



(Image credit: ChatGPT)

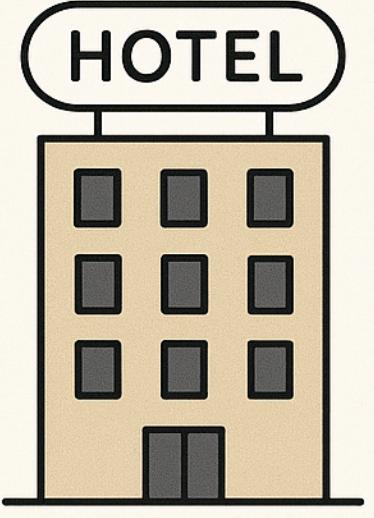
Beyond Logic and Math: Probabilistic Reasoning for Real-World Decision Making

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



So ... what did you reason about today?

So ... what did you reason about today?

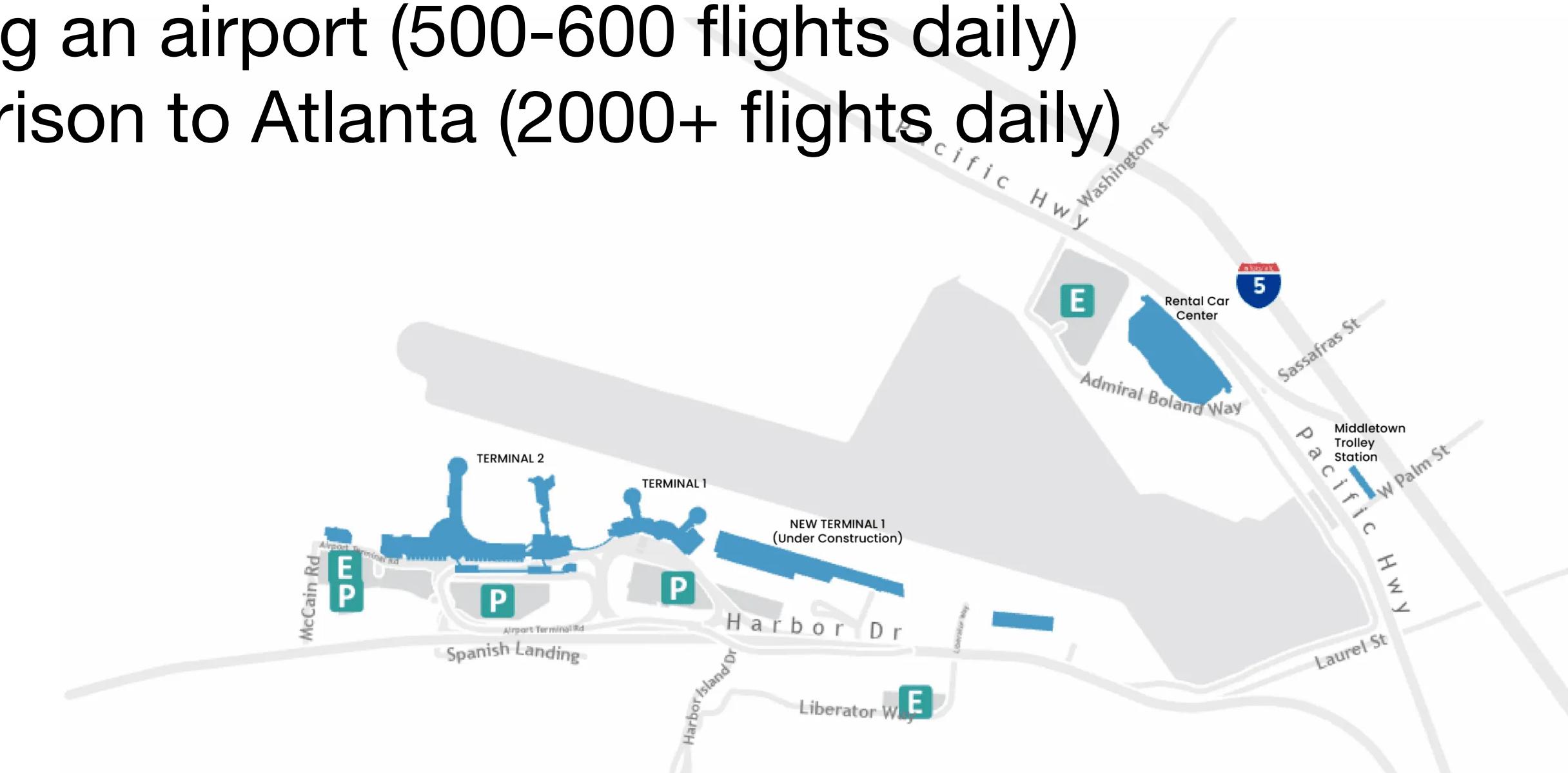


When should I leave my hotel near San Diego Convention Center tomorrow morning for the airport?

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

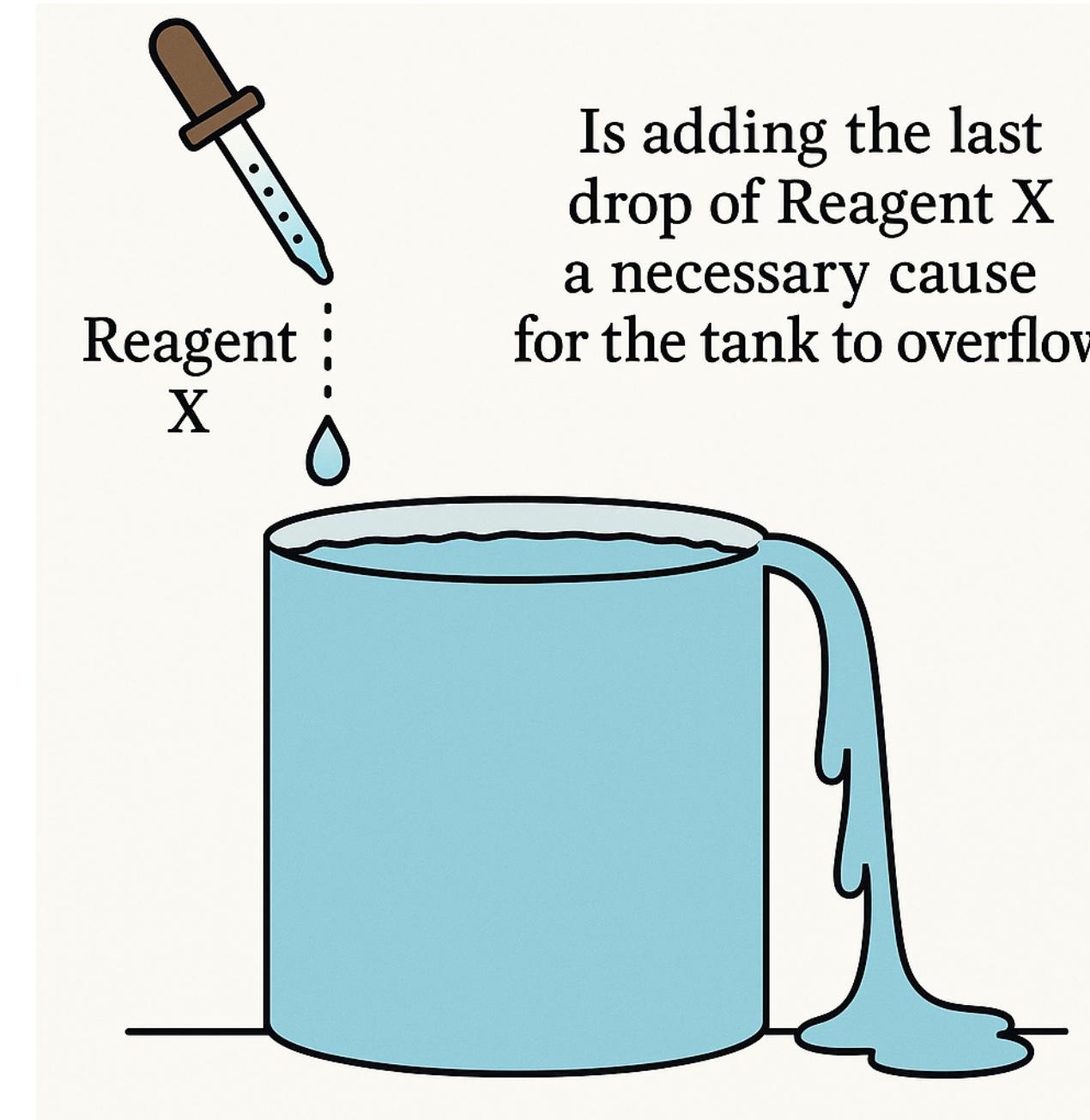


Your Daily Reasoning
(as a normal human being)

This vs. Current LLM reasoning benchmarks



Your Daily Reasoning
(as a normal human being)

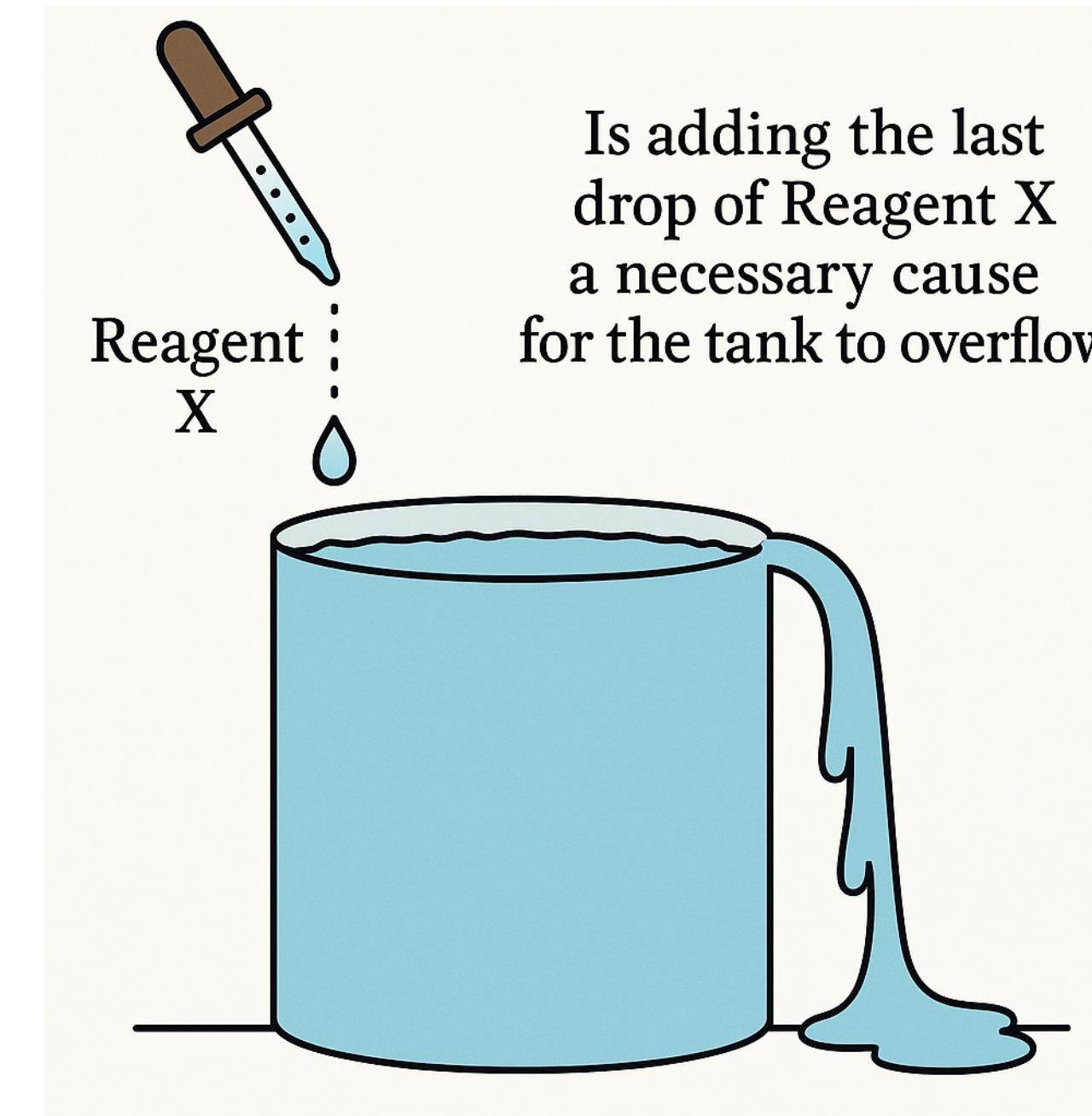


BIG-Bench Extra Hard
(used in OLMo3 evaluation)

