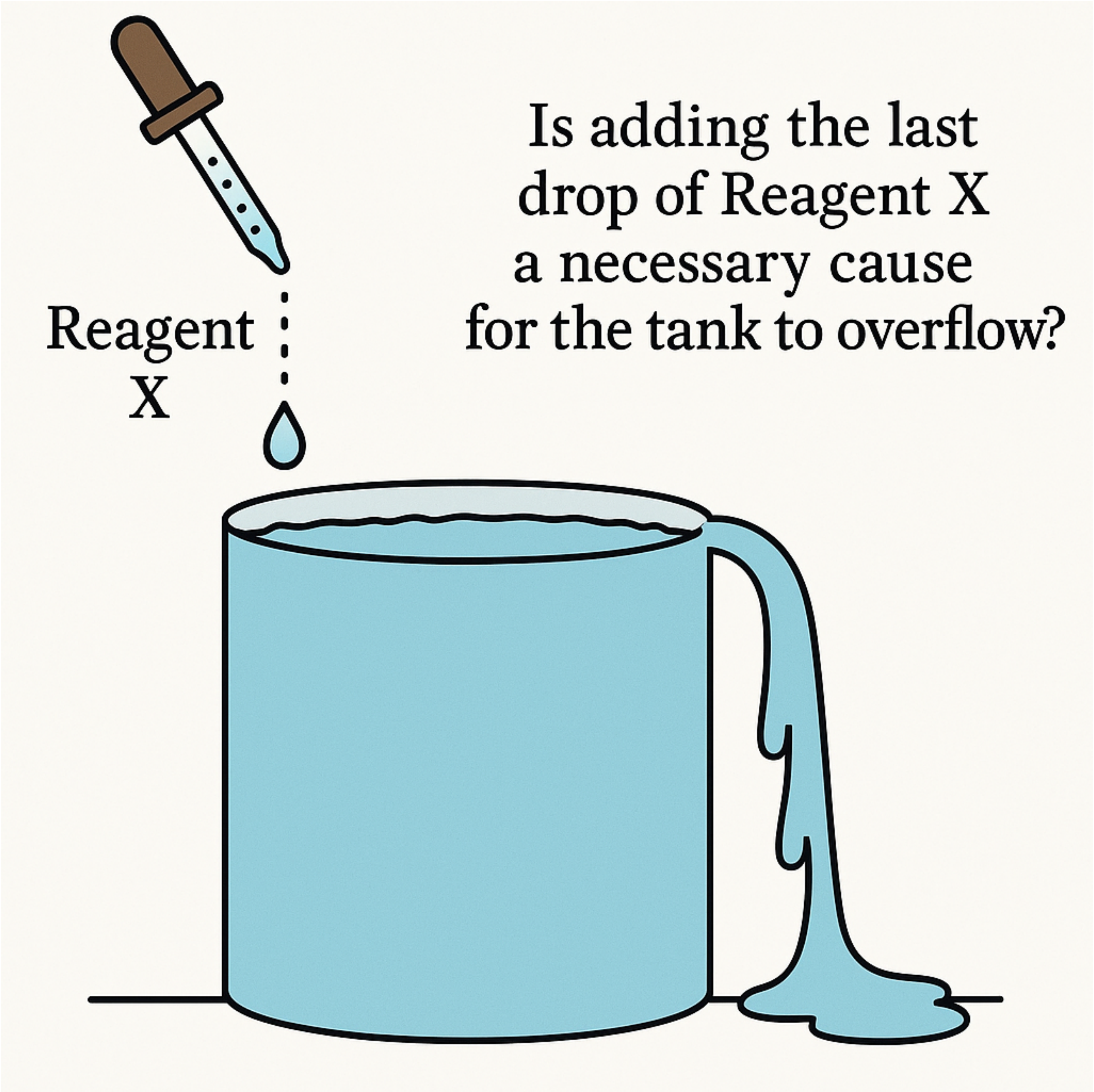


Your Daily Reasoning
(as a normal human being)

This vs. Current LLM reasoning benchmarks

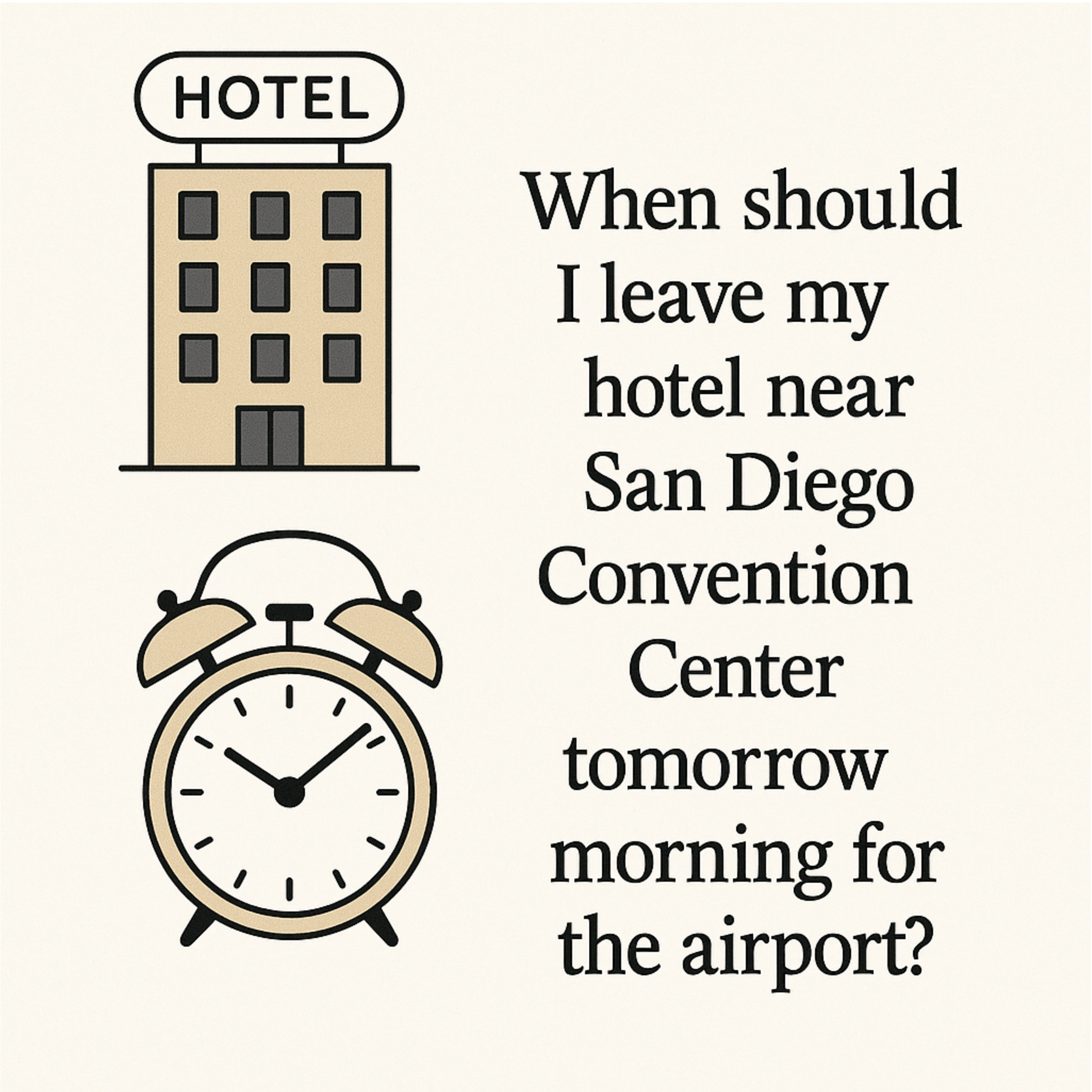


Your Daily Reasoning
(as a normal human being)

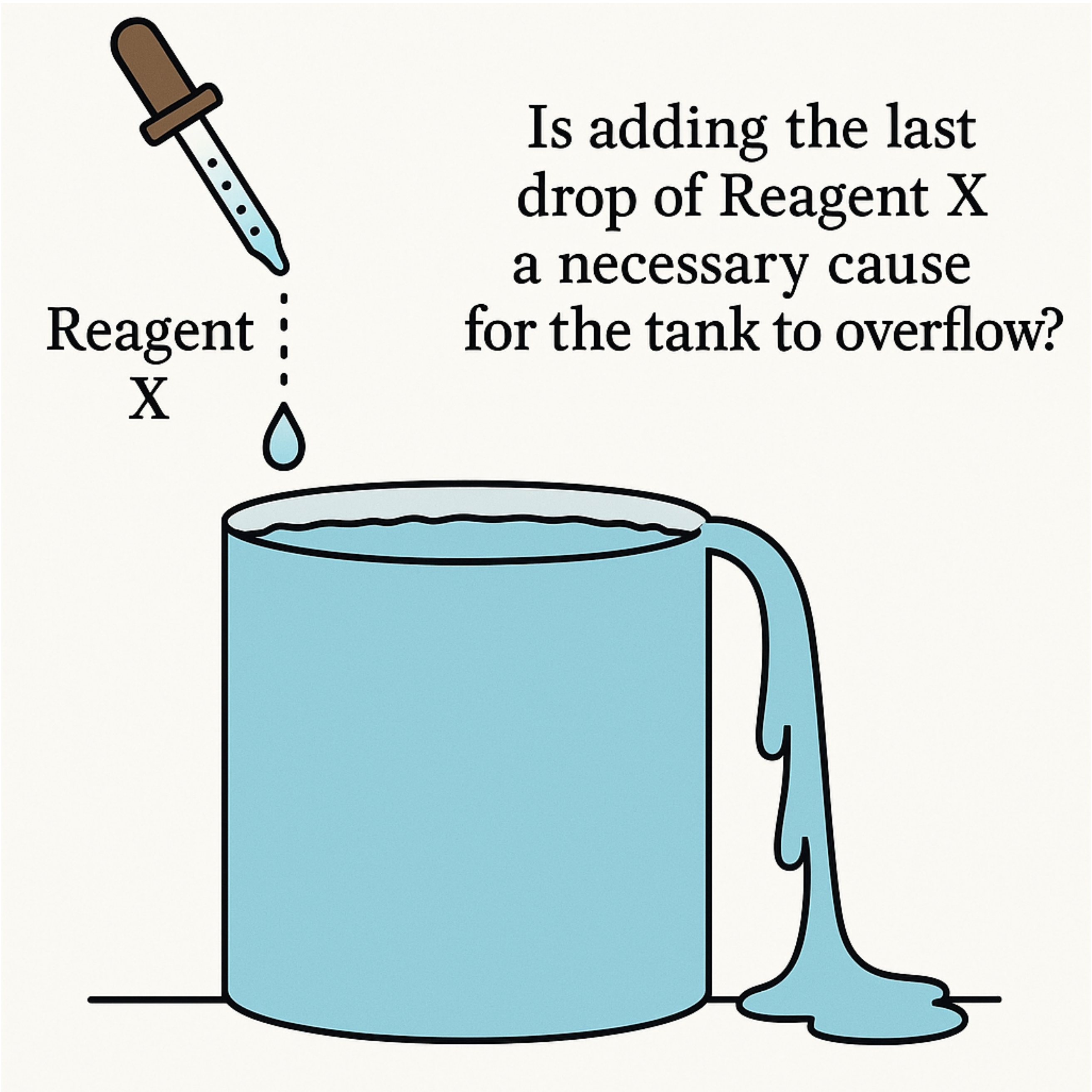


BIG-Bench Extra Hard
(used in OLMo3 evaluation)

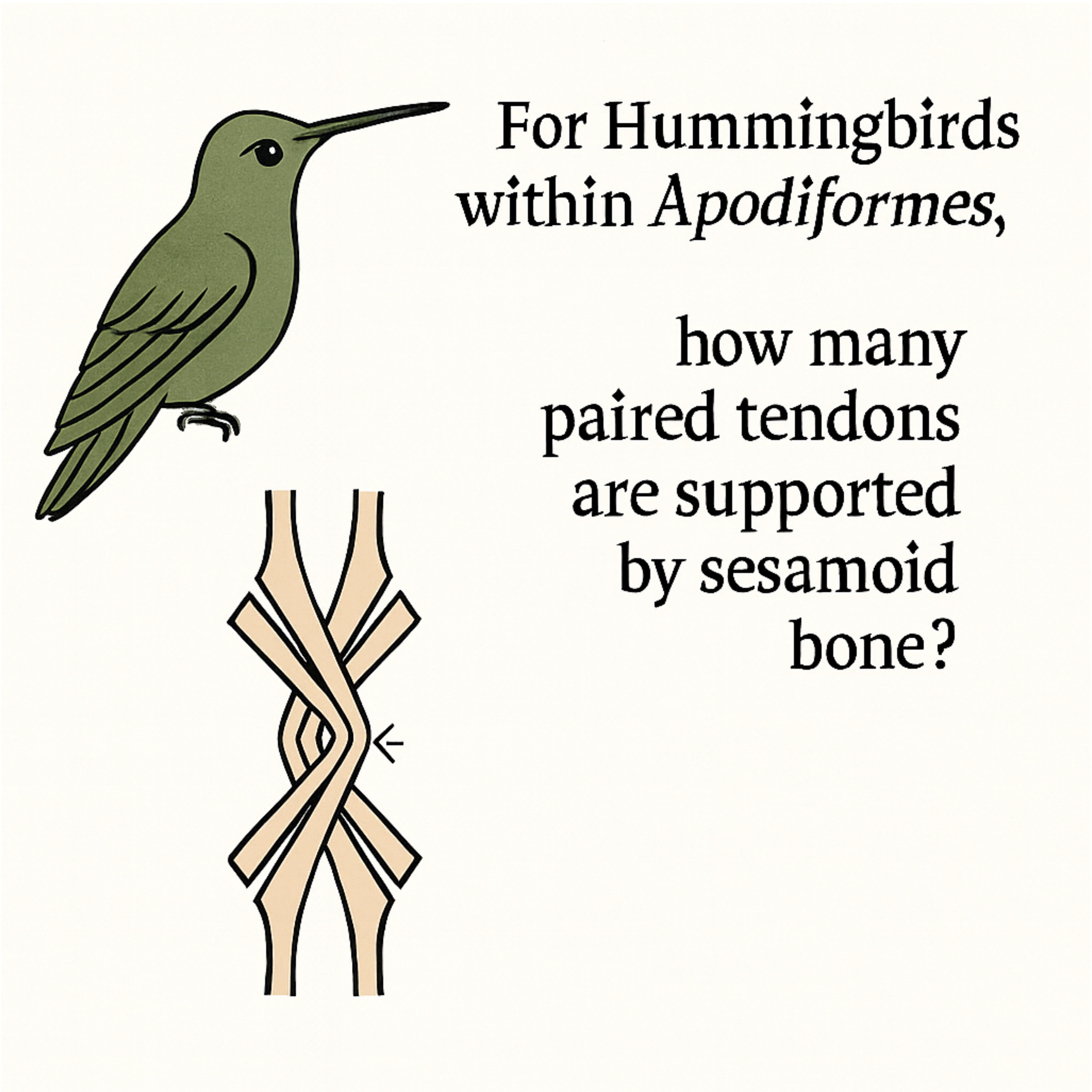
This vs. Current LLM reasoning benchmarks



Your Daily Reasoning
(as a normal human being)