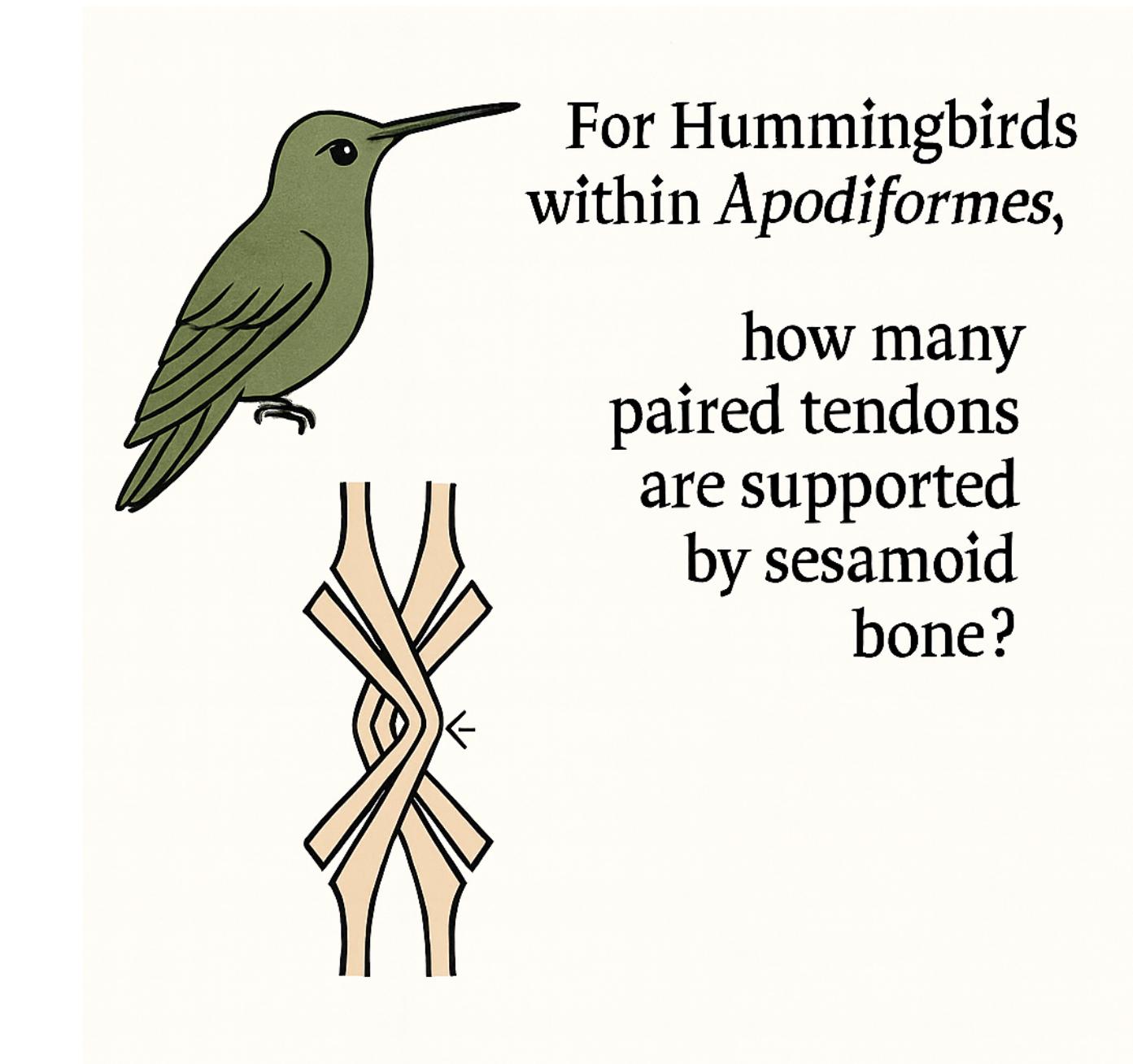
This vs. Current LLM reasoning benchmarks



Your Daily Reasoning
(as a normal human being)

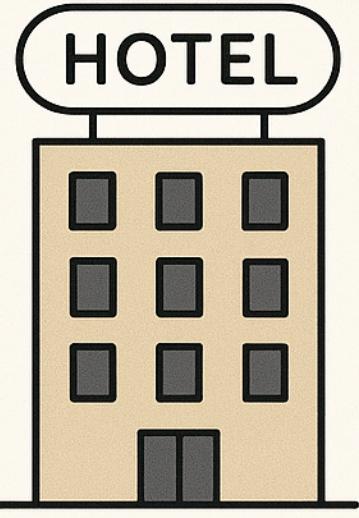


BIG-Bench Extra Hard
(used in OLMo3 evaluation)



Humanity's Last Exam
(used in Gemini3 evaluation)

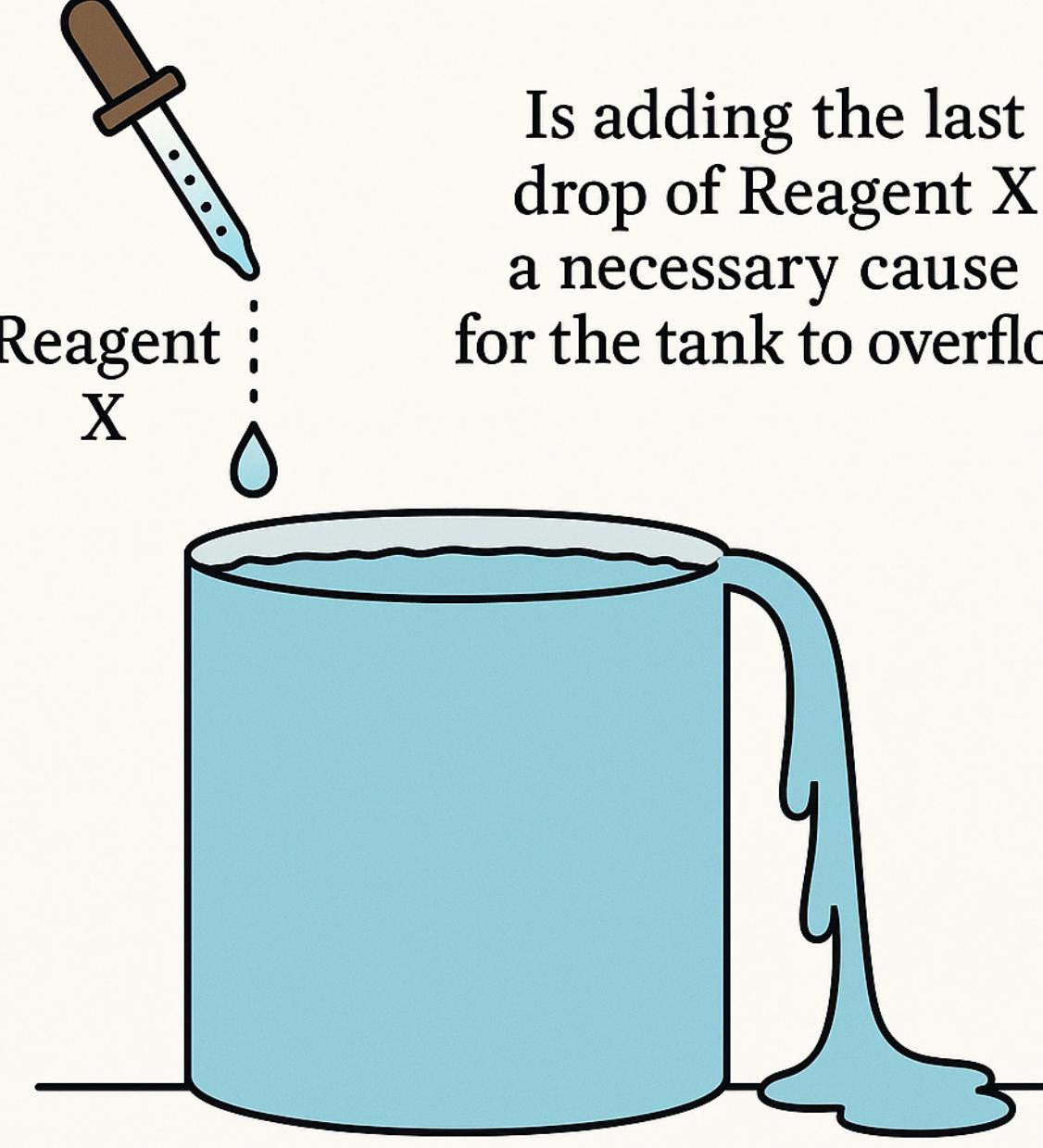
What's the difference?



When should I leave my hotel near San Diego Convention Center tomorrow morning for the airport?

Your Daily Reasoning
(as a normal human being)

Numerical, with some uncertainty



Is adding the last drop of Reagent X a necessary cause for the tank to overflow?

BIG-Bench Extra Hard
(used in OLMo3 evaluation)

Math/Logic tests on toy examples



For Hummingbirds within *Apodiformes*, how many paired tendons are supported by sesamoid bone?

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?

Legal Decision

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

Specialty sections:

Are Jurors Intuitive Statisticians? Bayesian Causal Reasoning in Legal Contexts

Tamara Shengelia^{1*} and David Lagnado²

¹ Department of Experimental Psychology, University College London, London, United Kingdom, ² Department of Experimental Psychology, University College London, London, United Kingdom

Decision-making in legal contexts often involves a degree of uncertainty and decision-making bias. This study examined how jurors move 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.

Marketing



W I R E D

8 NEWSLETTERS SUBSCRIBE

WILL KNIGHT BUSINESS JUN 30, 2025 9:00 AM

Microsoft Says Its New AI System Diagnosed Patients 4 Times More Accurately Than Human Doctors

The tech giant poached several top Google researchers to help build a powerful AI tool that can diagnose patients and potentially cut health care costs.



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

The government can build two charging stations: one in Area A (A1 or A2) and one in Area B (B1 or B2). They must choose either (A1,B2) or (A2,B1) to ensure balanced coverage and avoid overlap. Which option should they choose?

Location Characteristics	OpenAI o1 Confidence	OpenAI o1 Ranking	BIRD
Location A1 is situated on a busy highway in an area with a high concentration of EVs but currently lacks any existing charging stations.	90%	1	86.2 %
Location A2 is on a busy highway with no existing charging stations.	90%	2	74.3%
Location B1 is off main travel routes and has sufficient infrastructure to support a charging station.	60%	3	65.5 %
Location B2 is off main travel routes and has numerous nearby amenities for rest and convenience.	60%	4	62.6 %
What's your final decision?	I cannot decide.		A1, B2

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!

Probabilistic Reasoning with LLMs for k-anonymity Estimation



Jonathan Zheng



Sauvik Das



Alan Ritter



Wei Xu

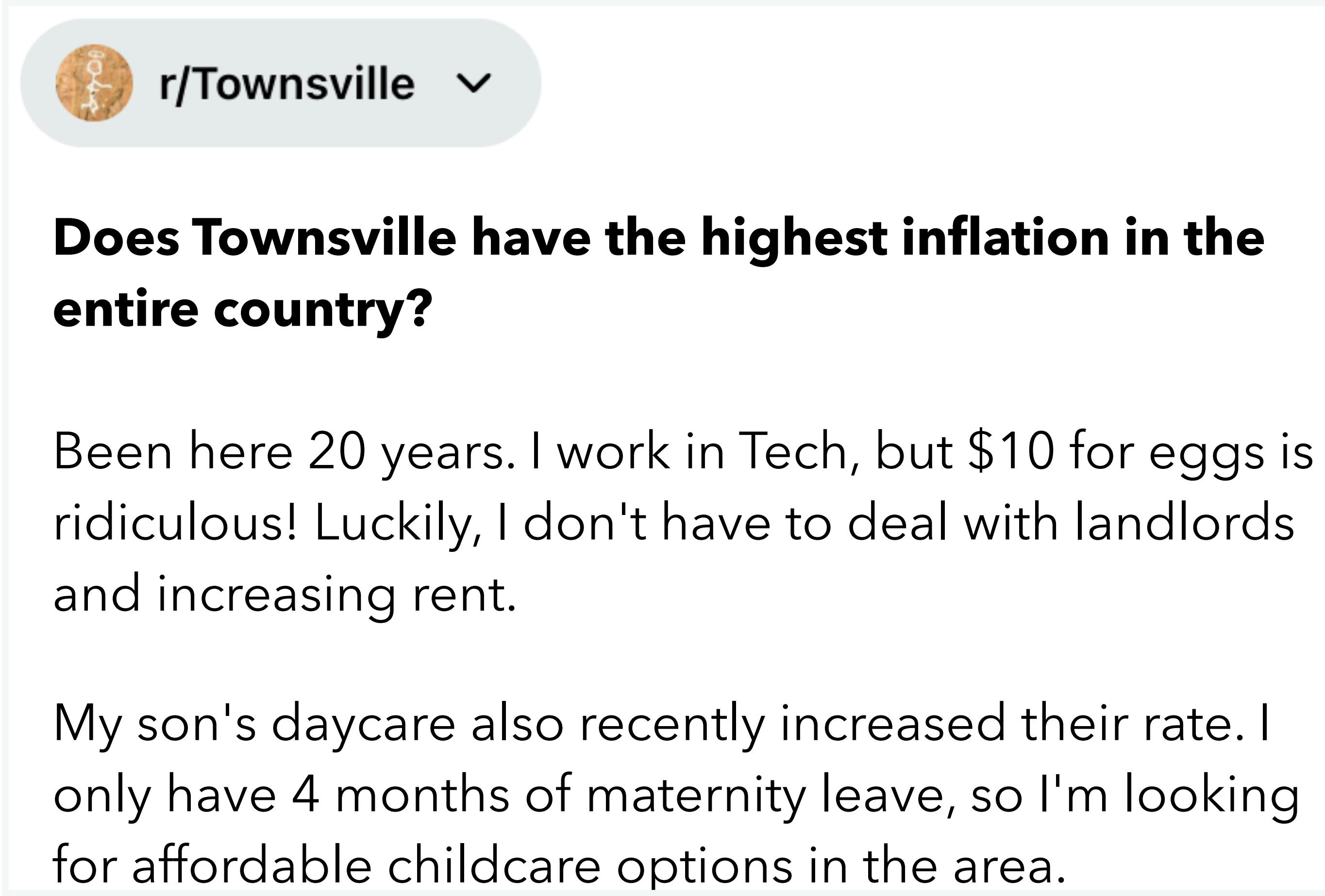


Georgia Tech®

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?



The screenshot shows a post from the subreddit r/Townsville. The post title is "Does Townsville have the highest inflation in the entire country?". The post content includes two paragraphs: "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." and "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."

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.

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.

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

$$204540 \times 0.5 = 102270$$

Working in Tech: Nationally in Australia, about 10% of workers are in tech. Assume similar proportion in Townsville.

$$102270 \times 0.1 = 10227$$

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).

$$7158.9 \times 0.05 = 204.54$$

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>

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

$$204540 \times 0.5 = 102270$$

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.
 $102270 \times 0.1 = 10227$

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).
 $7158.9 \times 0.05 = 204.54$

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>

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
 $204540 \times 0.5 = 102270$

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

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

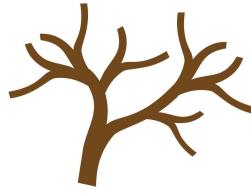
Has a child (with a son in daycare): Around 5% of women are new mothers or have toddlers (0-3 y/o).
 $7158.9 \times 0.05 = 204.54$

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>

In the contrary, the number should significantly further reduce



BRANCH: Probabilistic reasoning w/ Bayesian networks

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

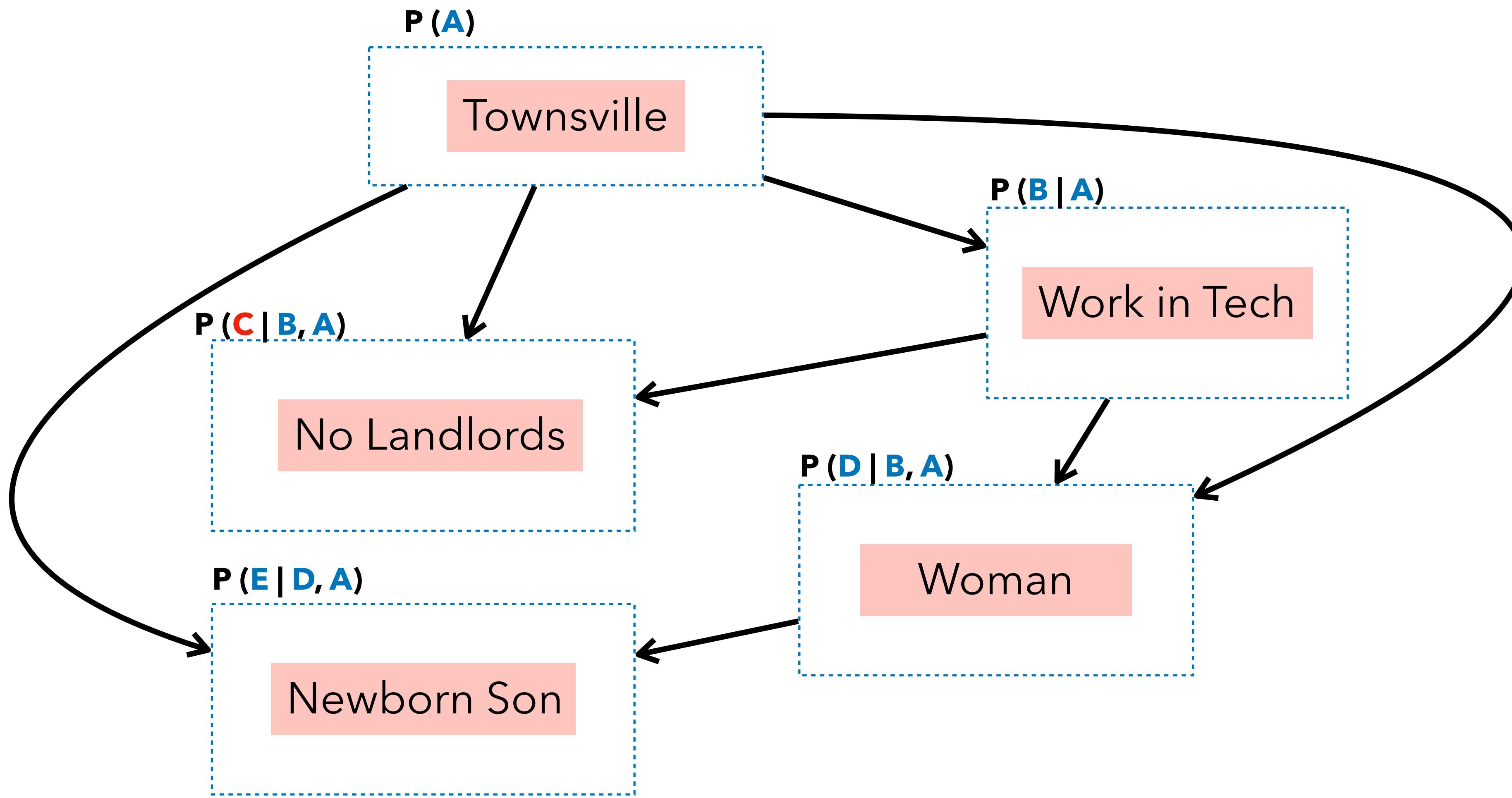
e.g., $P(\text{women} | \text{Work in Tech})$ is easier than $P(\text{Work in Tech} | \text{women})$