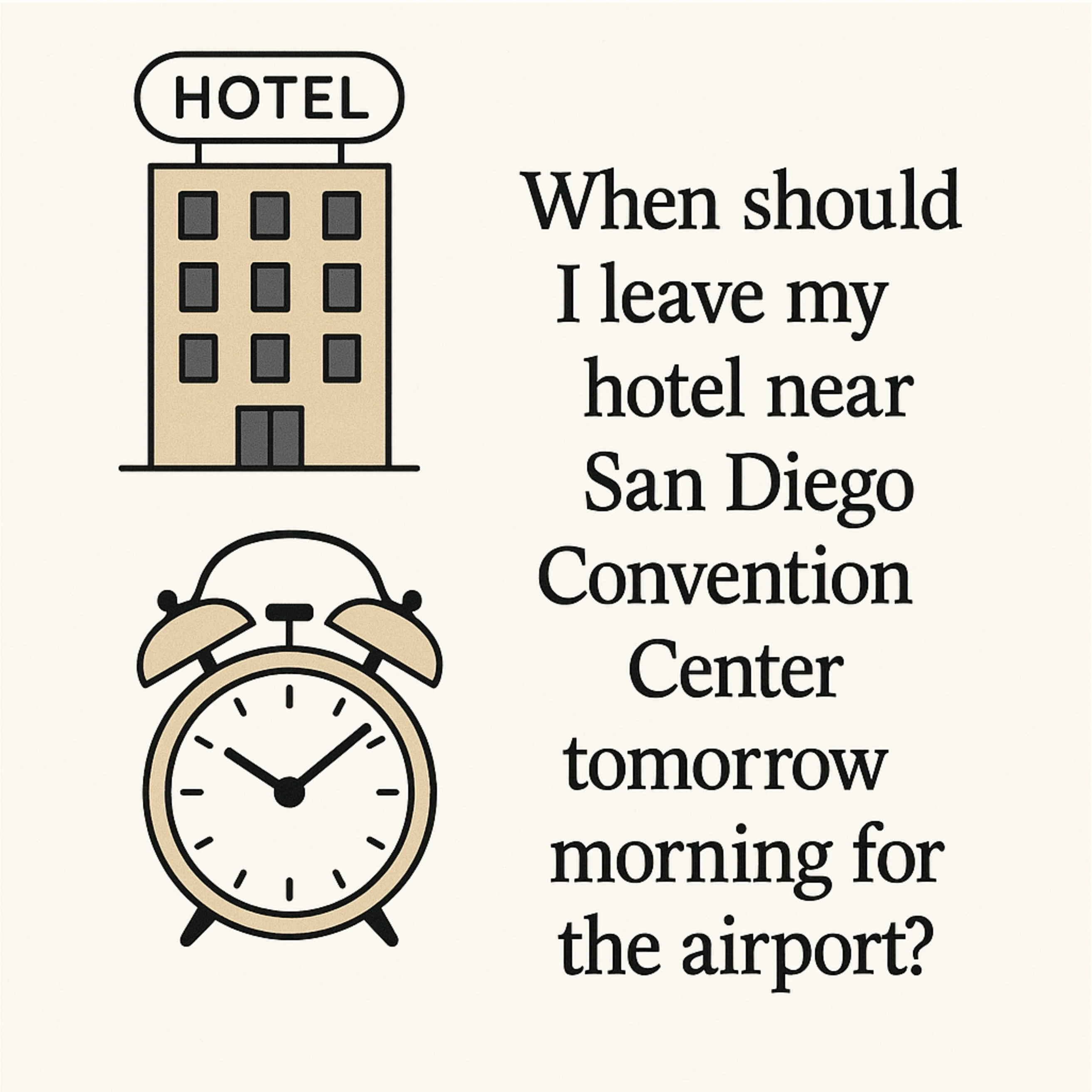


BIG-Bench Extra Hard
(used in OLMo3 evaluation)



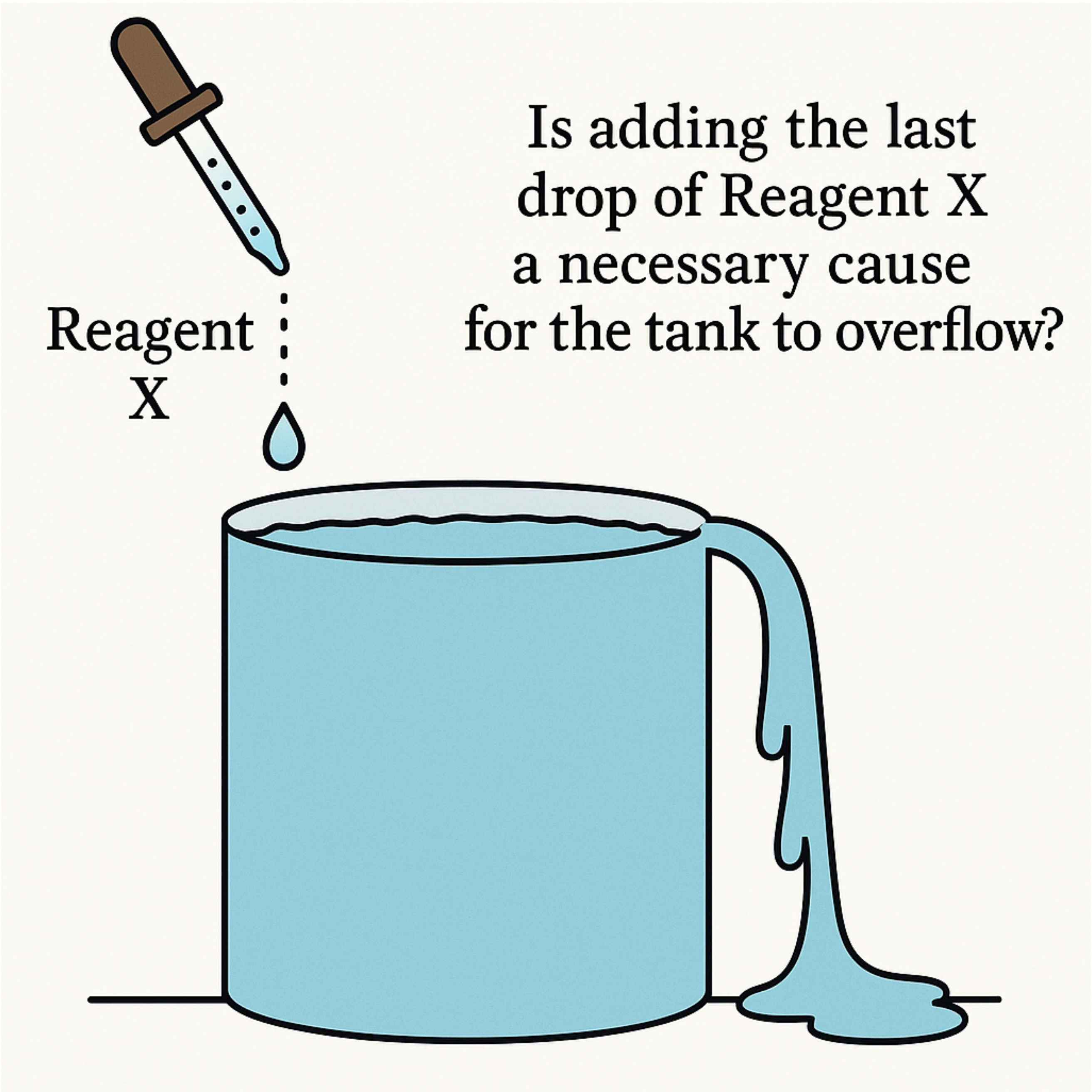
Humanity's Last Exam
(used in Gemini3 evaluation)

What's the difference?



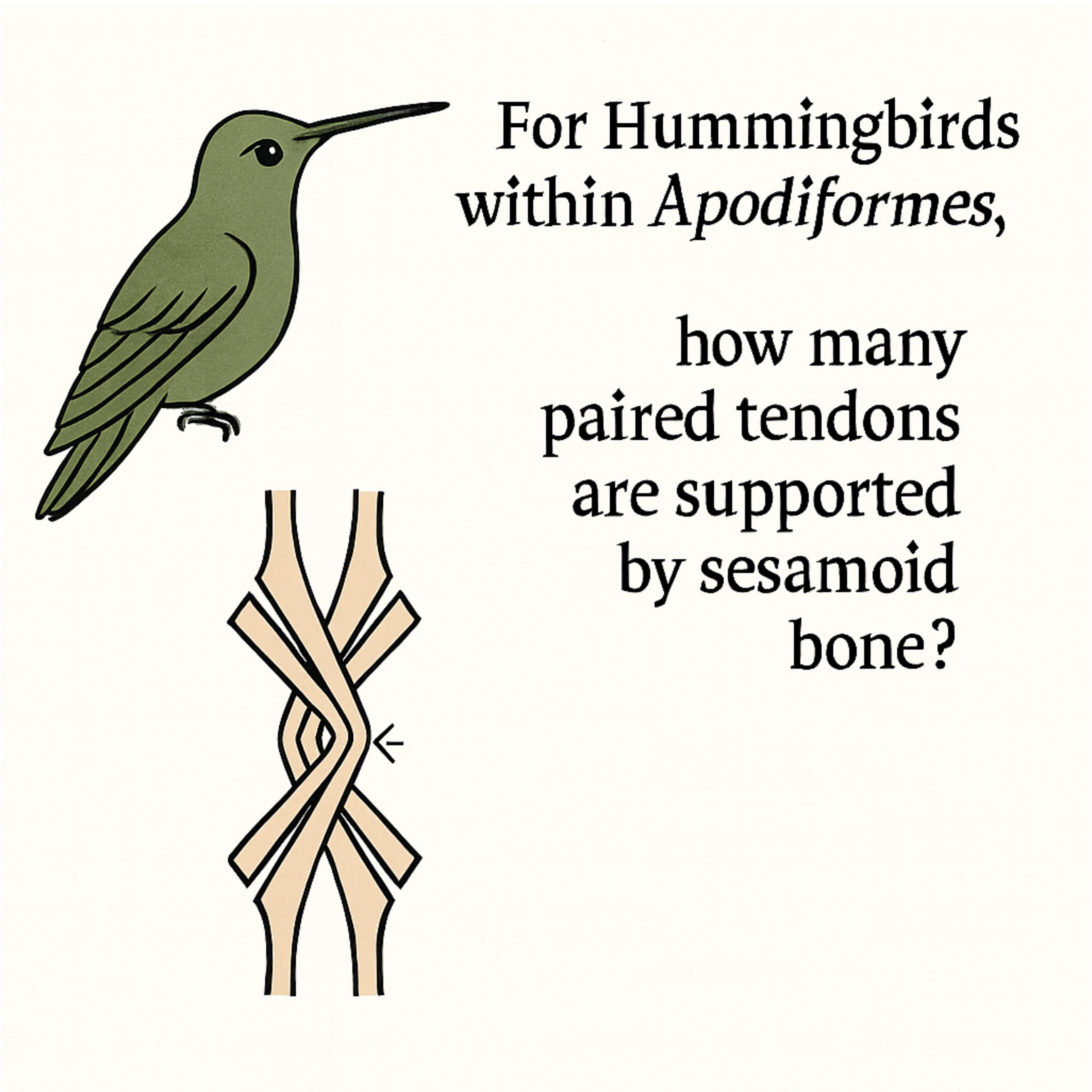
Your Daily Reasoning
(as a normal human being)

Numerical, with some uncertainty



BIG-Bench Extra Hard
(used in OLMo3 evaluation)

Math/Logic tests on toy examples



Humanity's Last Exam
(used in Gemini3 evaluation)

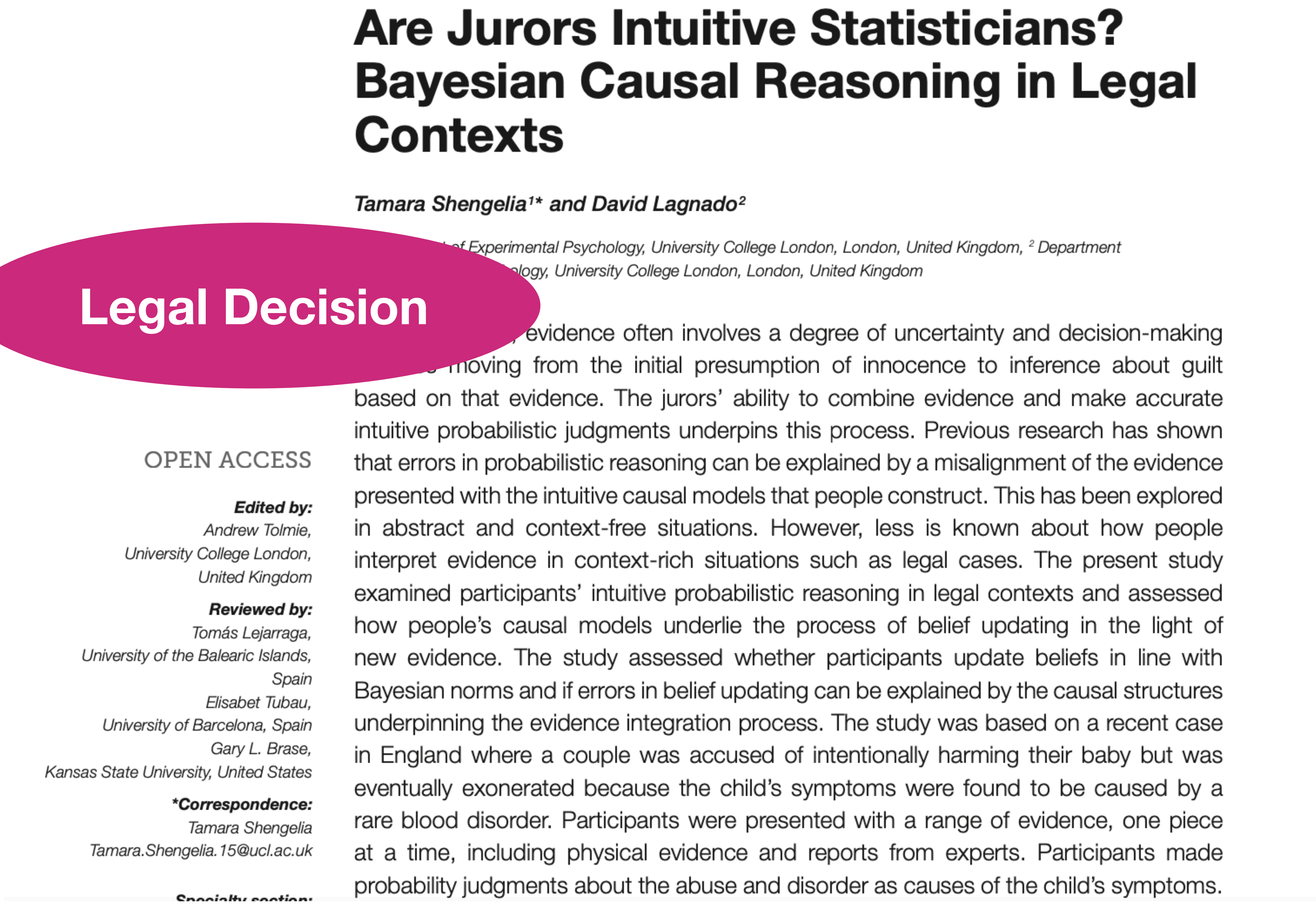
Numerical, but deterministic

Probabilistic Reasoning

Reasoning that deals with uncertainty by using probability theory to assess likelihoods, draw inferences, and make decisions, unlike deterministic logic.

e.g., predicting product success
How many purses in each color to manufacture?

Marketing



Are Jurors Intuitive Statisticians? Bayesian Causal Reasoning in Legal Contexts

Tamara Shengelia[1]* and David Lagnado[2]

[1] ... of Experimental Psychology, University College London, London, United Kingdom, [2] Department ... ology, University College London, London, United Kingdom

... evidence often involves a degree of uncertainty and decision-making ... moving from the initial presumption of innocence to inference about guilt based on that evidence. The jurors' ability to combine evidence and make accurate intuitive probabilistic judgments underpins this process. Previous research has shown that errors in probabilistic reasoning can be explained by a misalignment of the evidence presented with the intuitive causal models that people construct. This has been explored in abstract and context-free situations. However, less is known about how people interpret evidence in context-rich situations such as legal cases. The present study examined participants' intuitive probabilistic reasoning in legal contexts and assessed how people's causal models underlie the process of belief updating in the light of new evidence. The study assessed whether participants update beliefs in line with Bayesian norms and if errors in belief updating can be explained by the causal structures underpinning the evidence integration process. The study was based on a recent case in England where a couple was accused of intentionally harming their baby but was eventually exonerated because the child's symptoms were found to be caused by a rare blood disorder. Participants were presented with a range of evidence, one piece at a time, including physical evidence and reports from experts. Participants made probability judgments about the abuse and disorder as causes of the child's symptoms.

OPEN ACCESS

Edited by:
Andrew Tolmie,
University College London,
United Kingdom

Reviewed by:
Tomás Lejarraga,
University of the Balearic Islands,
Spain
Elisabet Tubau,
University of Barcelona, Spain
Gary L. Brase,
Kansas State University, United States

***Correspondence:**
Tamara Shengelia
Tamara.Shengelia.15@ucl.ac.uk

Legal Decision

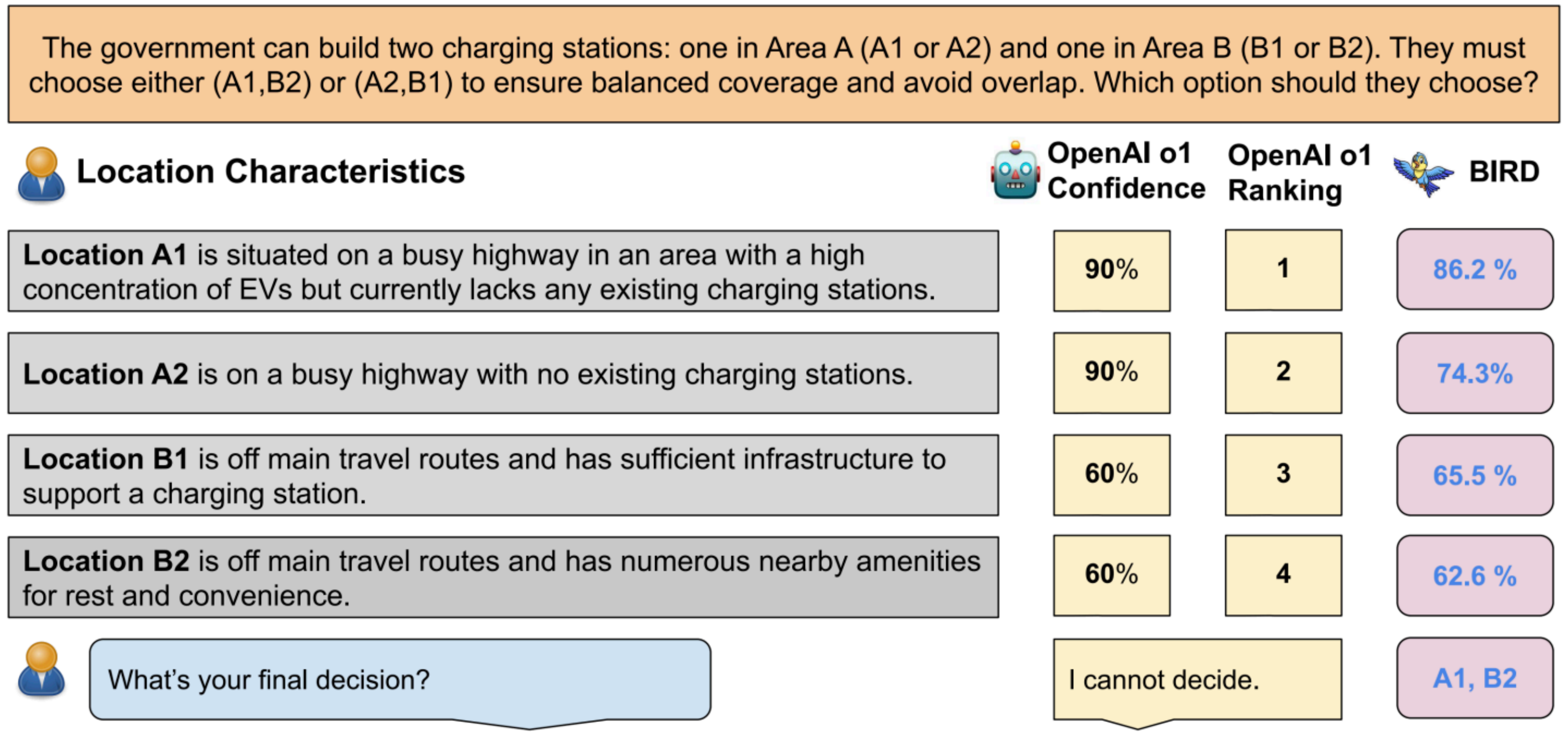
Medical Diagnosis

Why So Few LLM Probabilistic Reasoning Benchmarks?

- Hard to collect "ground-truth" for many decision making or predicting scenarios

"Should I open business at location A or B?"

- A lot of such data is in private collection
- Evaluation is inherently challenging, given the uncertainty



BIRD (Feng, Zhou, Lin, Roth; ICLR 2025)

LLM Probabilistic Reasoning Benchmark: Design Goals

1. Realistic — grounded in everyday scenarios that humans naturally reason about

2. Evaluable — uses public data where sufficiently reliable ground truth can be collected

3. Socially Meaningful — not just for publishing papers, but helpful for real users

We found one use case that fits all 3 criteria!

NEURAL INFORMATION
PROCESSING SYSTEMS
2025

Probabilistic Reasoning with LLMs for k-anonymity Estimation


Jonathan Zheng


Alan Ritter


Sauvik Das


Wei Xu

Georgia Tech

Carnegie Mellon University

How much privacy risk is posed to an reddit user?

i.e., how many (*K*) people in the entire world fit this description?

r/Townsville

Does Townsville have the highest inflation in the entire country?

Been here 20 years. I work in Tech, but $10 for eggs is ridiculous! Luckily, I don't have to deal with landlords and increasing rent.

My son's daycare also recently increased their rate. I only have 4 months of maternity leave, so I'm looking for affordable childcare options in the area.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, Wei Xu. "Reducing Privacy Risks in Online Self-Disclosures with Language Models" (ACL 2024)

How much privacy risk is posed to an reddit user?

i.e., how many (*K*) people in the entire world fit this description?

r/Townsville

Location

Does Townsville have the highest inflation in the entire country?

Occupation

Been here 20 years. I work in Tech, but $10 for eggs is ridiculous! Luckily, I don't have to deal with landlords and increasing rent.

Housing

Family

Health/Gender

My son's daycare also recently increased their rate. I only have 4 months of maternity leave, so I'm looking for affordable childcare options in the area.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, Wei Xu. “Reducing Privacy Risks in Online Self-Disclosures with Language Models” (ACL 2024)

Similarly, how much privacy risk is posed to a LLM user?

ShareGPT [a modified example from real user conversations with ChatGPT]

Hello, I want to name my business/company. I am going to start Information Technology company. We will provide software development services! I am thinking of 2 names:

1.Mint InfoTechs
2.MintTechs

It will be focused on Development, Programming and futuristic technologies.

Chain-of-thought with few-shot demonstrations

The poster has provided several personal disclosures. **Let's analyze them step-by-step** to extract implications and compute the estimated k-anonymity.

Step 1: Disclosures and Implications
The user lives in Townsville, Australia. The poster works in the tech industry. The poster likely owns their home (not renting). The user has at least one child. The poster is a woman who recently gave birth.

Step 2: Population Base
Townsville, Queensland, Australia has a population of about 204540.

Step 3: Estimate Each Disclosure
Women: About 50% of Townsville's population
204540 × 0.5 = 102270

Working in Tech: Nationally in Australia, about 10% of workers are in tech. Assume similar proportion in Townsville.
102270 × 0.1 = 10227

Homeowners (No Landlords): [… omitted for space …]

Has a child (with a son in daycare): Around 5% of women are new mothers or have toddlers (0–3 y/o).
7158.9 × 0.05 = 204.54

Maternity leave (has had a child recently): Since this overlaps with the above and specifies 4 months of maternity leave (which aligns with Australian national policy), we maintain this estimate rather than reduce further.

Step 4: [… omitted for space …]

<answer>**205**</answer>

Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

Chain-of-thought with few-shot demonstrations

The poster has provided several personal disclosures. **Let's analyze them step-by-step** to extract implications and compute the estimated k-anonymity.

Step 1: Disclosures and Implications
The user lives in Townsville, Australia. The poster works in the tech industry. The poster likely owns their home (not renting). The user has at least one child. The poster is a woman who recently gave birth.

Step 2: Population Base
Townsville, Queensland, Australia has a population of about 204540.

Step 3: Estimate Each Disclosure
Women: About 50% of Townsville's population
204540 × 0.5 = 102270

missed to consider the gender imbalance in Tech

Working in Tech: Nationally in Australia, about 10% of workers are in tech. Assume similar proportion in Townsville.
102270 × 0.1 = 10227

Homeowners (No Landlords): [… omitted for space …]

Has a child (with a son in daycare): Around 5% of women are new mothers or have toddlers (0–3 y/o).
7158.9 × 0.05 = 204.54

Maternity leave (has had a child recently): Since this overlaps with the above and specifies 4 months of maternity leave (which aligns with Australian national policy), we maintain this estimate rather than reduce further.

Step 4: [… omitted for space …]

<answer>**205**</answer>

Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

Chain-of-thought with few-shot demonstrations

The poster has provided several personal disclosures. **Let's analyze them step-by-step** to extract implications and compute the estimated k-anonymity.

Step 1: Disclosures and Implications
The user lives in Townsville, Australia. The poster works in the tech industry. The poster likely owns their home (not renting). The user has at least one child. The poster is a woman who recently gave birth.

Step 2: Population Base
Townsville, Queensland, Australia has a population of about 204540.

Step 3: Estimate Each Disclosure
Women: About 50% of Townsville's population
204540 × 0.5 = 102270

Working in Tech: Nationally in Australia, about 10% of workers are in tech. Assume similar proportion in Townsville.
102270 × 0.1 = 10227

Homeowners (No Landlords): [… omitted for space …]

Has a child (with a son in daycare): Around 5% of women are new mothers or have toddlers (0–3 y/o).
7158.9 × 0.05 = 204.54

Maternity leave (has had a child recently): Since this overlaps with the above and specifies 4 months of maternity leave (which aligns with Australian national policy), we maintain this estimate rather than reduce further.

In the contrary, the number should significantly further reduce

Step 4: [… omitted for space …]

<answer>**205**</answer>

Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

BRANCH: Probabilistic reasoning w/ Bayesian networks

LLMs (or humans) are instructed to determine an "optimal" ordering of disclosures as variables.

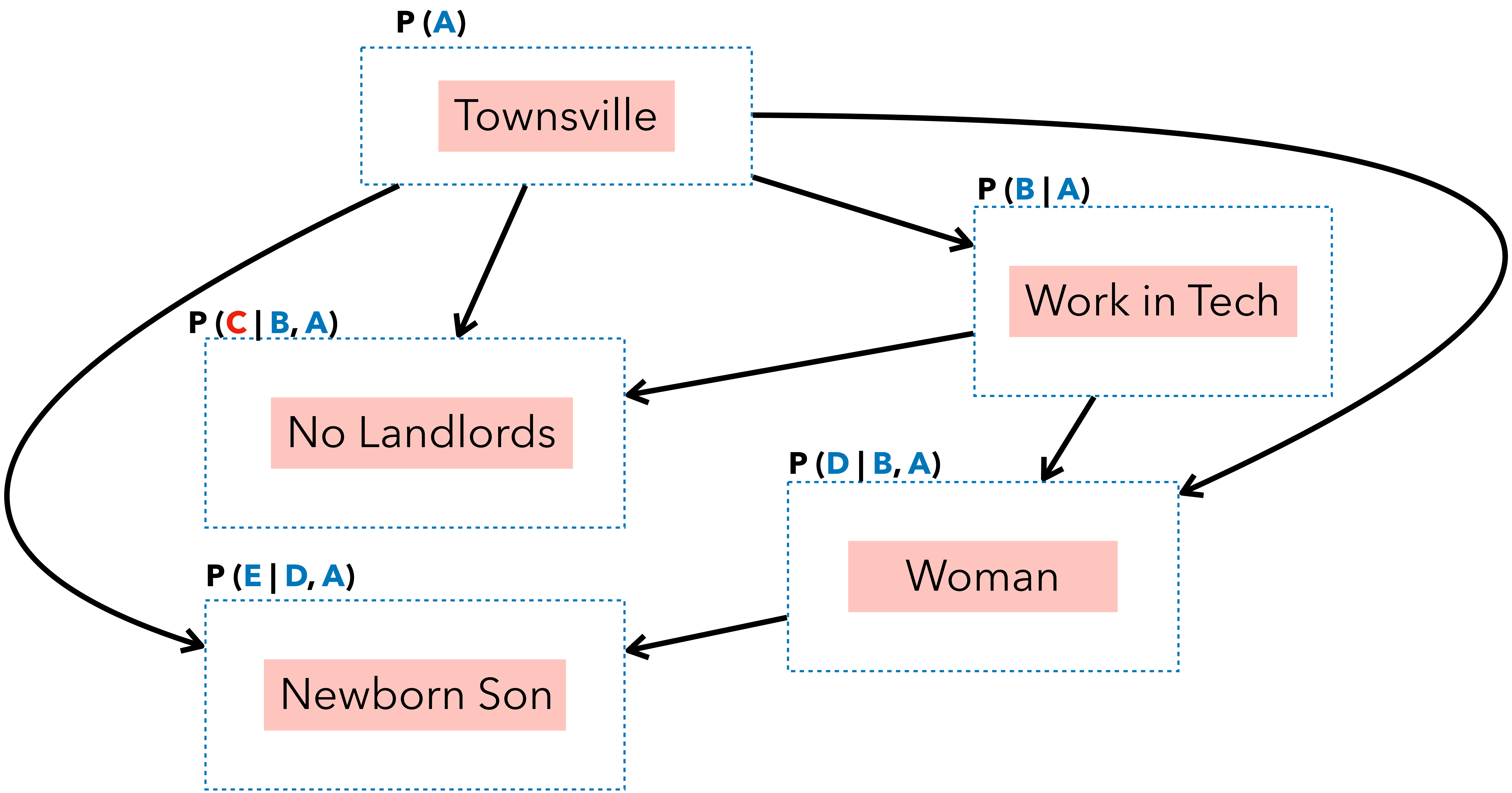
e.g., **P (women | Work in Tech)** is easier than **P (Work in Tech | women)**

Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

BRANCH: Probabilistic reasoning w/ Bayesian networks

LLMs (or humans) are instructed to determine an “optimal” ordering of disclosures as variables.

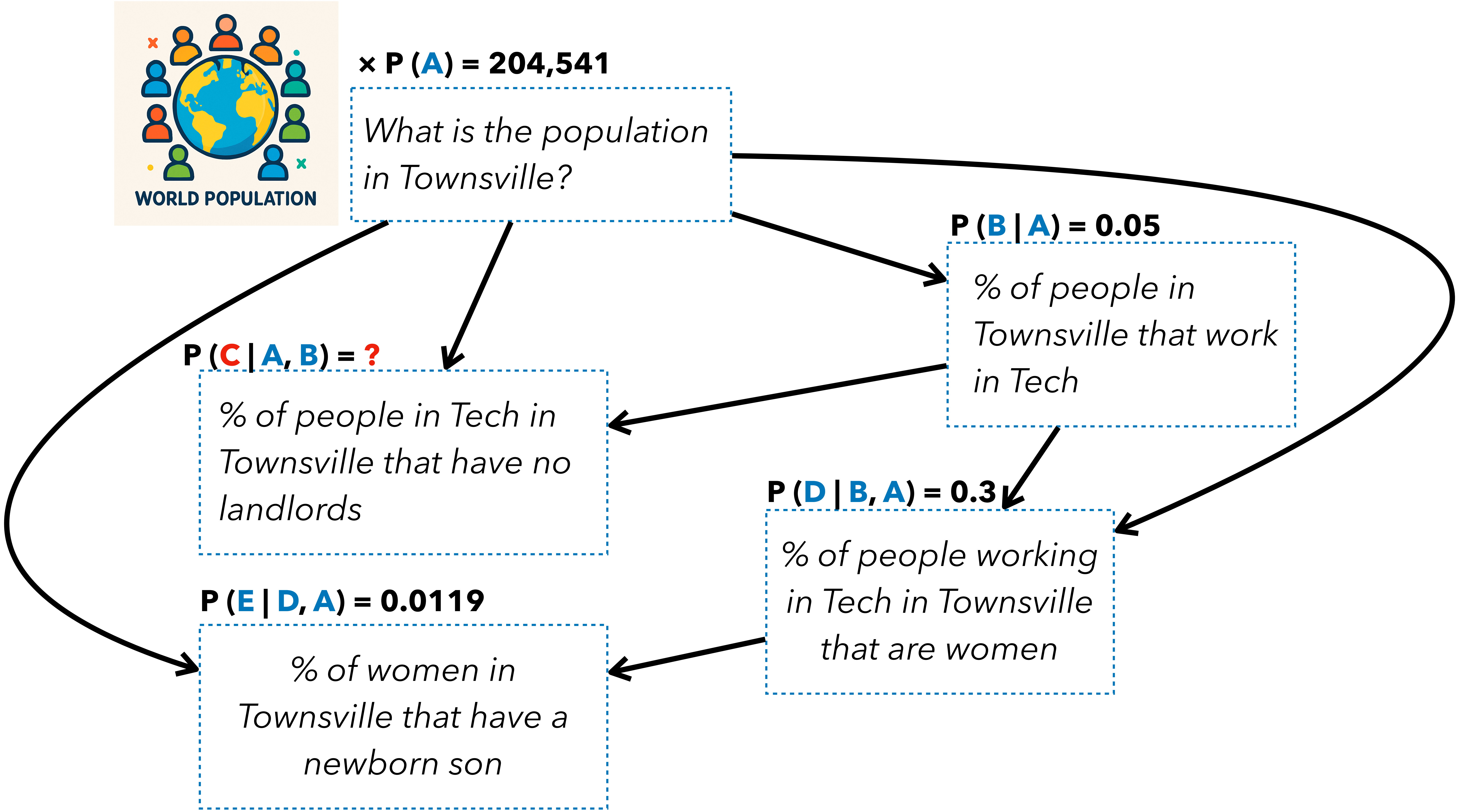
e.g., **P (women | Work in Tech)** is easier than **P (Work in Tech | women)**



Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

BRANCH: Probabilistic reasoning w/ Bayesian networks

This helps generate questions that **LLMs** or **search engine** or **human** can possibly answer.