BRANCH: Probabilistic reasoning w/ Bayesian networks

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

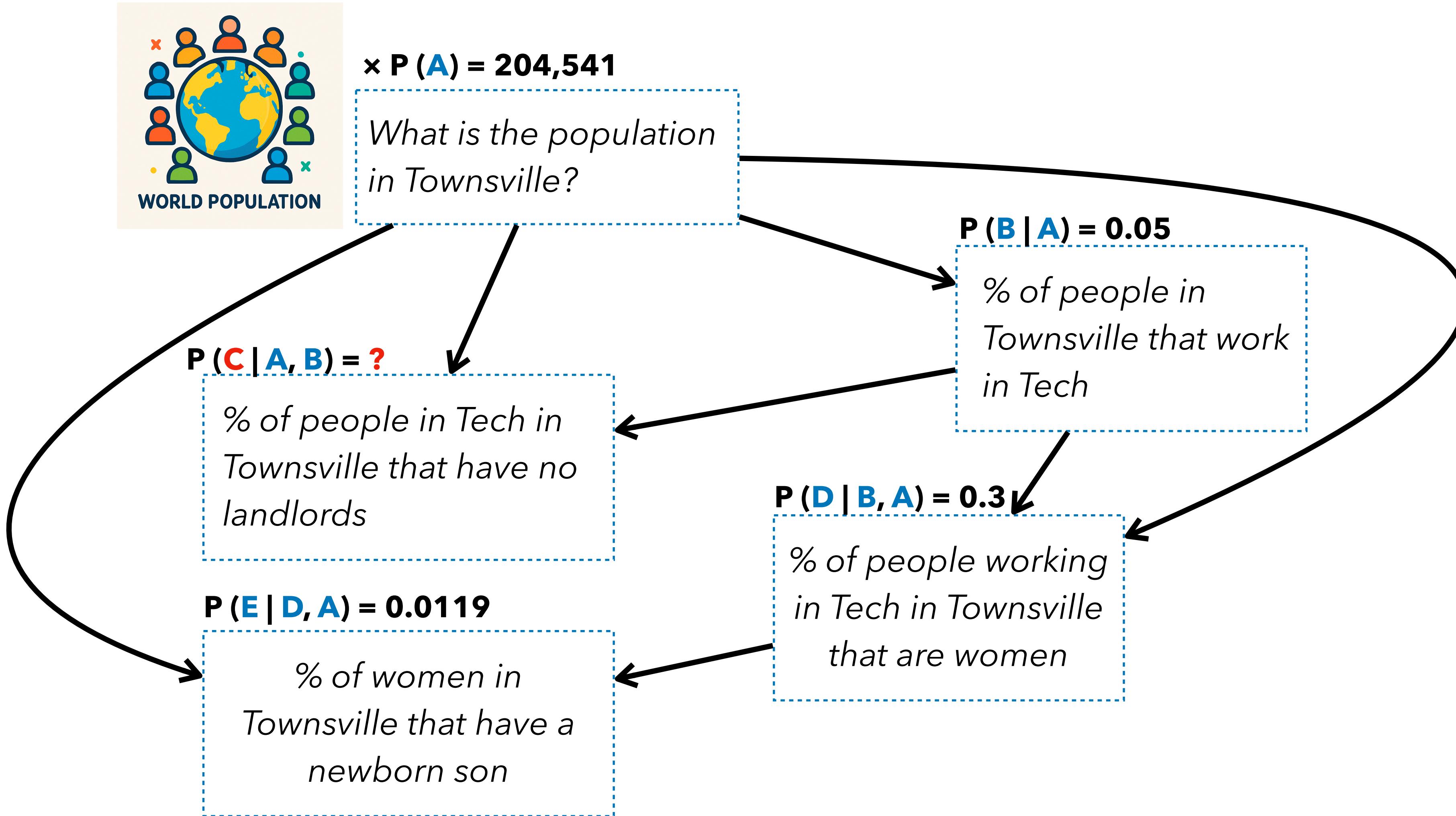
e.g., $P(\text{women} | \text{Work in Tech})$ is easier than $P(\text{Work in Tech} | \text{women})$





BRANCH: Probabilistic reasoning w/ Bayesian networks

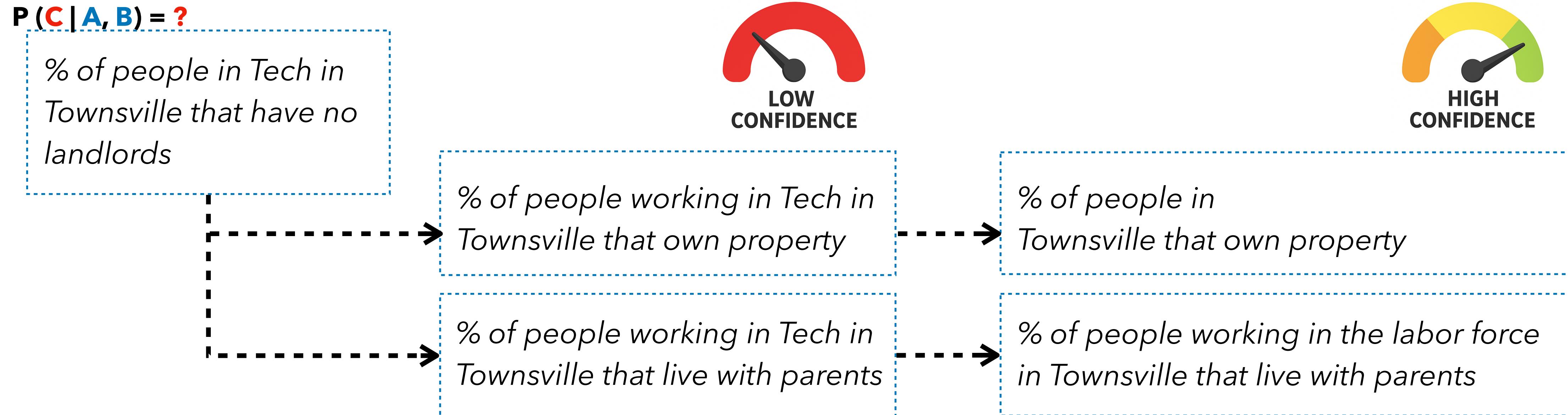
This helps generate questions that LLMs or search engine or human can possibly answer.





BRANCH: Probabilistic reasoning w/ Bayesian networks

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





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



% of people working in Tech in Townsville that own property

% of people working in Tech in Townsville that live with parents



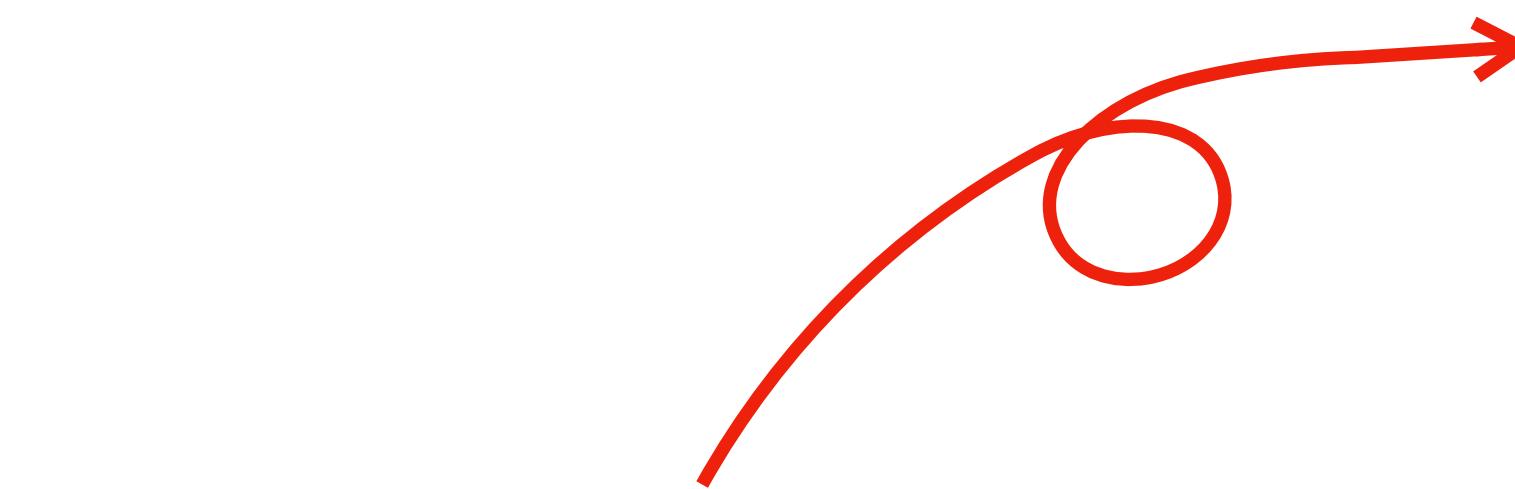
% of people in Townsville that own property

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



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

ChatGPT 4o ▾

To estimate the percentage of women in Townsville who have newborn sons, we can use available demographic and birth data:

1. Total Population of Townsville

As of 2024, Townsville's population is approximately 201,313 .

2. Estimated Number of Women

Assuming a roughly equal gender distribution, about 50% of the population are women:

- $201,313 \times 50\% = 100,656$ women

3. Annual Births in Townsville

In 2023, Townsville University Hospital recorded 2,345 births . Considering this is the primary birthing facility in the region, we'll use this number for our estimate.

4. Number of Newborn Sons

Globally, approximately 51% of newborns are male. Applying this to Townsville:

- $2,345 \text{ births} \times 51\% \approx 1,196$ newborn boys

5. Percentage of Women with Newborn Sons

To find the percentage of women who have a newborn son:

- $(1,196 / 100,656) \times 100\% \approx 1.19\%$

Conclusion

Approximately 1.19% of women in Townsville have a newborn son at any given time, based on 2023 data. This estimate assumes:

- All births occur at Townsville University Hospital
- An even gender distribution in the population
- A consistent male birth ratio of 51%

Please note that actual figures may vary due to factors like births at other facilities, home births, and demographic shifts.

W Sources

Ask anything

+ Search Deep research Create image ...

Share More WE PLUS

More

- ABS Australian Bureau of Statistics
Regional population, 2023-24 financial year
March 26, 2025 — Statistics about the population and components of change (births, deaths, migration) fo...
- ABS Australian Bureau of Statistics
Births, Australia, 2023 - Australian Bureau of Statistics
October 15, 2024 — Women aged 30 to 34 years had the highest fertility rate at 105.2 babies per 1,000...
- AIHW Aboriginal and Torres Strait Islander mothers and babies , Birth rate
October 30, 2023 — In comparison the birth rate for non-Indigenous females ranged between 61 per 1,00...
- The Guardian The Australia of Australia: Townsville is the most popular destination ...
May 31, 2023 — Townsville's population was 198,577 as of June 2022, a 1.5% increase on the previous ye...
- population.net.au Townsville Population 2025 ❤️
2023, 202,480, 1.14%. 2024, 204,970, 1.23%. Demographics of Townsville. According to census,...
- ABS Australian Bureau of Statistics
1318.3 - Qld Stats, Nov 2008 - Australian Bureau of Statistics
Brisbane (C) recorded the lowest rate of 1.65 babies per 1,000 women followed by the Gold Coast (C) wit...
- Townsville City Council Living in Townsville
Known as the Second Capital of Queensland, with a population of approximately 198,143 people,...
- Townsville Dashboard Your City - Townsville Dashboard
Townsville is a thriving tropical home to over 200,000 people who interact and enjoy our city in so many...
- Townsville City Council [PDF] City of Townsville
From 2016 to 2021, City of Townsville's population increased by 6,020 people (3.2%). ... Analysis of the...
- W Wikipedia Demographics of Australia - Wikipedia



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		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.