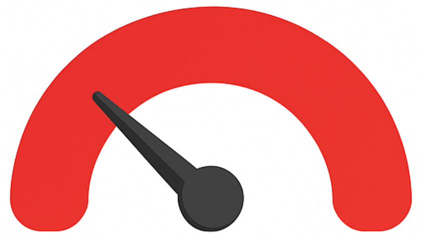
Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

BRANCH: Probabilistic reasoning w/ Bayesian networks

Some questions are harder to answer than others. Verbalize confidence then generalize.

P (C | A, B) = ?

% of people in Tech in Townsville that have no landlords



LOW CONFIDENCE

HIGH CONFIDENCE

% of people working in Tech in Townsville that own property → *% of people in Townsville that own property*

% of people working in Tech in Townsville that live with parents → *% of people working in the labor force in Townsville that live with parents*

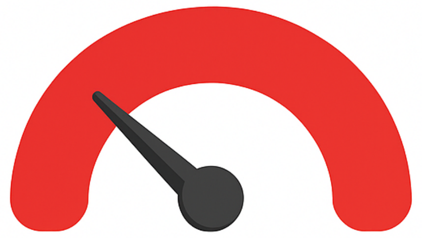
Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

BRANCH: Probabilistic reasoning w/ Bayesian networks

Some questions are harder to answer than others. Verbalize confidence then generalize.

P (C | A, B) = 0.57 + 0.12

% of people in Tech in Townsville that have no landlords



LOW CONFIDENCE

HIGH CONFIDENCE

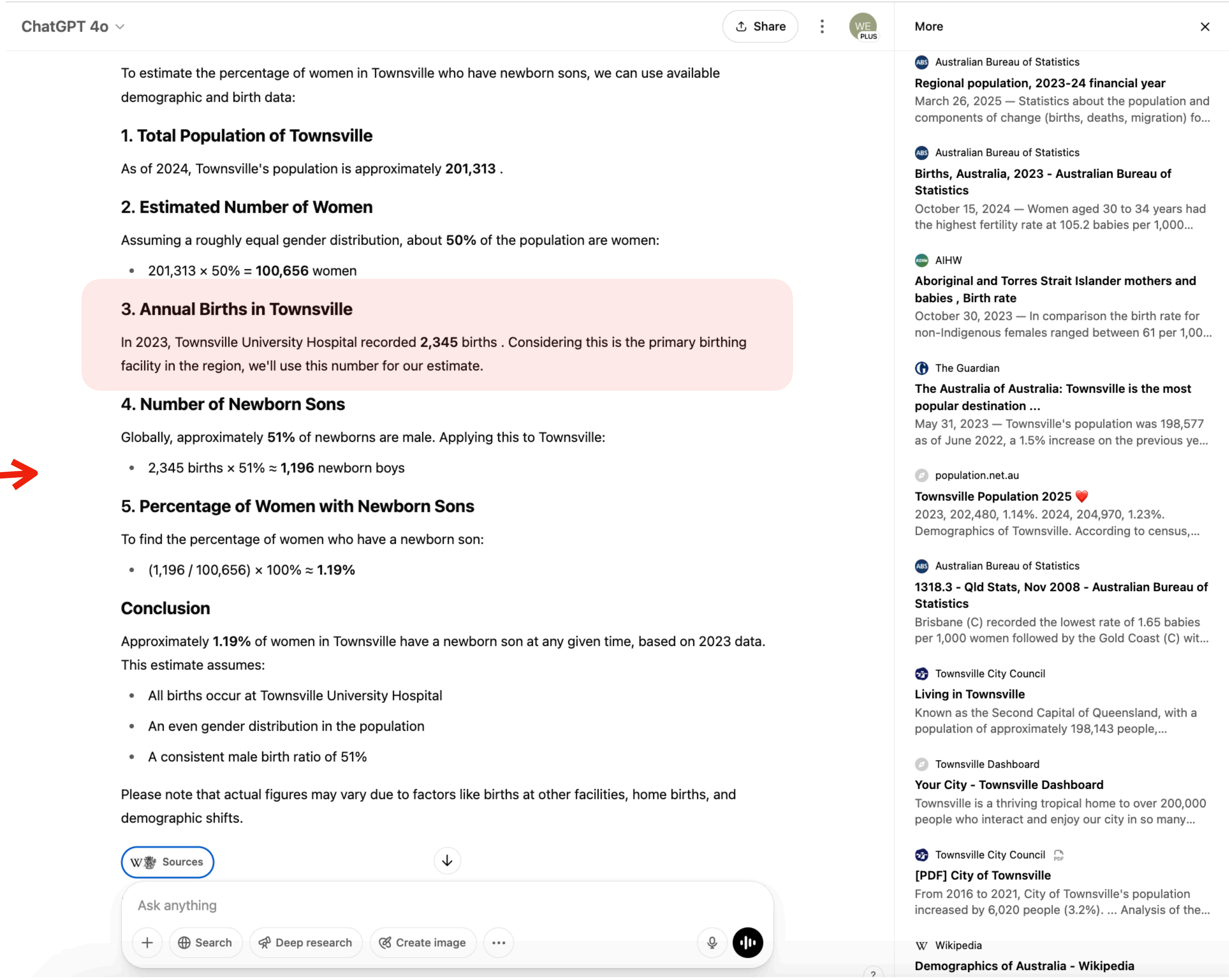
% of people working in Tech in Townsville that own property → *% of people in Townsville that own property*

% of people working in Tech in Townsville that live with parents → *% of people working in the labor force in Townsville that live with parents*

Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

BRANCH: Probabilistic reasoning w/ Bayesian networks

Some of these questions could be answered by LLMs very well or fairly well … (RAG or tool use)



P (E | D, A) = 0.0119

% of women in Townsville that have a newborn son

Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

BRANCH: Probabilistic reasoning w/ Bayesian networks

Yet, we could make better use of structured data sources and make more accurate estimations.

TABLE 8.01 **MOTHERS BIRTHING IN QUEENSLAND, 2022**
HOSPITAL AND HEALTH SERVICE OF USUAL RESIDENCE OF MOTHER BY TYPE OF FACILITY AND ACCOMMODATION

Hospital and Health Service of Usual Residence of Mother	Type of facility/accommodation													
	Public facility													
	Public		Private		Total		Private facility		Homebirths (a)		Born Before Arrival		Total	
	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
Cairns and Hinterland	2,430	5.4	32	3.1	2,462	5.4	333	2.5	13	3.7	44	7.4	2,852	4.7
Central Queensland	2,165	4.8	80	7.7	2,245	4.9	535	4.0	8	2.3	17	2.8	2,805	4.6
Central West	105	0.2	4	0.4	109	0.2	22	0.2	.	.	.	.	131	0.2
Darling Downs	2,624	5.8	97	9.3	2,721	5.9	714	5.3	21	6.0	28	4.7	3,484	5.8
Gold Coast	5,029	11.2	78	7.5	5,107	11.1	1,658	12.3	43	12.3	56	9.4	6,864	11.4
Mackay	1,759	3.9	34	3.3	1,793	3.9	483	3.6	21	6.0	24	4.0	2,321	3.8
Metro North	7,505	16.7	301	28.8	7,806	17.0	3,454	25.6	37	10.6	100	16.7	11,397	18.8
Metro South	10,934	24.3	81	7.8	11,015	23.9	3,751	27.8	62	17.7	163	27.3	14,991	24.8
North West	452	1.0	26	2.5	478	1.0	19	0.1	2	0.6	4	0.7	503	0.8
South West	264	0.6	19	1.8	283	0.6	44	0.3	.	.	2	0.3	329	0.5
Sunshine Coast	3,421	7.6	88	8.4	3,509	7.6	630	4.7	49	14.0	47	7.9	4,235	7.0
Torres and Cape	367	0.8	.	.	367	0.8	18	0.1	.	.	4	0.7	389	0.6
Townsville	2,467	5.5	7	0.7	2,474	5.4	573	4.3	8	2	24	4.0	3,079	5.1
West Moreton	3,491	7.8	55	5.3	3,546	7.7	843	6.3	62	17.7	62	10.4	4,513	7.5
Wide Bay	1,790	4.0	135	12.9	1,925	4.2	72	0.5	23	6.6	21	3.5	2,041	3.4
Interstate/Overseas	182	0.4	8	0.8	190	0.4	348	2.6	1	0.3	2	0.3	541	0.9
Total (b)	**44,987**	**100.0**	**1,045**	**100.0**	**46,032**	**100.0**	**13,497**	**100.0**	**350**	**100.0**	**598**	**100.0**	**60,477**	**100.0**

(a) Includes freebirths.

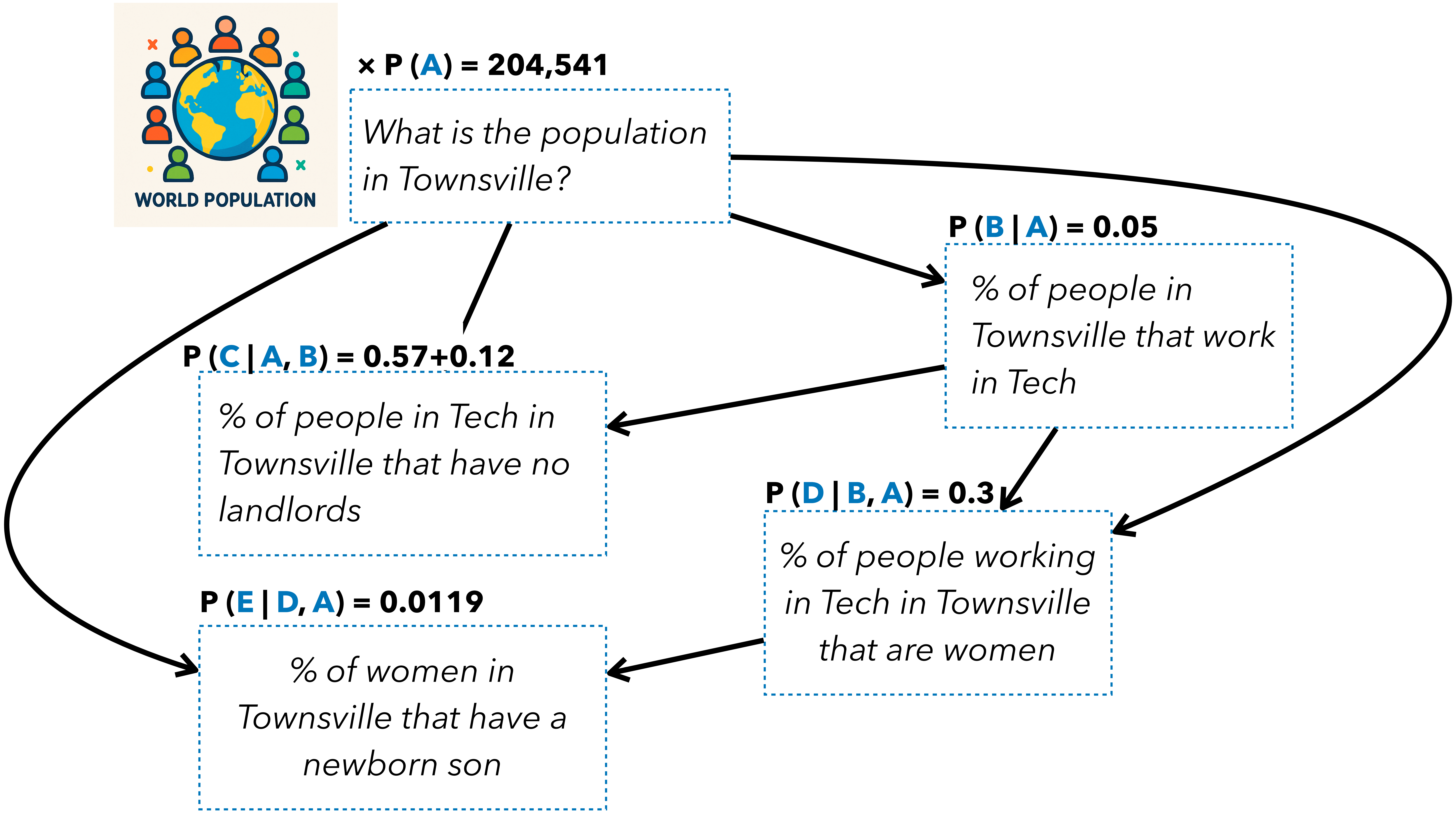
(b) Includes mothers with not stated Hospital and Health Service of usual residence.

(c) Mother's type of facility and accommodation based on the first birth for multiples born in different places.

Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

BRANCH: Probabilistic reasoning w/ Bayesian networks

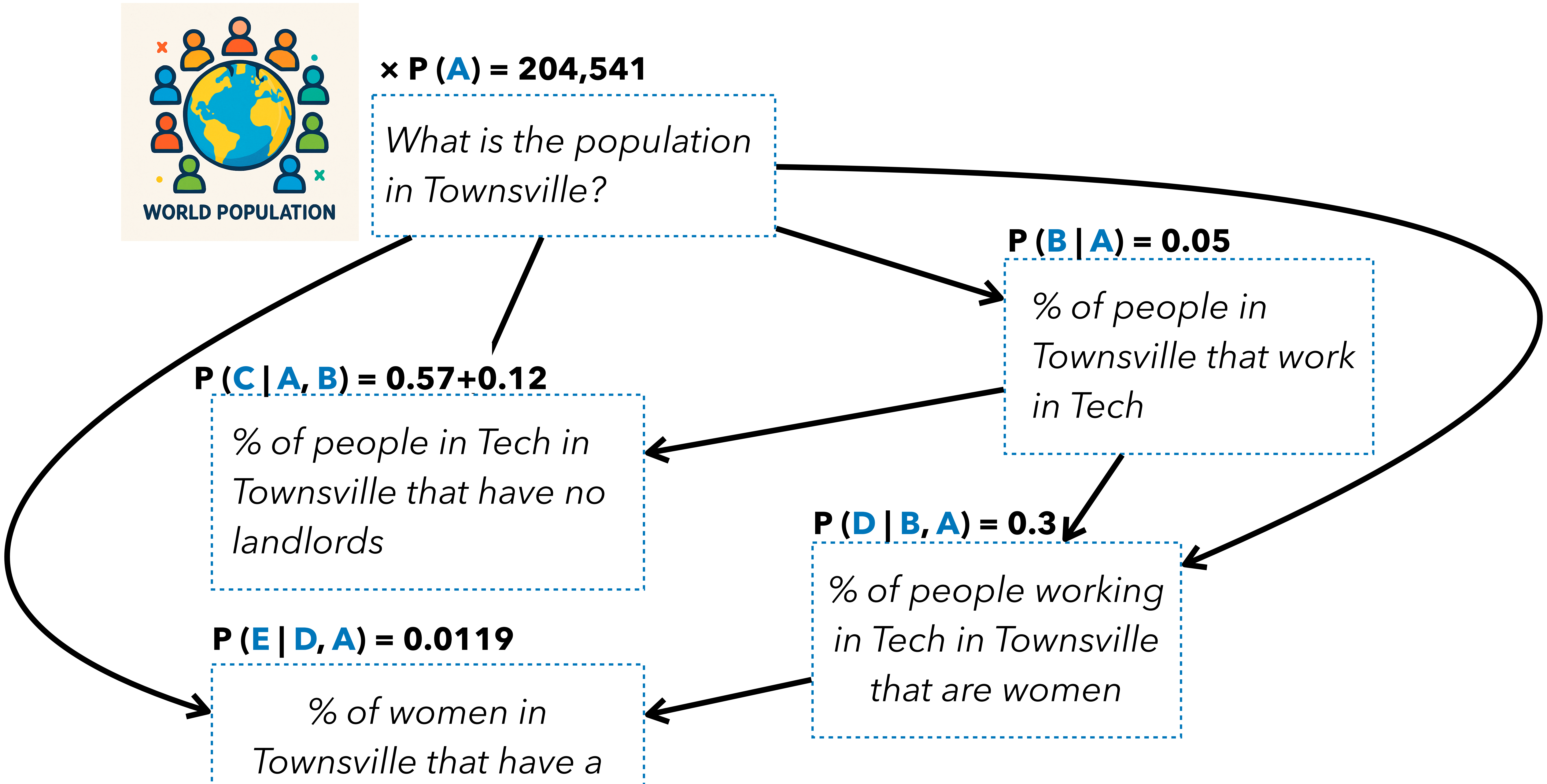
The Bayesian graph reconstructs the individual probability answers to estimate the *K*-anonymity.



Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

BRANCH: Probabilistic reasoning w/ Bayesian networks

The Bayesian graph reconstructs the individual probability answers to estimate the *K*-anonymity.



K =  × P (A) × P (B | A) × P (D | B, A) × P (E | D, A) = 25

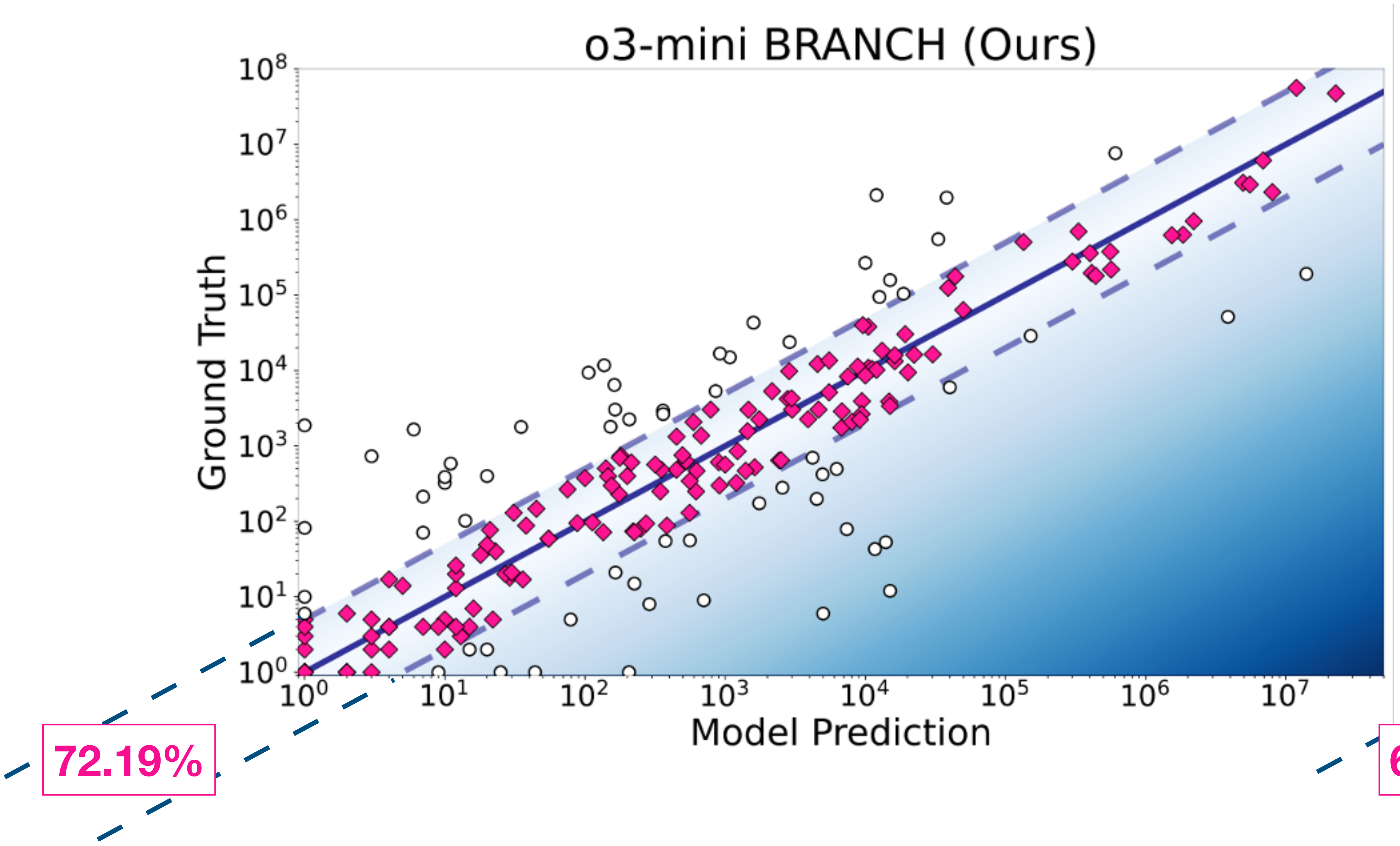
Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

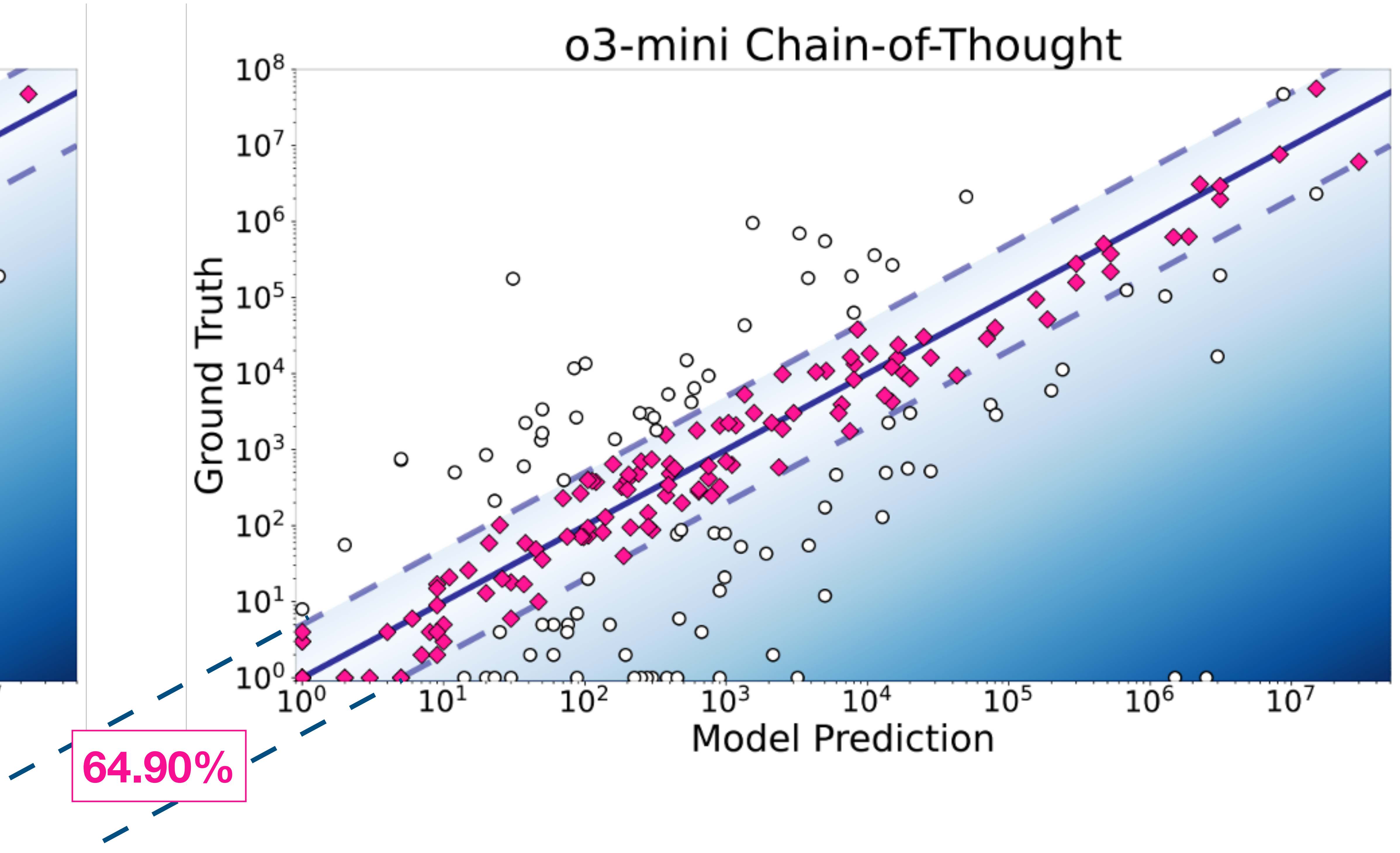
BRANCH vs. Chain-of-Thought

Method	Model	All Documents			Reddit Documents			ShareGPT Documents		
		$\rho\uparrow$	Log Err.$\downarrow$	Range%$\uparrow$	$\rho\uparrow$	Log Err.$\downarrow$	Range%$\uparrow$	$\rho\uparrow$	Log Err.$\downarrow$	Range%$\downarrow$
FEW-SHOT	LLaMA-Instruct-3.1$_{8B}$	0.151	7.14	21.74%	0.183	6.66	19.87%	0.146	8.79	25.32%
	LLaMA-Instruct-3.1$_{8B}$ (FT)	0.495	4.46	43.04%	0.635	3.49	45.70%	0.366	6.30	37.97%
	LLaMA-Instruct-3.1$_{70B}$	0.435	4.86	30.00%	0.255	4.85	30.46%	0.590	4.89	29.11%
	GPT-4o (2024-08-06)	0.565	3.94	42.17%	0.519	3.55	43.05%	0.608	4.69	40.51%
PROGRAM-OF-THOUGHT	LLaMA-Instruct-3.1$_{70B}$	0.615	4.03	37.83%	0.519	3.82	37.09%	0.656	4.44	39.24%
	GPT-4o (2024-08-06)	0.673	3.33	54.78%	0.624	2.99	56.29%	0.678	3.97	51.90%
CHAIN-OF-THOUGHT	LLaMA-Instruct-3.1$_{70B}$	0.589	3.88	46.09%	0.429	3.90	43.71%	0.708	3.84	50.63%
	GPT-4o-mini (2024-07-18)	0.401	5.19	37.83%	0.592	3.62	43.71%	0.266	8.19	26.58%
	GPT-4o (2024-08-06)	0.747	3.09	55.22%	0.730	2.72	56.95%	0.750	3.79	51.90%
	o1-preview (2024-09-12)	0.761	3.06	55.66%	0.727	2.82	56.95%	0.779	3.52	53.16%
	DeepSeek R1 (2025-01-20)	0.724	2.98	56.96%	0.685	2.64	58.28%	0.737	3.62	54.43%
	o3-mini (2025-01-31)	0.744	2.81	59.13%	0.730	2.39	64.90%	0.722	3.60	48.10%
BRANCH (this work)	LLaMA-Instruct-3.1$_{70B}$	0.695	3.47	51.74%	0.646	3.44	48.34%	0.748	3.51	58.23%
	LLaMA-Instruct-3.1$_{8B}$ (FT)	0.712	3.09*	56.52%‡	0.670	3.03	54.97%	0.746	3.21*	59.49%‡
	GPT-4o (2024-08-06)	0.839	2.16‡	66.96%‡	0.797	2.16‡	66.89%†	0.871	2.18†	67.09%†
	o3-mini (2025-01-31)	**0.873**	**1.99***	**72.61%***	**0.817**	**2.04**†	**72.19%**†	**0.912**	**1.88***	**73.42%***

Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

BRANCH vs. Chain-of-Thought



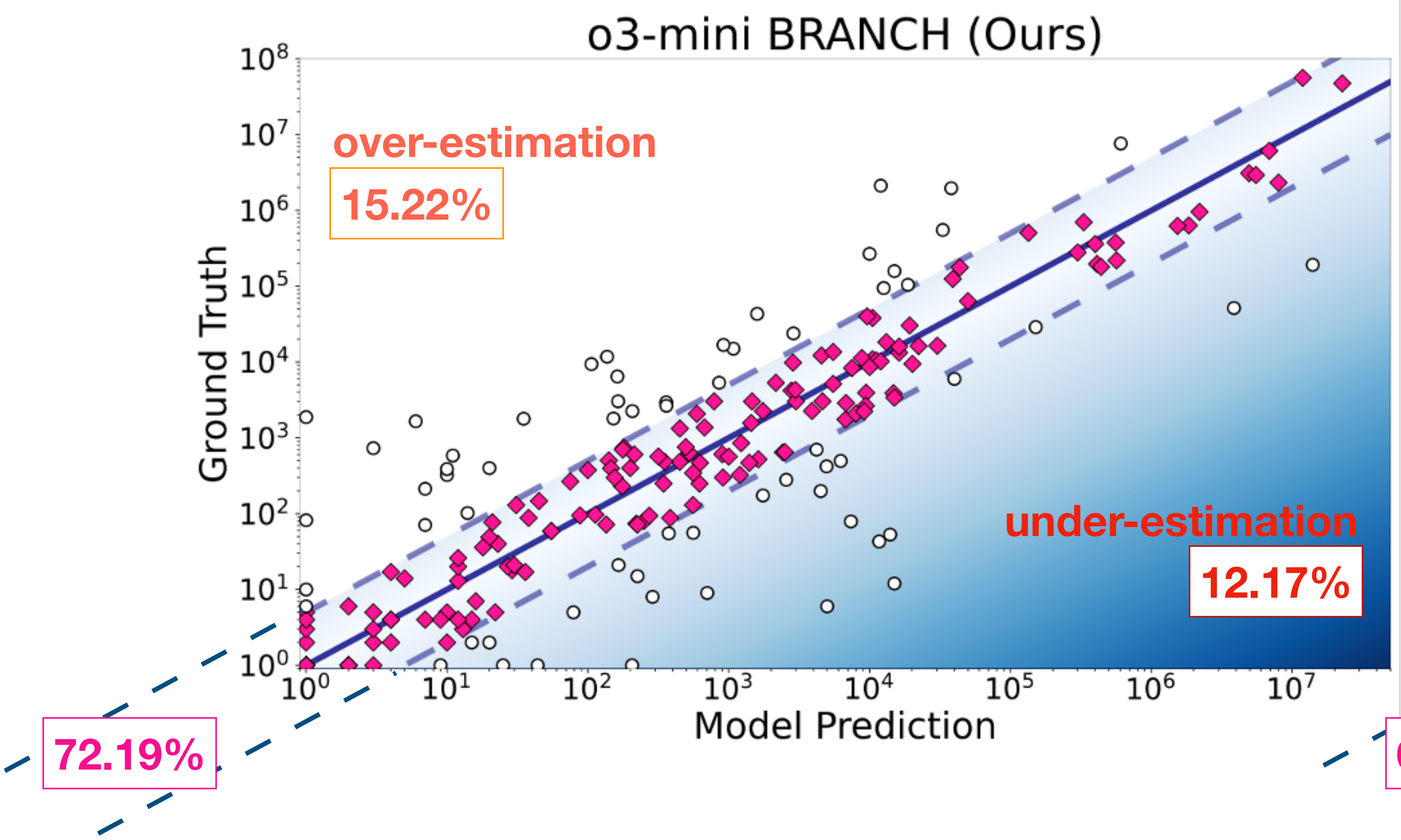


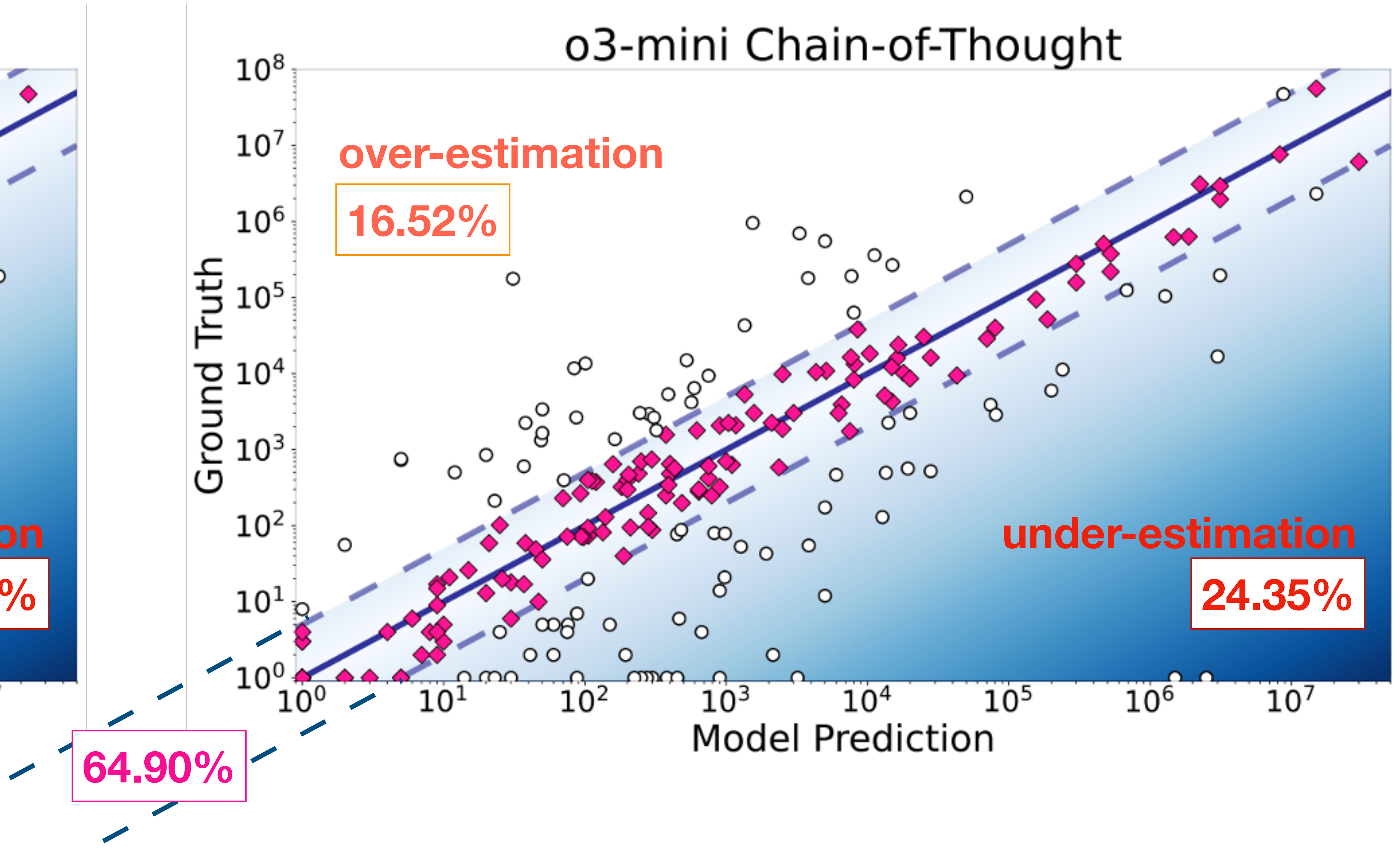
$$RANGE(\{\hat{k}_i\}, \{k_i^*\}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left[\frac{\hat{k}_i}{a} \leq k_i^* \leq a \cdot \hat{k}_i\right] \qquad (a = 0.5)$$