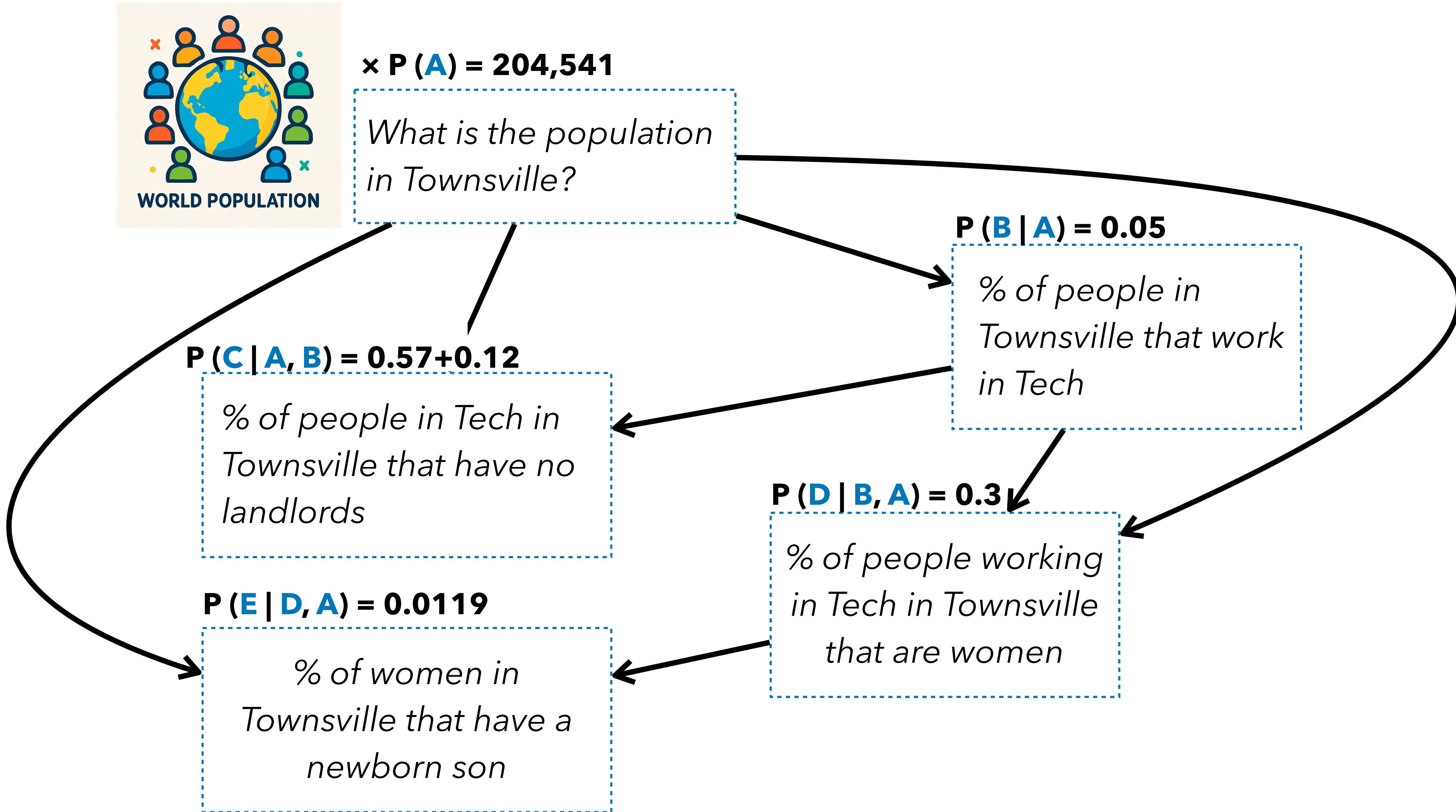
(b) Includes mothers with not stated Hospital and Health Service of usual residence.

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



BRANCH: Probabilistic reasoning w/ Bayesian networks

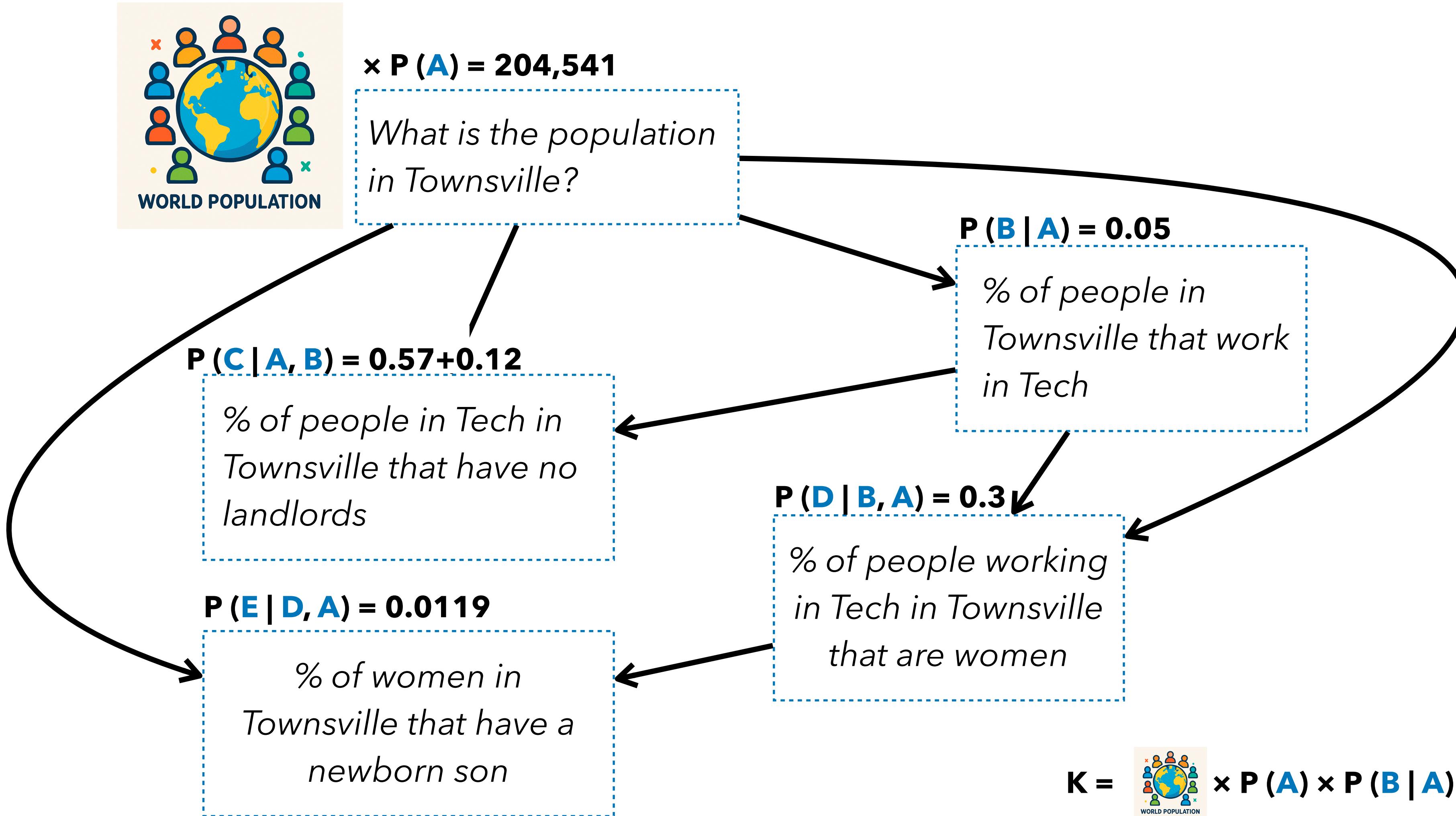
The Bayesian graph reconstructs the individual probability answers to estimate the K -anonymity.