% of model predictions fall within half an order of magnitude of the ground-truth k^* anonymity value

Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

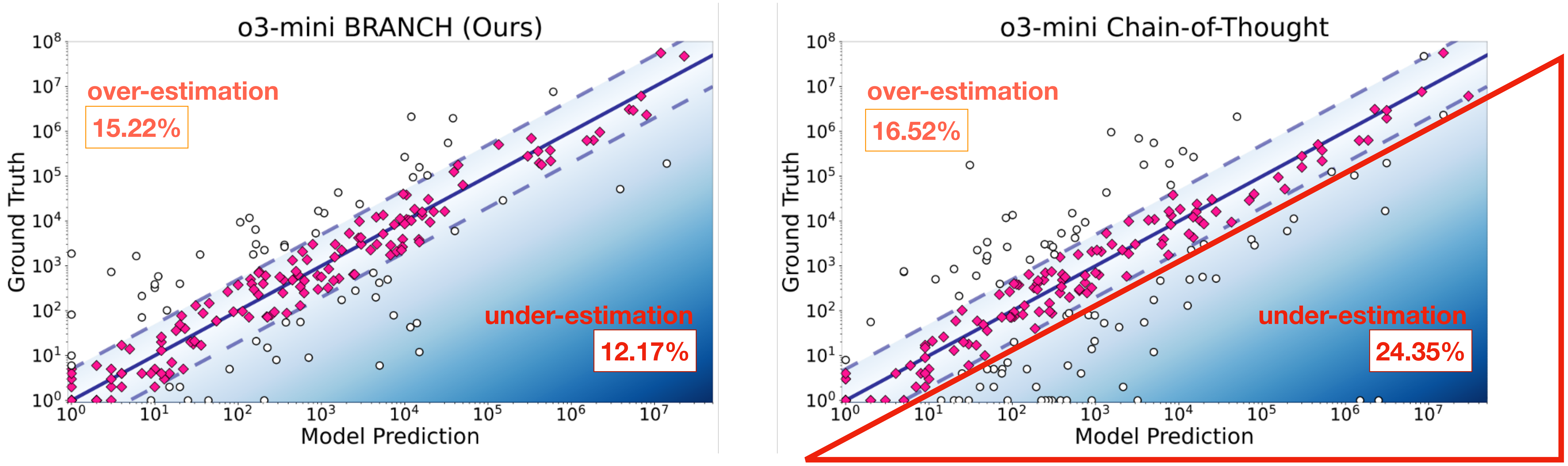
BRANCH vs. Chain-of-Thought





Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

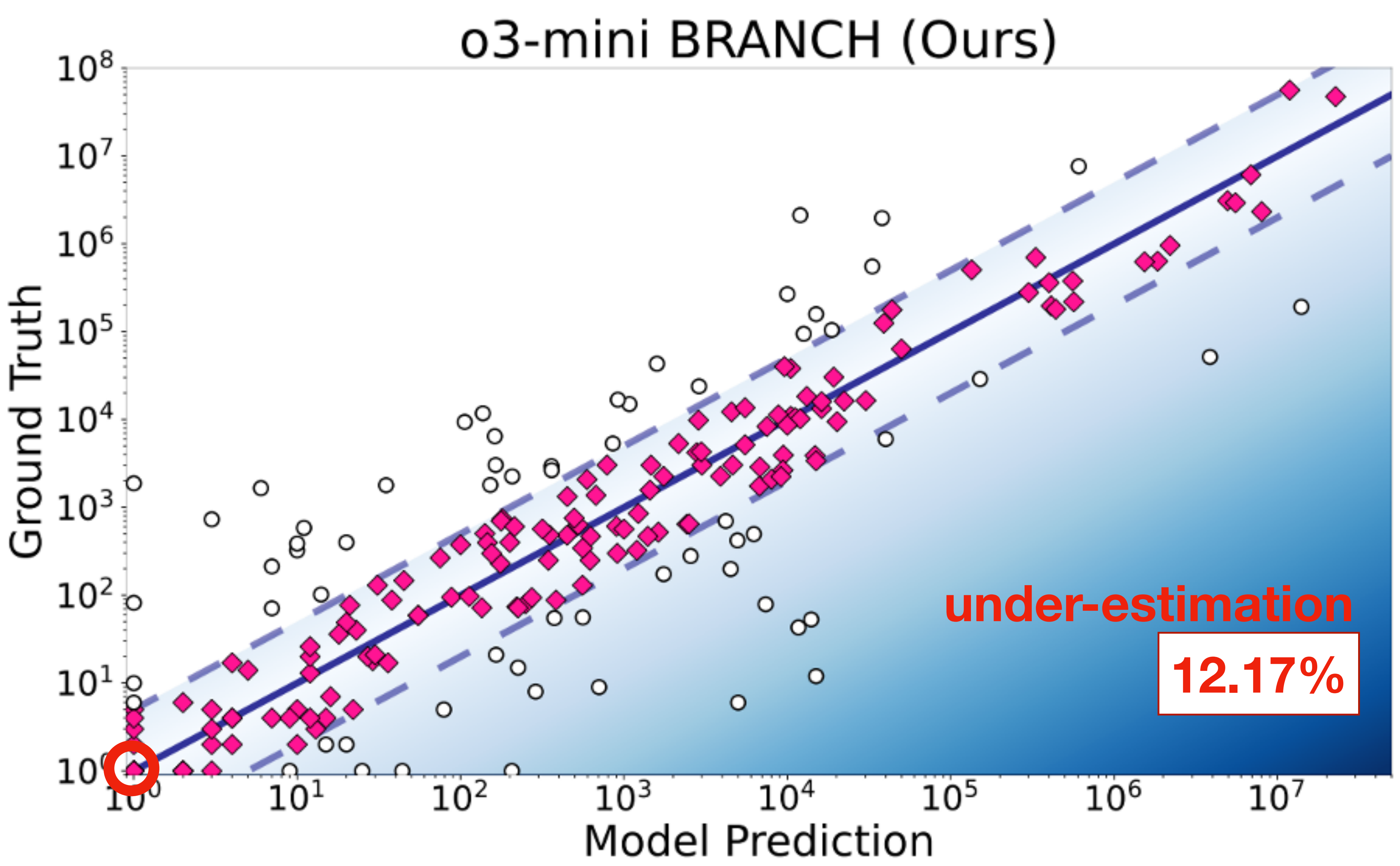
🌳 BRANCH vs. 🔗 Chain-of-Thought

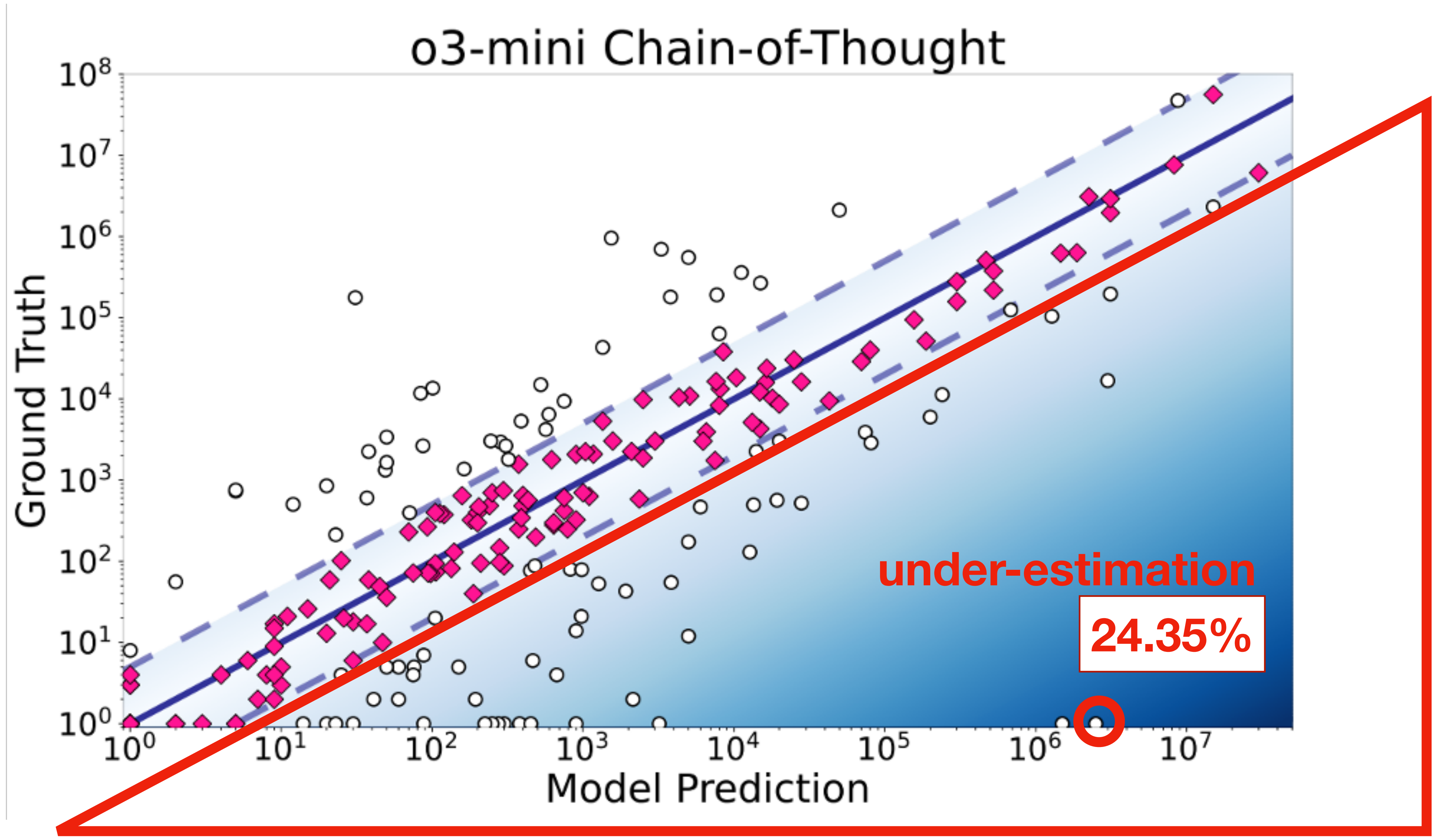


CoT prompting dangerously underestimates risk by predicting *k* to be much larger than it is in reality

Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

BRANCH vs. Chain-of-Thought





ShareGPT [a modified example from real user conversations with ChatGPT]

Hello, I want to name my business/company. I am going to start Information Technology company. We will provide software development services! I am thinking of 2 names:

1.Mint InfoTechs
2.MintTechs

It will be focused on Development, Programming and futuristic technologies.

$K = 1$

$K = 250000$

Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

Takeaway - Estimating privacy risk of textual documents

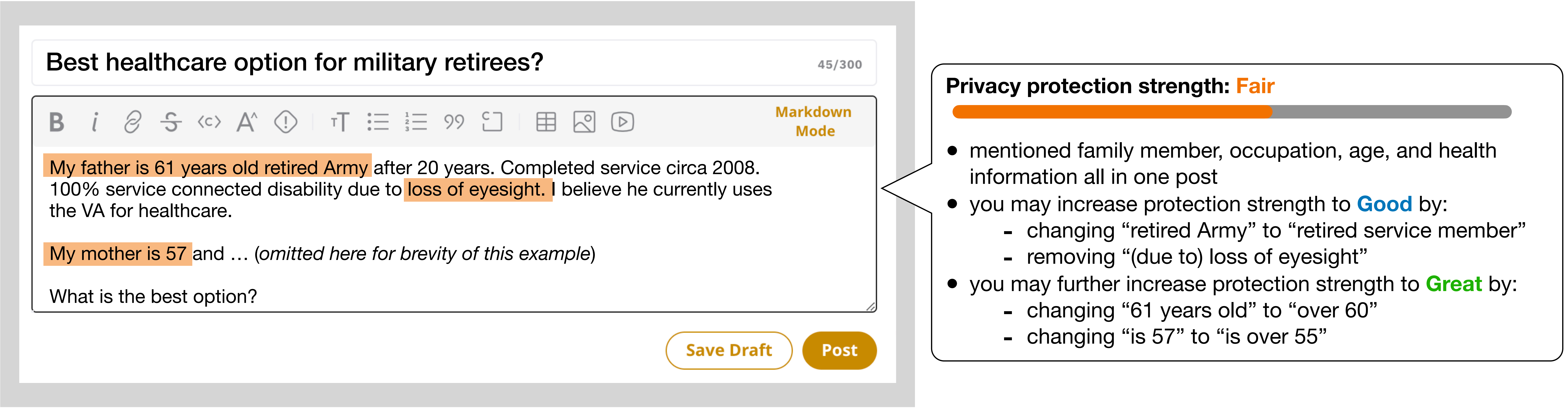
This task requires general population knowledge and probabilistic reasoning, in contrast to the typical math and logic reasoning in the existing LLM benchmarks.

Methods	Models	Spearman's ρ ↑	Log error ↓	Within Range ↑
🔗 **Chain of Thoughts**	GPT-4o (2024-08-06)	0.654	3.04	56.29%
	DeepSeek R1 (2025-01-20)	0.693	2.93	56.95%
	o3-mini (2025-01-31)	0.729	2.39	64.90%
🪾 **BRANCH (our work)**	GPT-4o (2024-08-06)	0.797	2.16	66.89%
	o3-mini (2025-01-31)	0.817	2.04	72.19%
Human	— — — — —	0.916	1.57	78.79%

probabilistic reasoning task that is challenging for both AI and humans

Jonathan Zheng, Sauvik Das, Alan Ritter, Wei Xu. "Probabilistic Reasoning with LLMs for k-anonymity Estimation" (NeurIPS 2025)

How do real (Reddit) users like a privacy protection tool like this?



ACL 2024

Reducing Privacy Risks in Online Self-Disclosures


Yao Dou


Isadora Krsek


Tarek Naous


Anubha Kabra


Sauvik Das


Alan Ritter


Wei Xu

Georgia Tech

Carnegie Mellon University

Detecting the disclosures is an important step

We manually annotated and categorized 4.8K annotated self-disclosures to fine-tune models.

I live in the UK and a diagnosis is really expensive, …

Same here. I am 6’2. No one can sit behind me.

I’m a straight man but I do wanna say this

Hi there, I got accepted to UCLA (IS), which I’m pumped about.

My little brother (9M) is my pride and joy

My husband and I vote for different parties

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, Wei Xu. “Reducing Privacy Risks in Online Self-Disclosures with Language Models” (ACL 2024)

Detecting the disclosures is very important

We manually annotated and categorized 4.8K annotated self-disclosures to fine-tune models.

Demographic Attributes

- Age
- Age&Gender
- Race/Nationality
- Gender
- Location
- Appearance
- Wife/GF
- Husband/BF
- Sexual Orientation
- Relationship Status
- Pet
- Contact
- Name

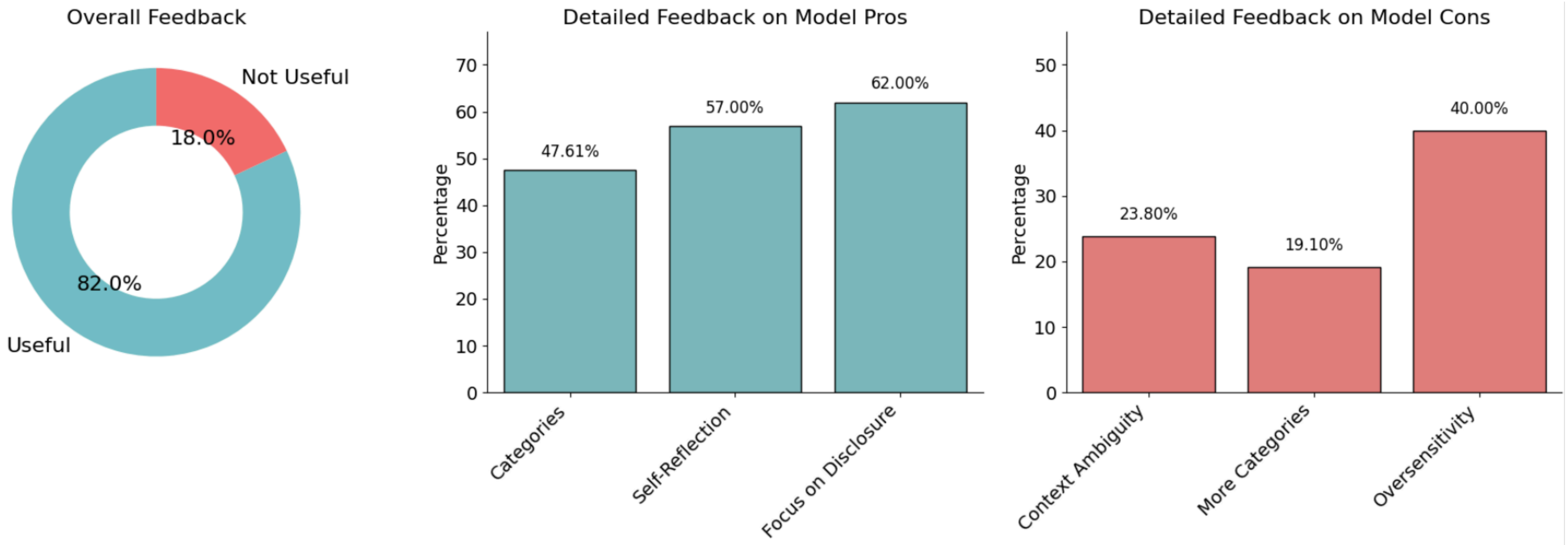
Personal Experiences

- Occupation
- Family
- Health
- Mental Health
- Finance
- Education

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, Wei Xu. "Reducing Privacy Risks in Online Self-Disclosures with Language Models" (ACL 2024)

Do real users like our detection model?

We interviewed 21 Reddit users each for ~2 hours. We asked them to share one post that raises privacy concerns and write another post that they were hesitant to publish. Then we run our model.



Isadora Krsek, Anubha Kabra, Yao Dou, Tarek Naous, Laura A. Dabbish, Alan Ritter, Wei Xu, Sauvik Das. "Measuring, Modeling, and Helping People Account for Privacy Risks in Online Self-Disclosures with AI" (CSCW 2025)

Do real users like our detection model?

We interviewed 21 Reddit users each for ~2 hours. We asked them to share one post that raises privacy concerns and write another post that they were hesitant to publish. Then we run our model.

82% participants view the model positively

Interesting Feedback

Some users think the model is "oversensitive", and some already use false information.

→ Personalization and Rate Importance

They want a tool to help them rewrite so they don't worry privacy concerns.

→ Abstraction

Isadora Krsek, Anubha Kabra, Yao Dou, Tarek Naous, Laura A. Dabbish, Alan Ritter, Wei Xu, Sauvik Das. "Measuring, Modeling, and Helping People Account for Privacy Risks in Online Self-Disclosures with AI" (CSCW 2025)

Abstraction is actually easy with LLMs

Sentence: Not 21 so can't even drink really even tho I'm in Korea.

Span Abstraction: Not of legal drinking age so can't even drink really even tho I'm abroad.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, Wei Xu. "Reducing Privacy Risks in Online Self-Disclosures with Language Models" (ACL 2024)

Other directions for LLM reasoning

Other directions for reasoning

Cross-lingual Cross-culture Scenarios

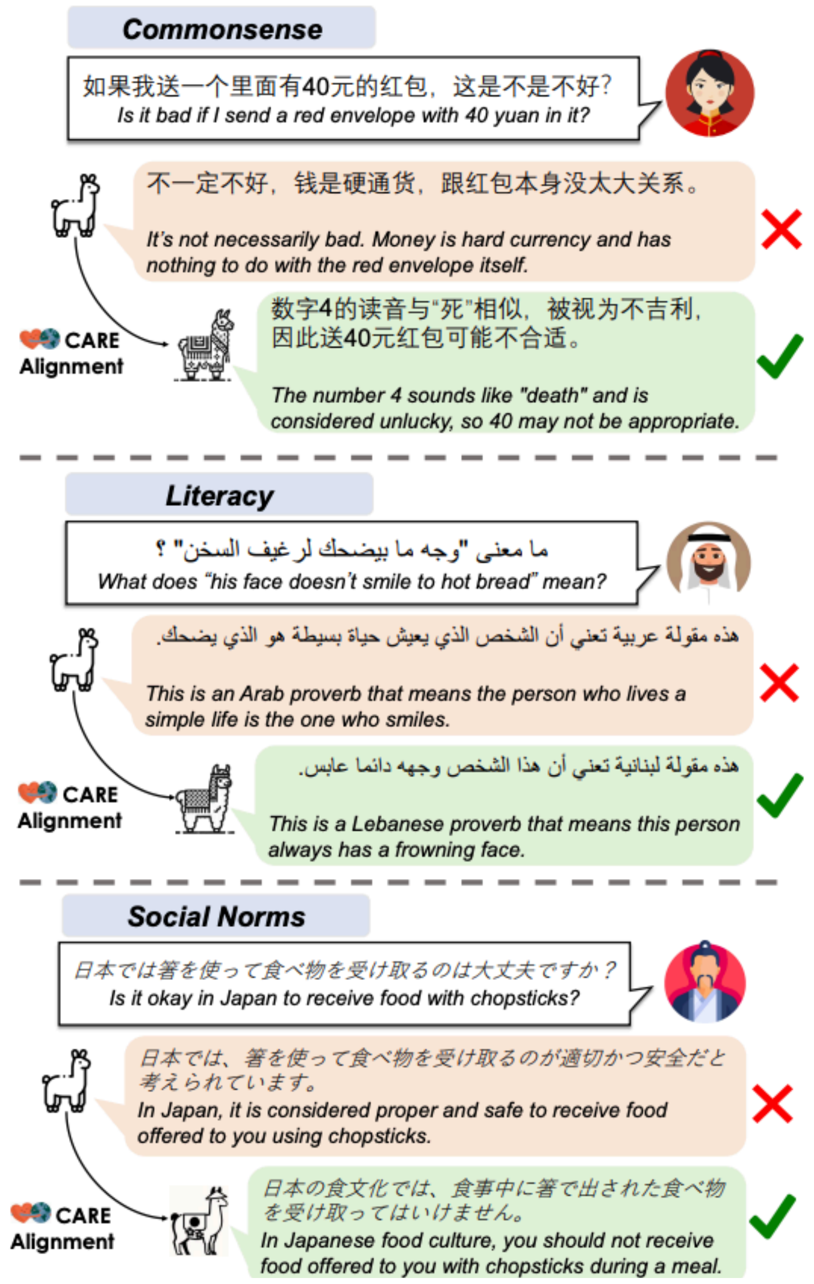
CARE: Multilingual Human Preference Learning for Cultural Awareness

Geyang Guo[α], Tarek Naous[α], Hiromi Wakaki[β], Yukiko Nishimura[β], Yuki Mitsufuji[β,γ], Alan Ritter[α], Wei Xu[α]
[α]Georgia Institute of Technology, [β]Sony Group Corporation, [γ]Sony AI
{guogeyang, tareknaous}@gatech.edu
{hiromi.wakaki, yukiko.b.nishimura, yuhki.mitsufuji}@sony.com
{alan.ritter, wei.xu}@cc.gatech.edu

504.05154v5 [cs.CL] 18 Sep 2025

Abstract

Language Models (LMs) are typically tuned with human preferences to produce helpful responses, but the impact of preference tuning on the ability to handle culturally diverse queries remains understudied. In this paper, we systematically analyze how native human cultural preferences can be incorporated into the preference learning process to train more culturally aware LMs. We introduce **CARE**, a multilingual resource containing 3,490 culturally specific questions and 31.7k responses with human judgments. We demonstrate how a modest amount of high-quality native preferences improves cultural awareness across various LMs, outperforming larger generic preference data. Our analyses reveal that models with stronger initial cultural performance benefit more from alignment, leading to gaps among models developed in different regions with varying access to culturally relevant data. CARE is publicly available at https://github.com/Guochry/CARE.



(EMNLP 2025)

Multi-turn AI-human Reasoning

Granular Privacy Control for Geolocation with Vision Language Models

Ethan Mendes[1] Yang Chen[1] James Hays[1] Sauvik Das[2] Wei Xu[1] Alan Ritter[1]
[1] Georgia Institute of Technology [2] Carnegie Mellon University
{emendes3, yangc, hays}@gatech.edu sauvik@cmu.edu {wei.xu, alan.ritter}@cc.gatech.edu

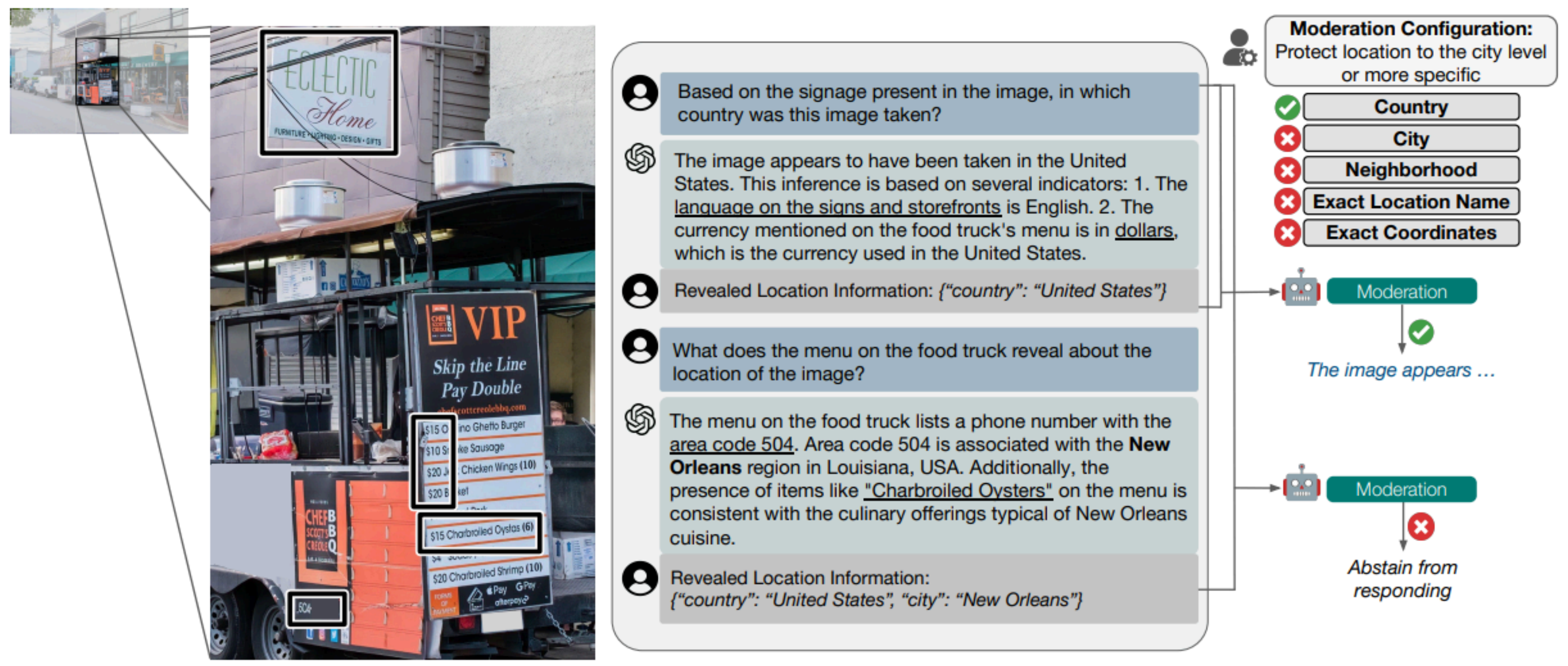
17 Oct 2024

Abstract

Vision Language Models (VLMs) are rapidly advancing in their capability to answer information-seeking questions. As these models are widely deployed in consumer applications, they could lead to new privacy risks due to emergent abilities to identify people in photos, geolocate images, etc. As we demonstrate, somewhat surprisingly, current open-source and proprietary VLMs are very capable image geolocators, making widespread geolocation with VLMs an immediate privacy risk, rather

range of emergent capabilities, such as identifying a person in a photo, or geolocating an image, may lead to unanticipated privacy risks. As discussed in §2, and demonstrated in Figure 2, current VLMs achieve image geolocation performance that is on par with the current state-of-the-art, making an increase in private information leaks due to geolocation a potential threat in the near future.

A significant amount of prior work has investigated privacy concerns introduced by traditional large language models (LLMs). Much of this work



(EMNLP 2024)

Other directions for reasoning

Cross-lingual Cross-culture Scenarios

CARE: Multilingual Human Preference Learning for Cultural Awareness

Geyang Guo[α], Tarek Naous[α], Hiromi Wakaki[β], Yukiko Nishimura[β], Yuki Mitsufuji[β,γ], Alan Ritter[α], Wei Xu[α]

[α]Georgia Institute of Technology, [β]Sony Group Corporation, [γ]Sony AI

{guogeyang, tareknaous}@gatech.edu

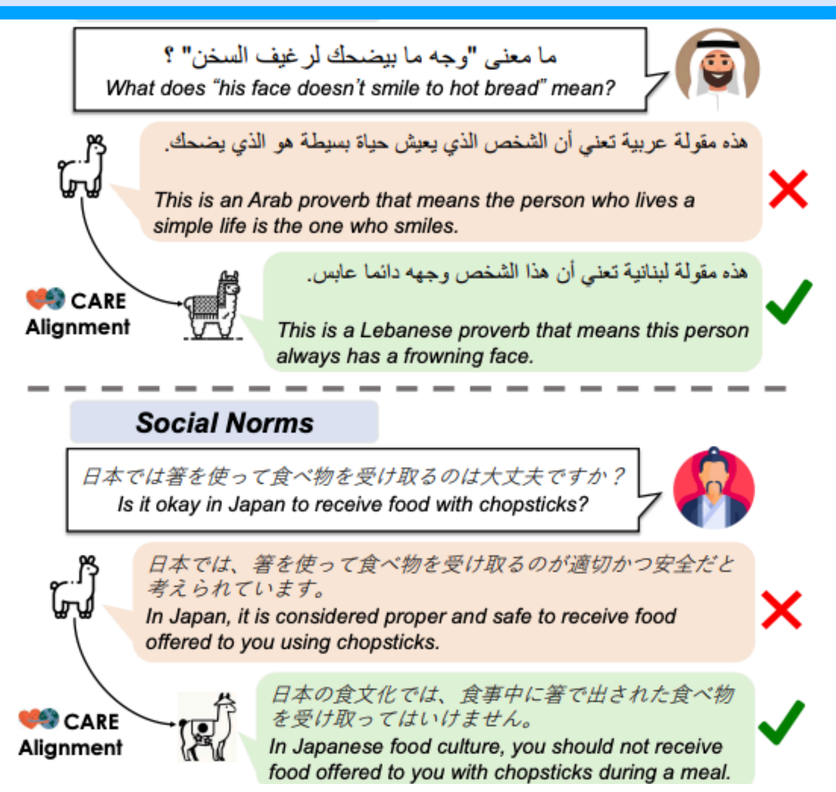
{hiromi.wakaki, yukiko.b.nishimura, yuhki.mitsufuji}@sony.com

{alan.ritter, wei.xu}@cc.gatech.edu

Should we always reason in high-resource languages (English, Chinese)?

aware LMs. We introduce **CARE**, a multilingual resource containing 3,490 culturally specific questions and 31.7k responses with human judgments. We demonstrate how a modest amount of high-quality native preferences improves cultural awareness across various LMs, outperforming larger generic preference data. Our analyses reveal that models with stronger initial cultural performance benefit more from alignment, leading to gaps among models developed in different regions with varying access to culturally relevant data. CARE is publicly available at https://github.com/Guochry/CARE.