BRANCH: Probabilistic reasoning w/ Bayesian networks

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



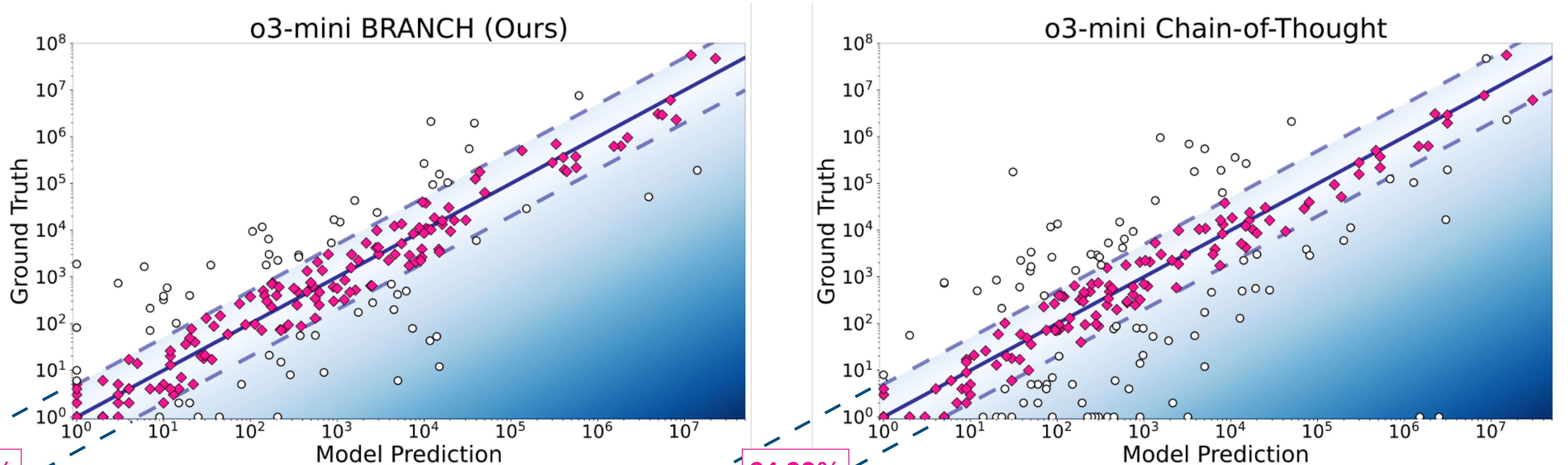


BRANCH vs. Chain-of-Thought

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%*



BRANCH vs. Chain-of-Thought

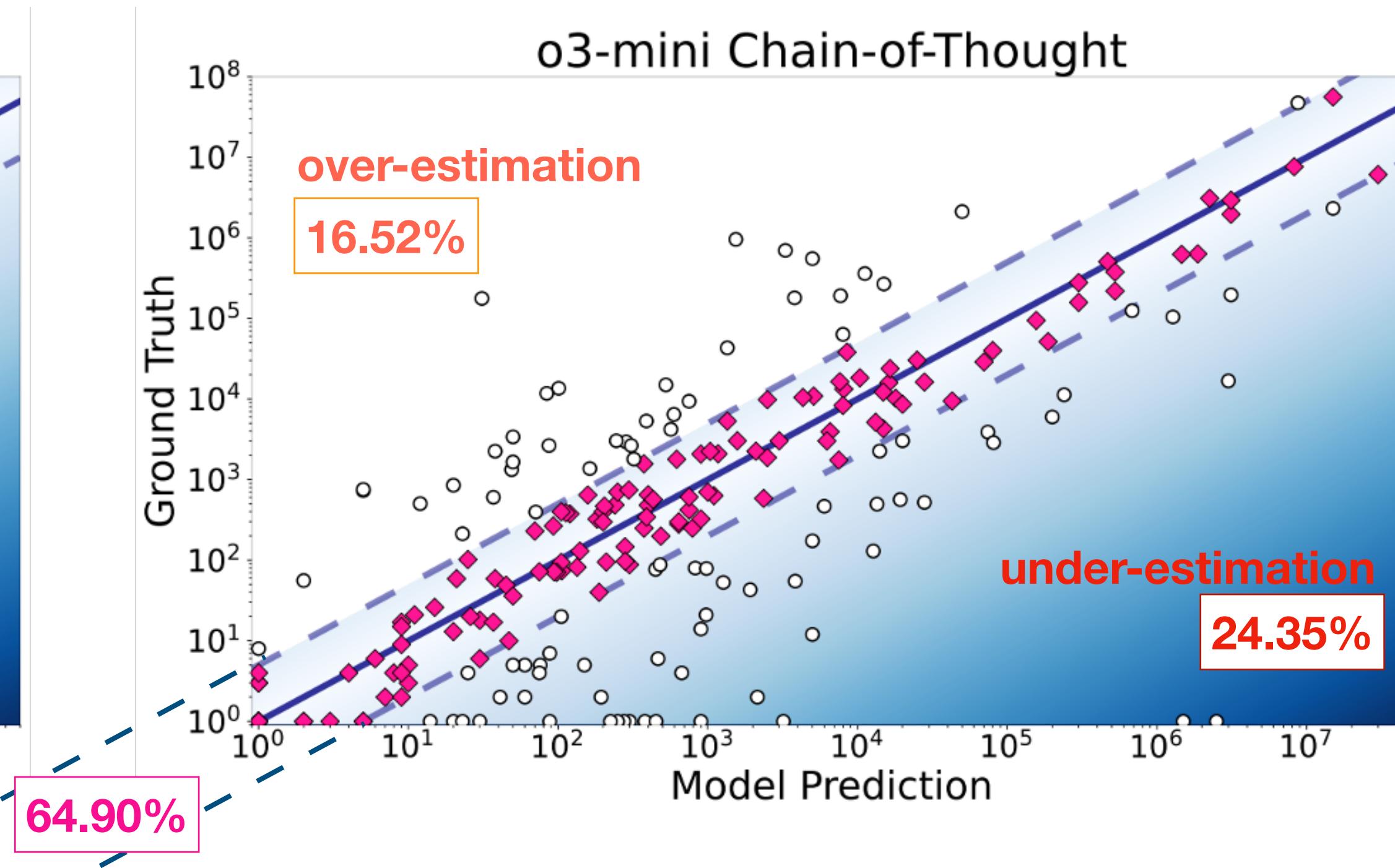
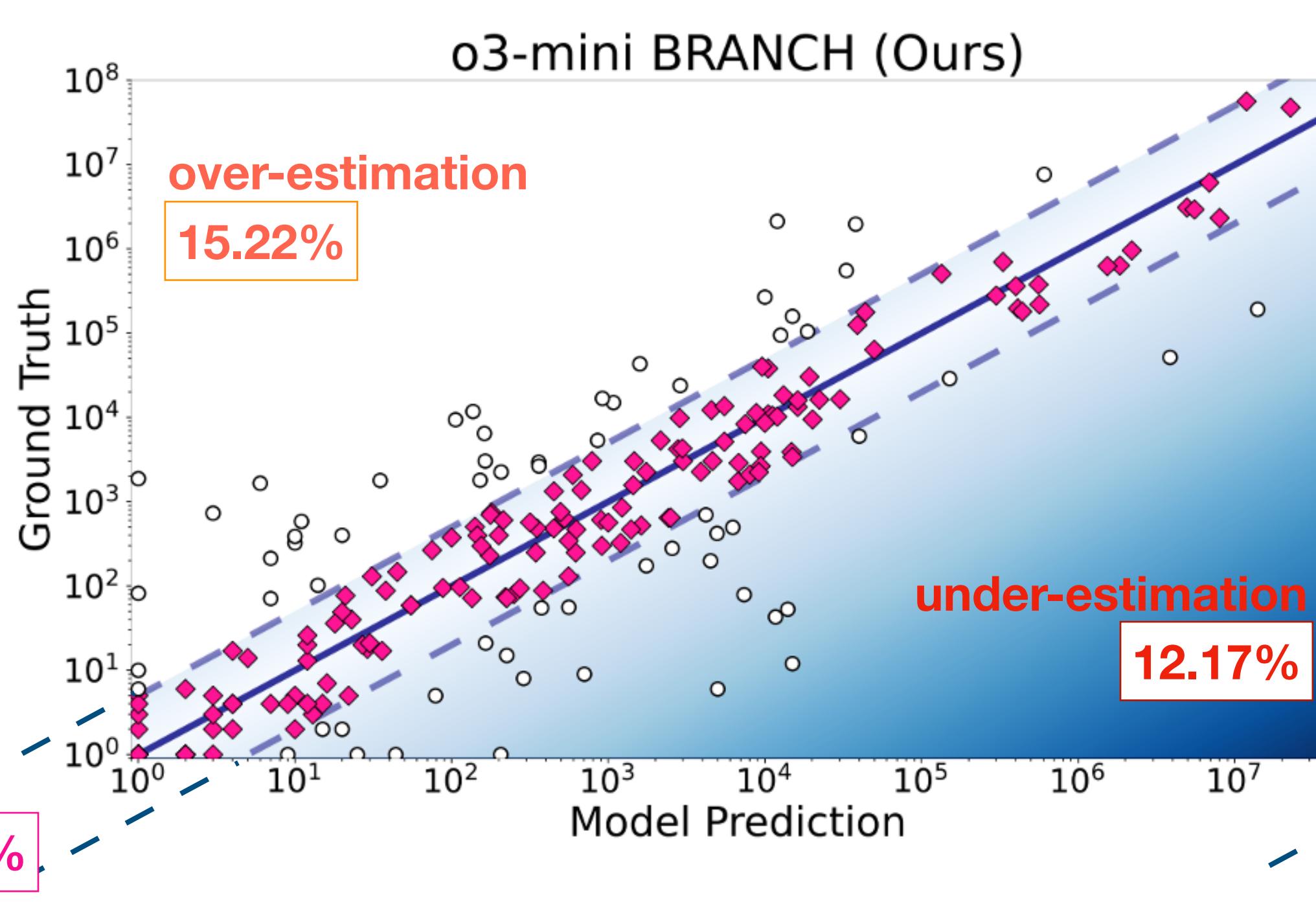


$$RANGE(\{\hat{k}_i\}, \{k_i^*\}) = \frac{1}{n} \sum_{i=1}^n \mathbb{1} \left[\frac{\hat{k}_i}{a} \leq k_i^* \leq a \cdot \hat{k}_i \right] \quad (a = 0.5)$$

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

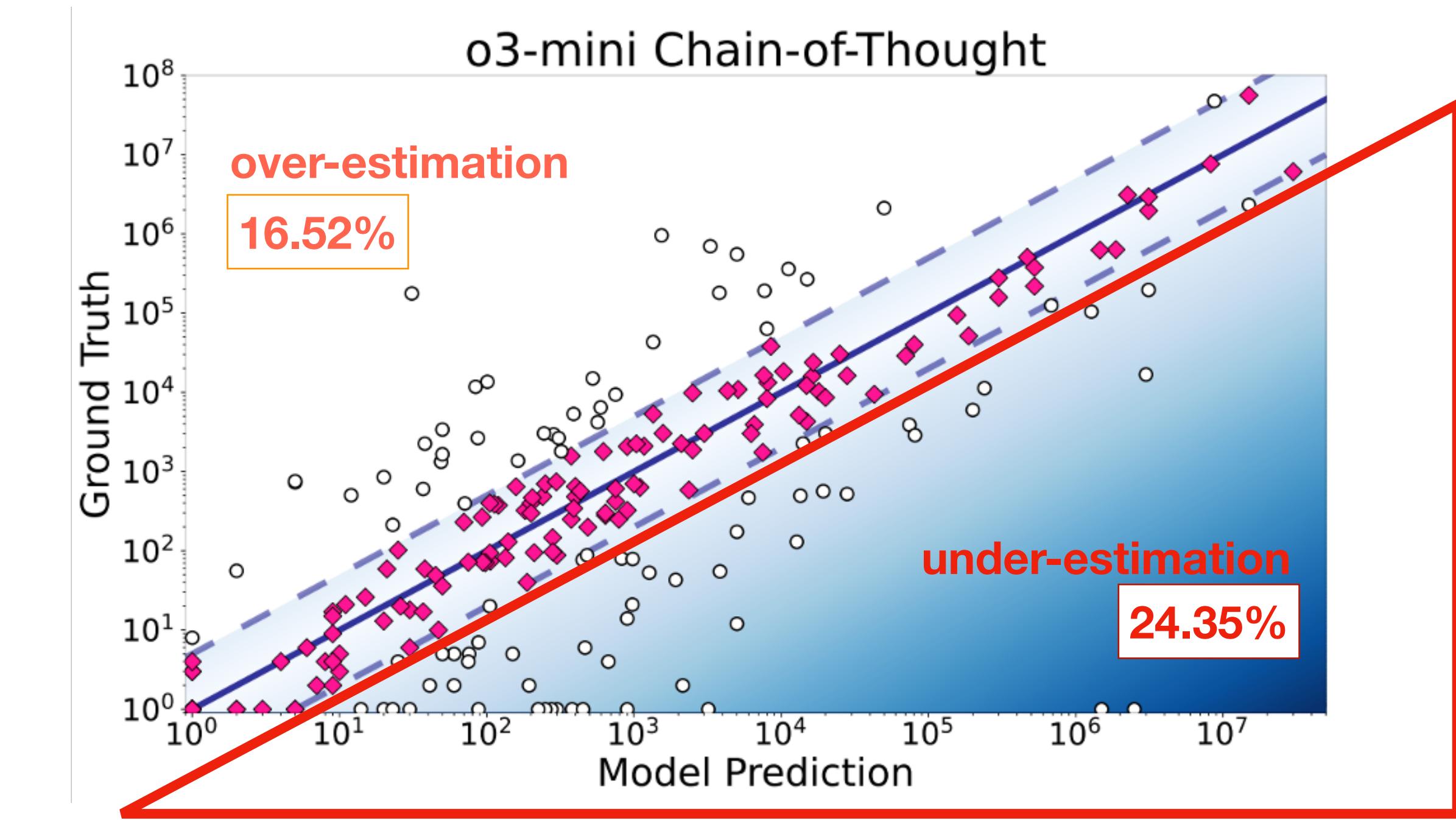
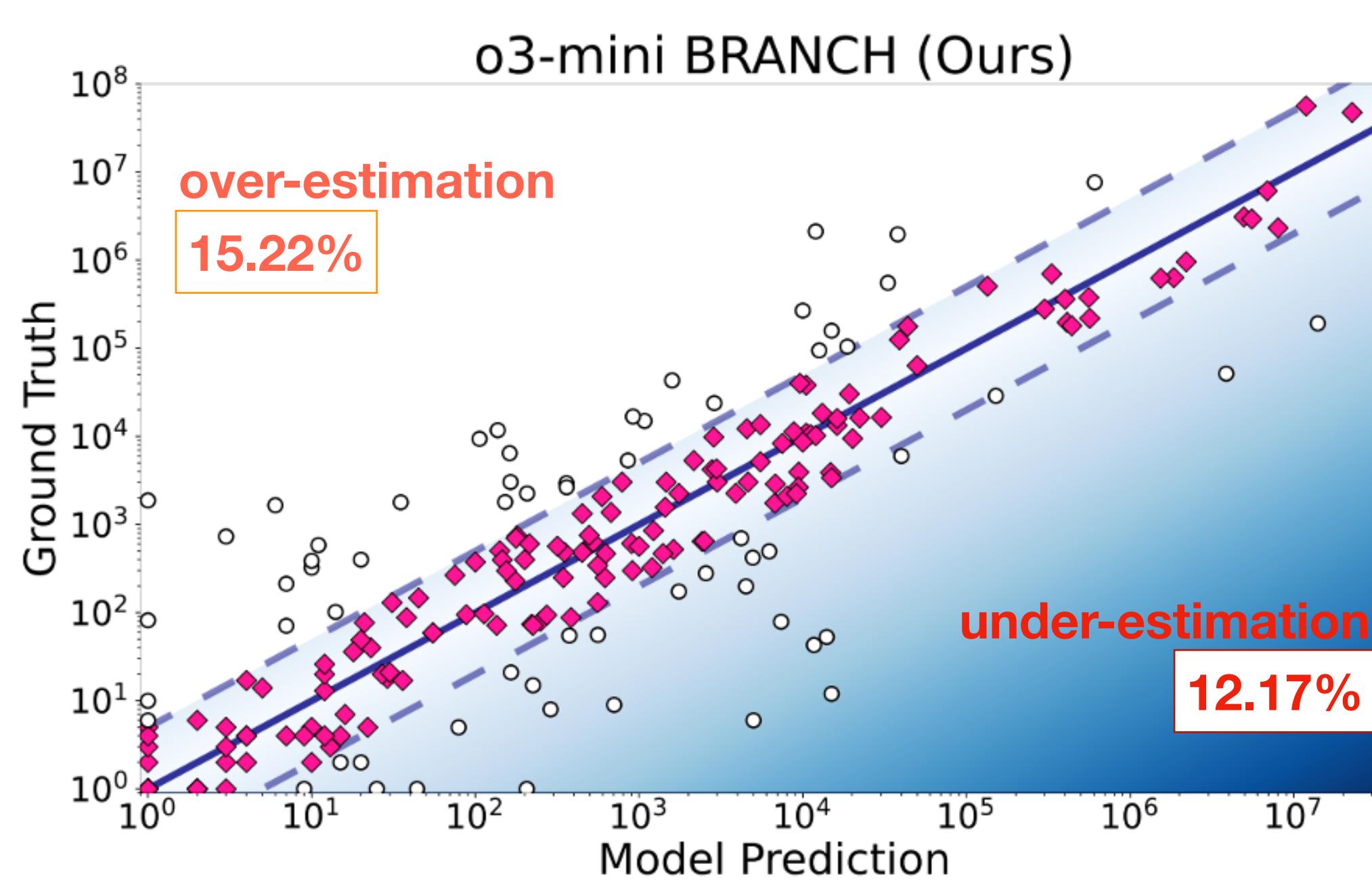