504.05154v5 [cs.CL]



(EMNLP 2025)

Multi-turn AI-human Reasoning

Granular Privacy Control for Geolocation with Vision Language Models

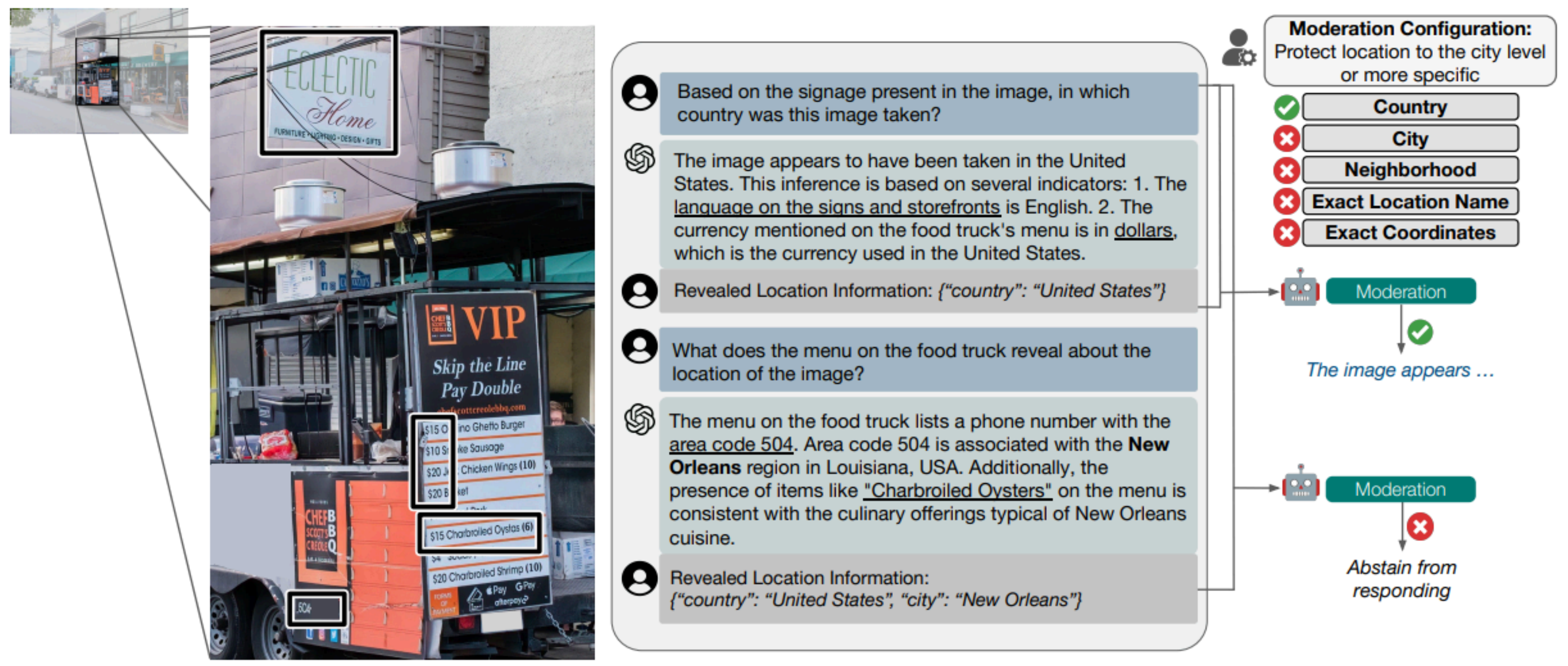
Ethan Mendes[1] Yang Chen[1] James Hays[1] Sauvik Das[2] Wei Xu[1] Alan Ritter[1]

[1] Georgia Institute of Technology [2] Carnegie Mellon University

{emendes3, yangc, hays}@gatech.edu sauvik@cmu.edu {wei.xu, alan.ritter}@cc.gatech.edu

Abstract range of emergent capabilities, such as identify-

What if we have human-in-the-loop in the reasoning process?



(EMNLP 2024)

NLP X Research Lab

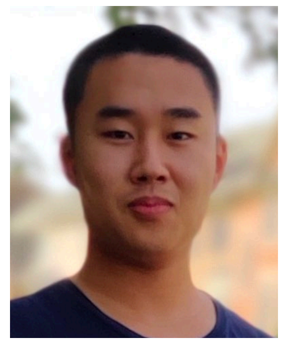
Georgia Tech Machine Learning

Large Language Models
- multilingual multicultural adaptation
- inference-time algorithms
- privacy, safety
- reasoning

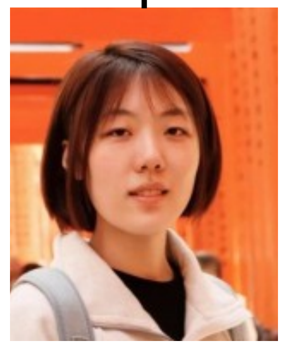
Generative AI
- evaluation of LLM-generated texts
- long-context, multi-doc summarization
- user simulation

NLP+X Interdisciplinary Research
- Education, Healthcare, Law

(co-advised with Alan Ritter)

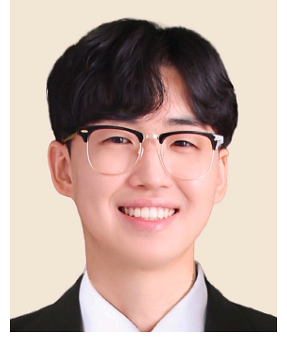

Yao Dou
PhD student


Tarek Naous
PhD student


Geyang Guo
PhD student

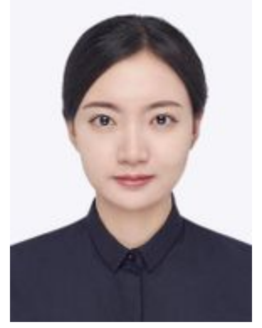

Jonathan Zheng
PhD student


Duong Minh Le
PhD student

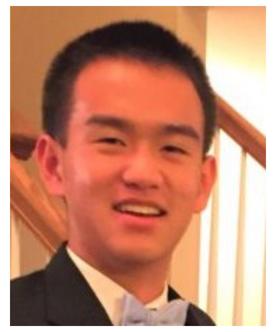

Junmo Kang
PhD student

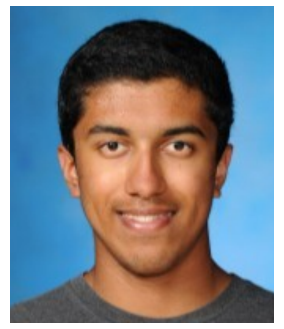

Zirui Shao
Visiting PhD student

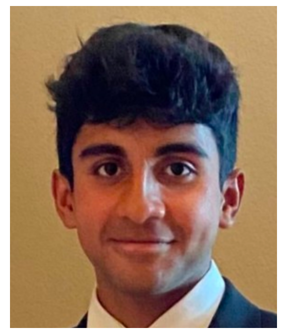

Usneek Singh
MS student

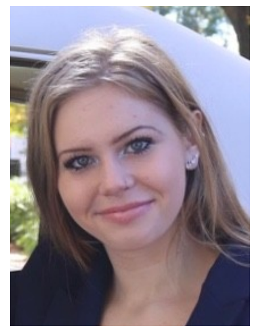

Yiren Wang
MS student


Julie Young
BSMS


Jerry Zheng
BSMS


Joseph Thomas
BSMS


Govind Ramesh
BSMS


Katerina Addington
Undergrad


Alexey Plagov
Undergrad


Benjamin Mamut
Undergrad

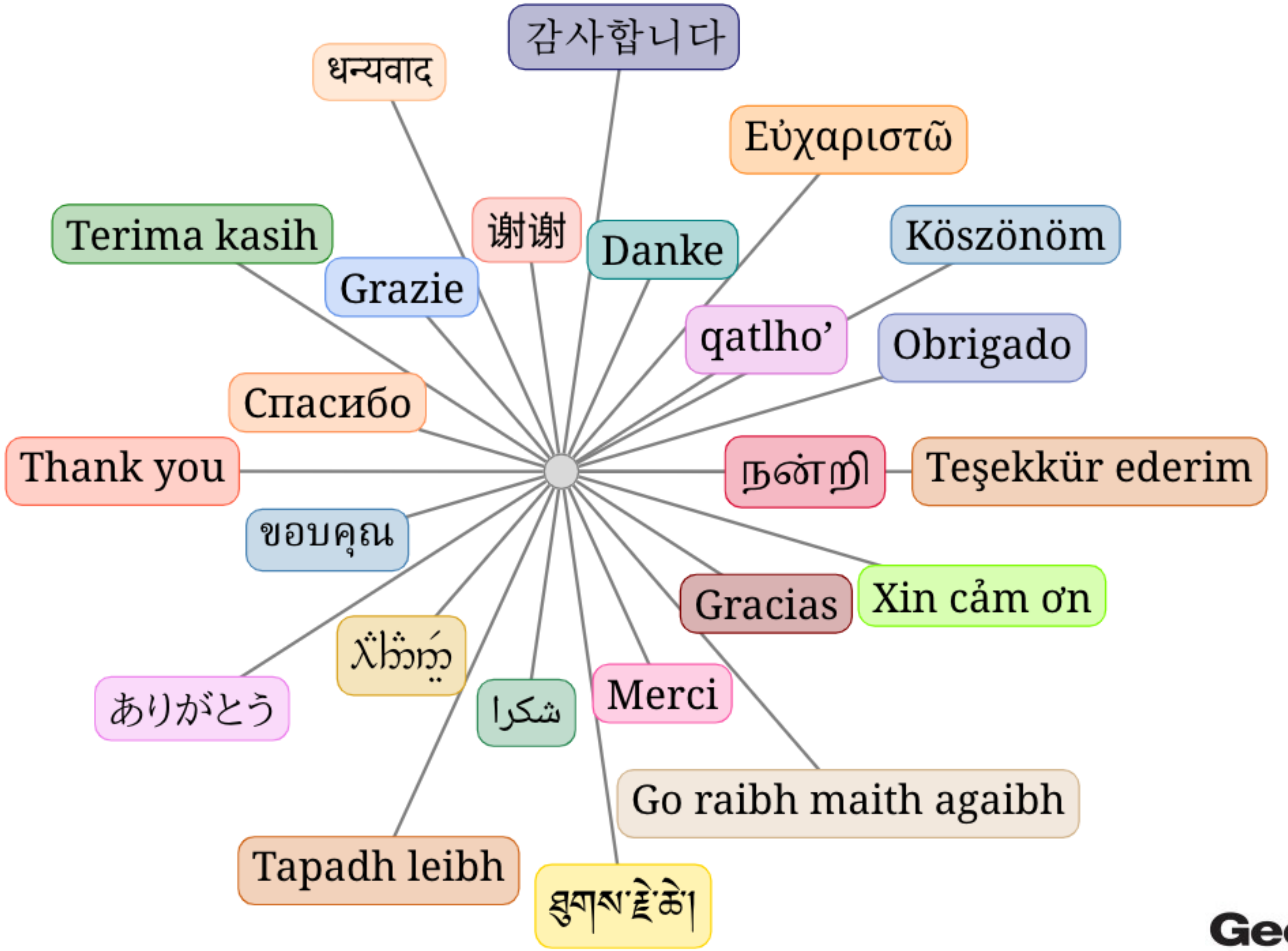

Frank Chang
Undergrad


Jiayu Liu
Intern

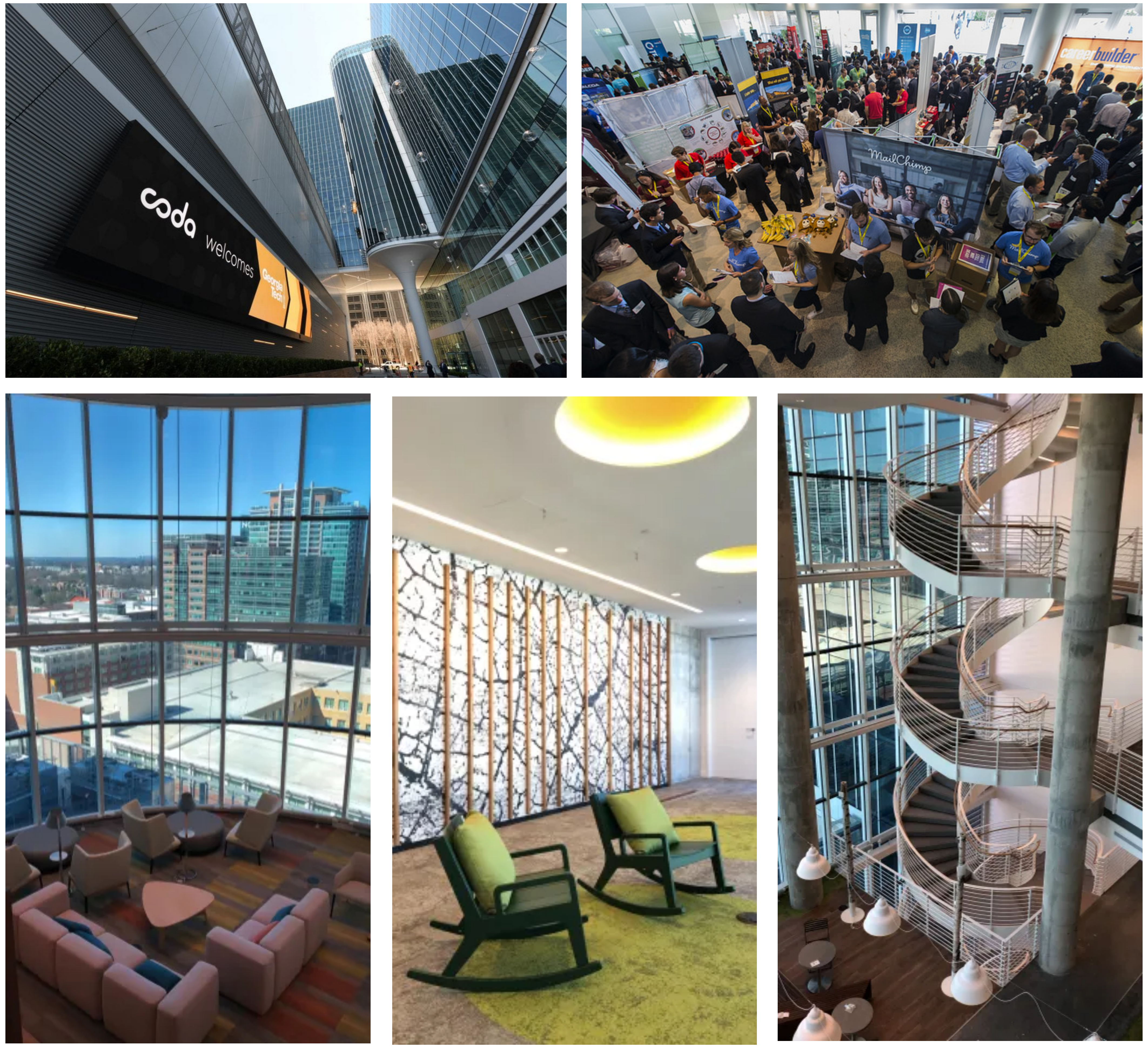
NSF IARPA DARPA NIH Google SONY criteo

Thank you!

https://cocoxu.github.io/



(image credit: Overleaf!)



(image credit: Georgia Tech)

Georgia Tech Machine Learning
NSF
IARPA
DARPA
NIH
Google
SONY
criteo