BRANCH vs. Chain-of-Thought





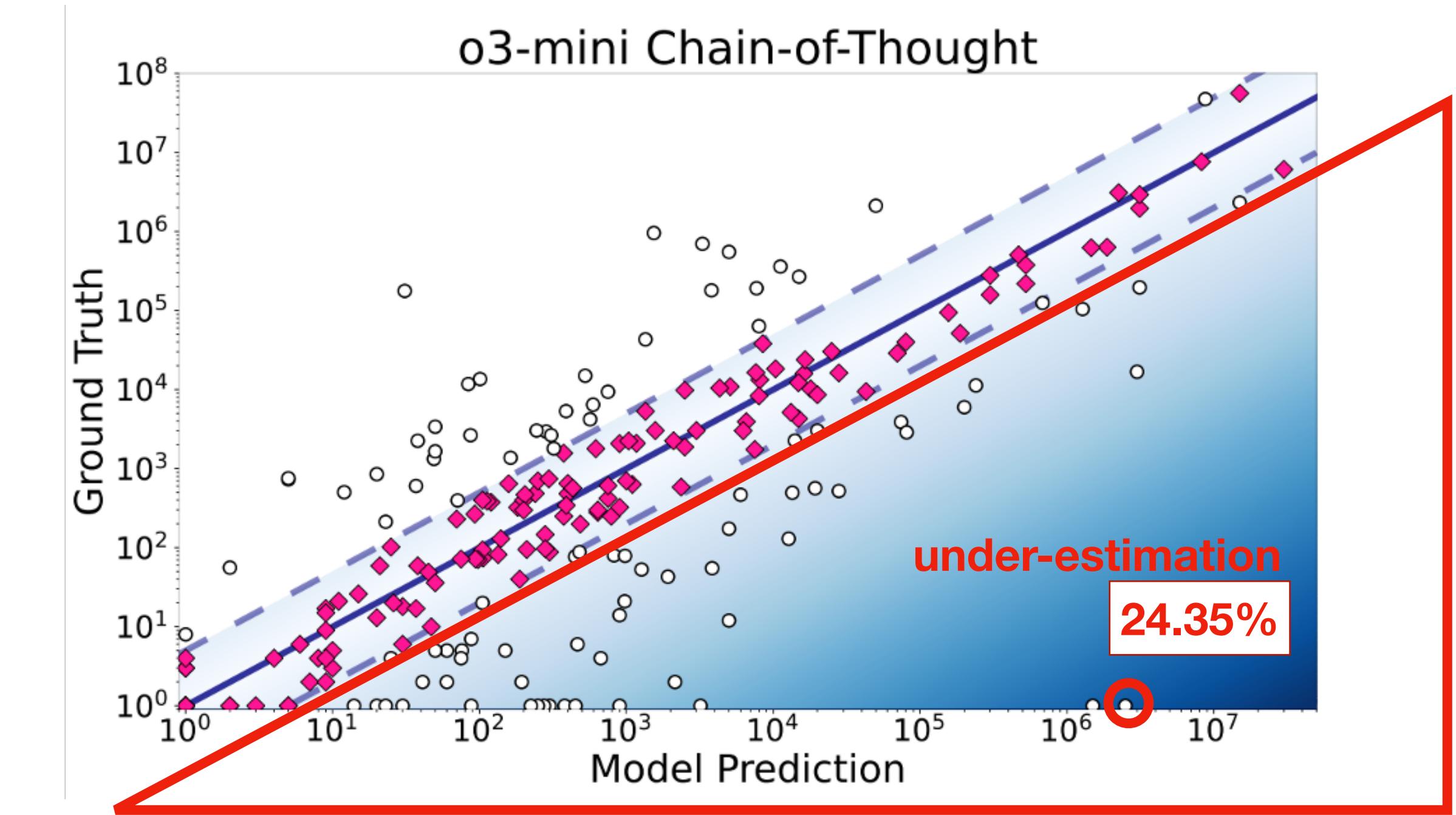
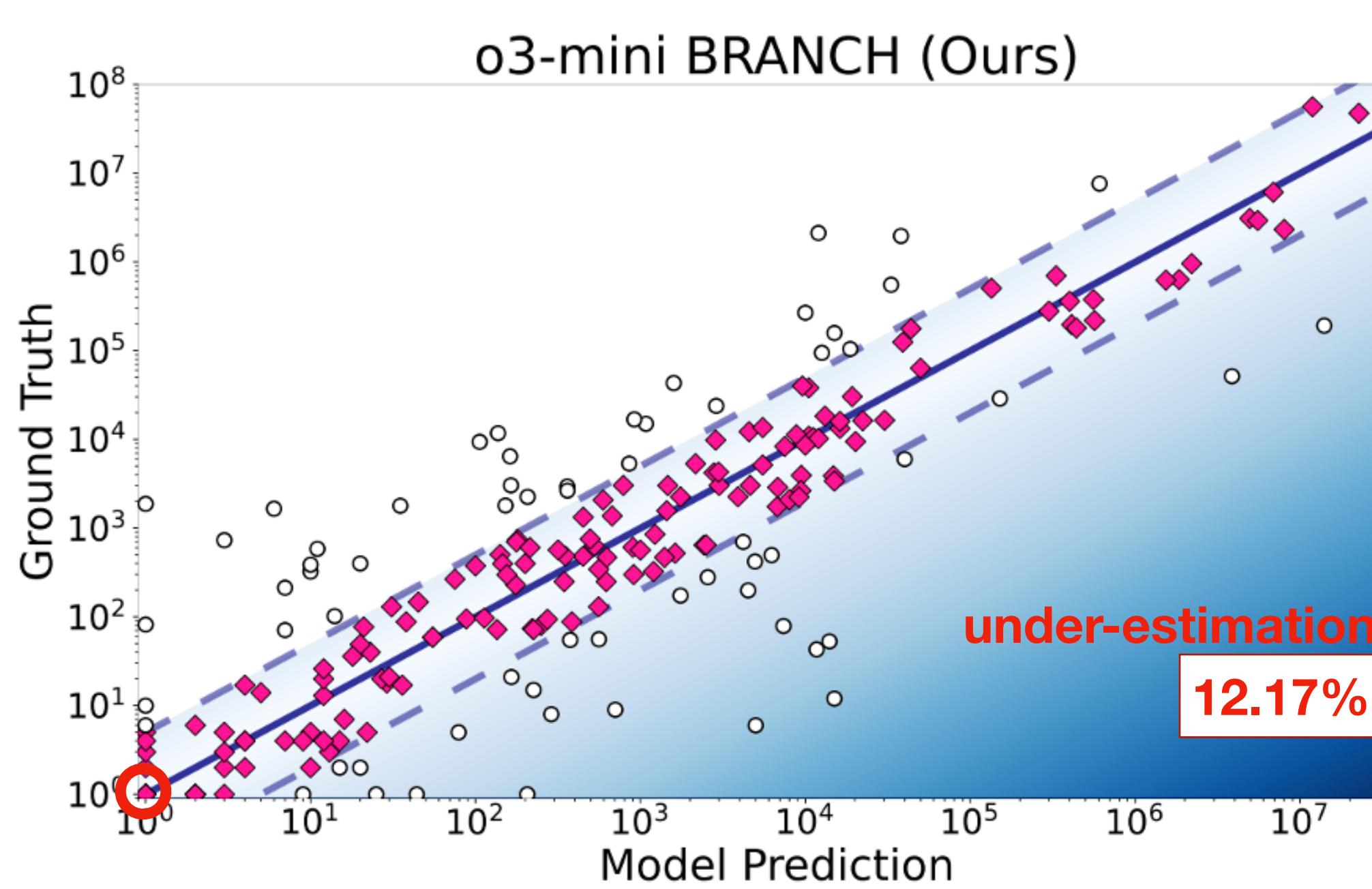
BRANCH vs. Chain-of-Thought



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



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



$K = 1$



$K = 250000$

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

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 $\rho \uparrow$	Log error \downarrow	Within Range \uparrow
🔗 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

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

Best healthcare option for military retirees? 45/300

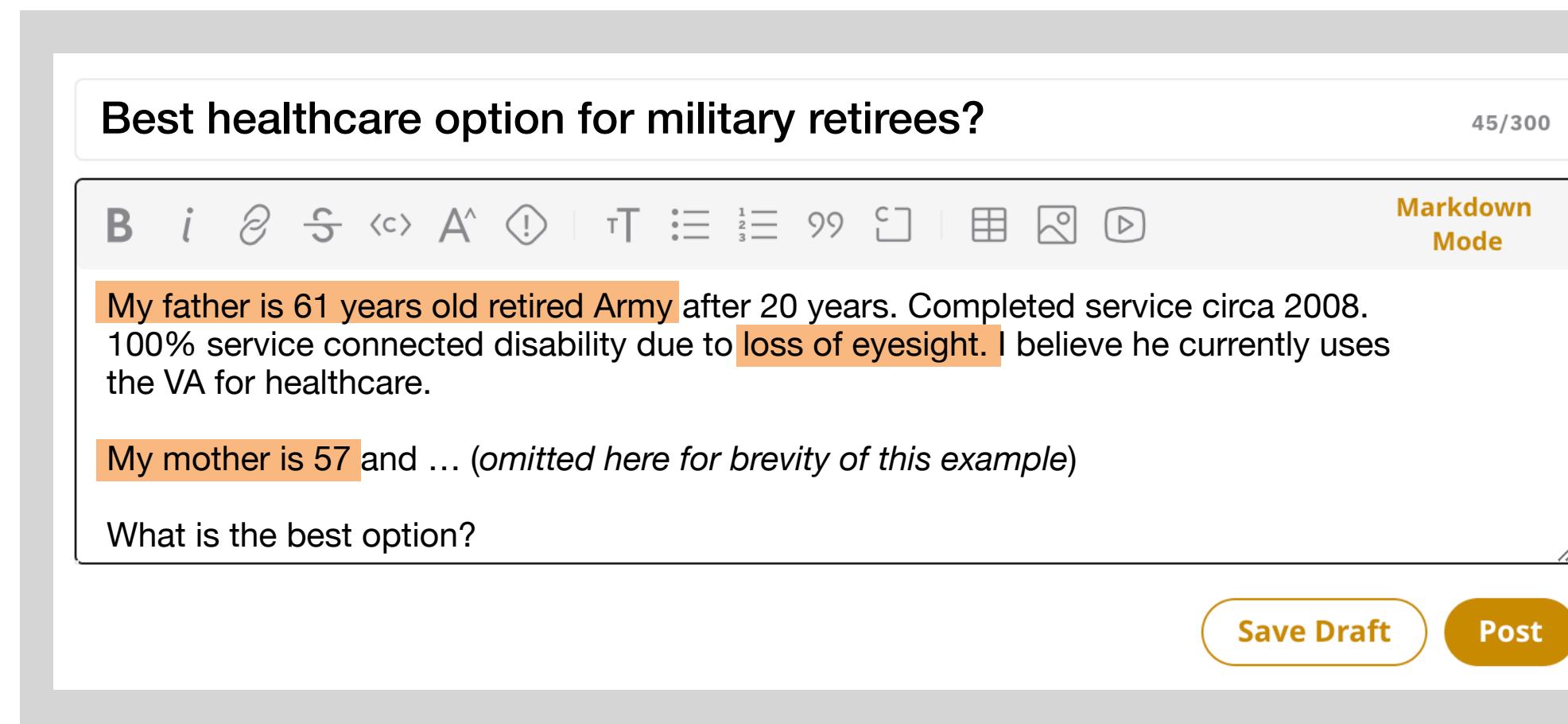
B i ∂ S $\leftarrow\rightleftharpoons$ A[^] ! T : 1 2 3 99 C | # Markdown Mode

My father is 61 years old retired Army after 20 years. Completed service circa 2008. 100% service connected disability due to loss of eyesight. I believe he currently uses the VA for healthcare.

My mother is 57 and ... (omitted here for brevity of this example)

What is the best option?

Save Draft **Post**

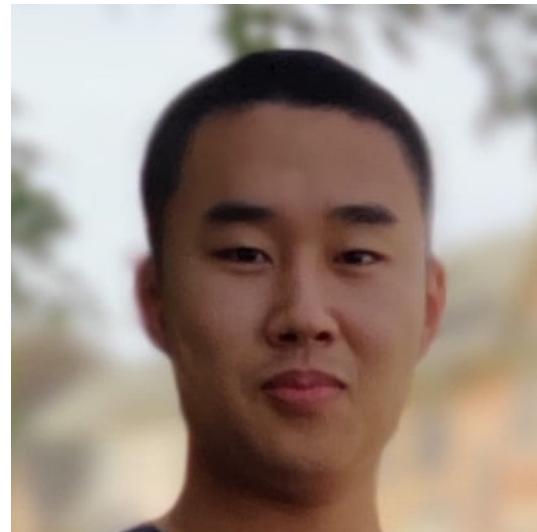


Privacy protection strength: **Fair**



- mentioned family member, occupation, age, and health information all in one post
- you may increase protection strength to **Good** by:
 - changing “retired Army” to “retired service member”
 - removing “(due to) loss of eyesight”
- you may further increase protection strength to **Great** by:
 - changing “61 years old” to “over 60”
 - changing “is 57” to “is over 55”

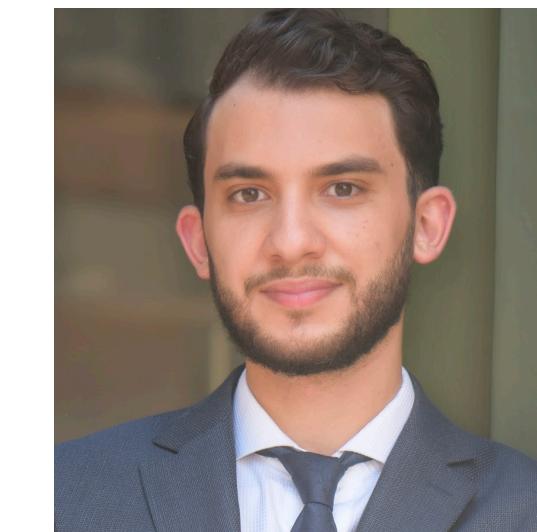
Reducing Privacy Risks in Online Self-Disclosures



Yao Dou



Isadora Krsek



Tarek Naous



Anubha Kabra



Sauvik Das



Alan Ritter



Wei Xu

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

Detecting the disclosures is very important

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

Demographic Attributes

Age Wife/GF

Age&Gender Husband/BF

Race/Nationality Sexual Orientation

Gender Relationship Status

Location Pet

Appearance Contact

Name

Personal Experiences

Occupation

Family

Health

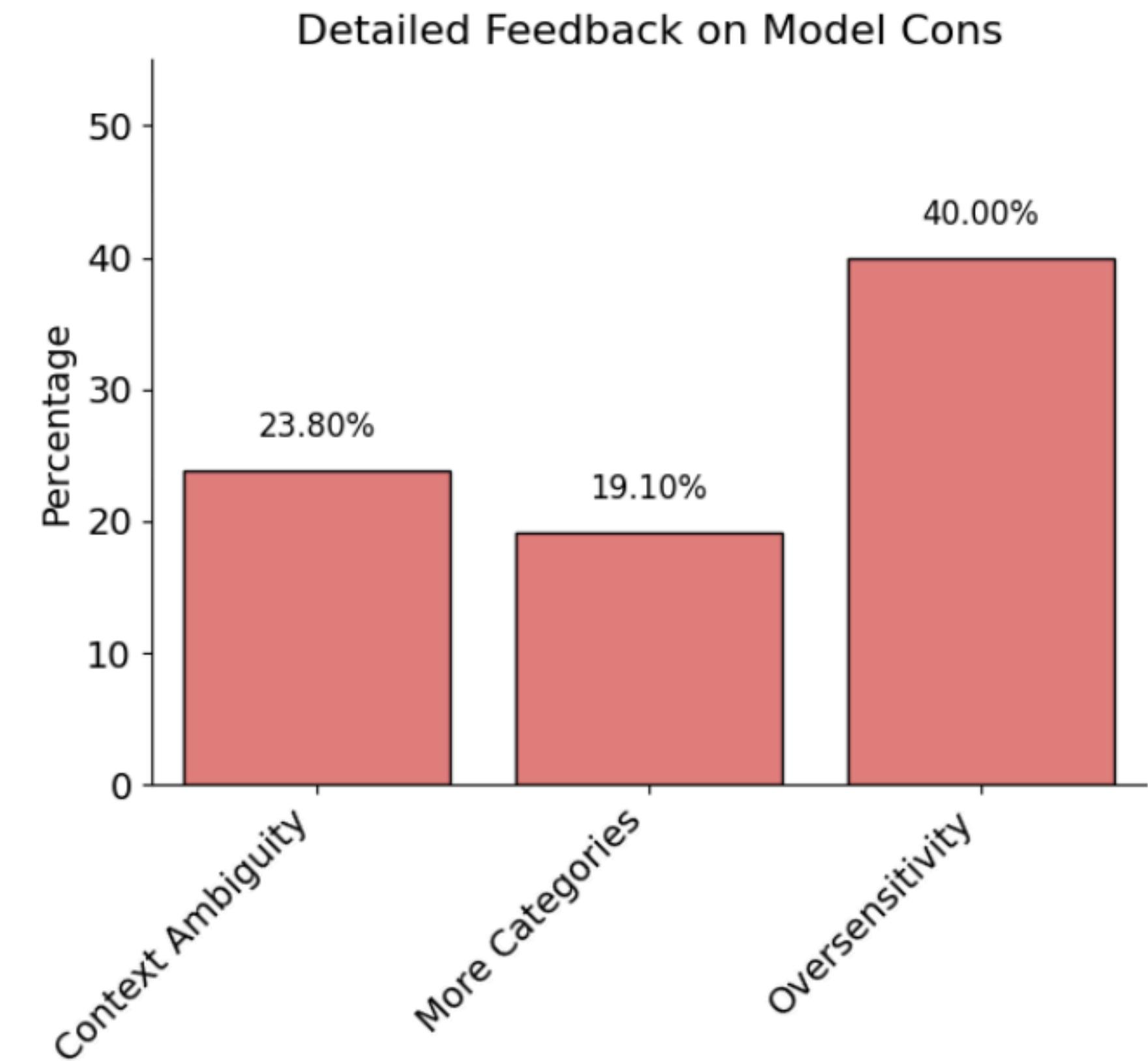
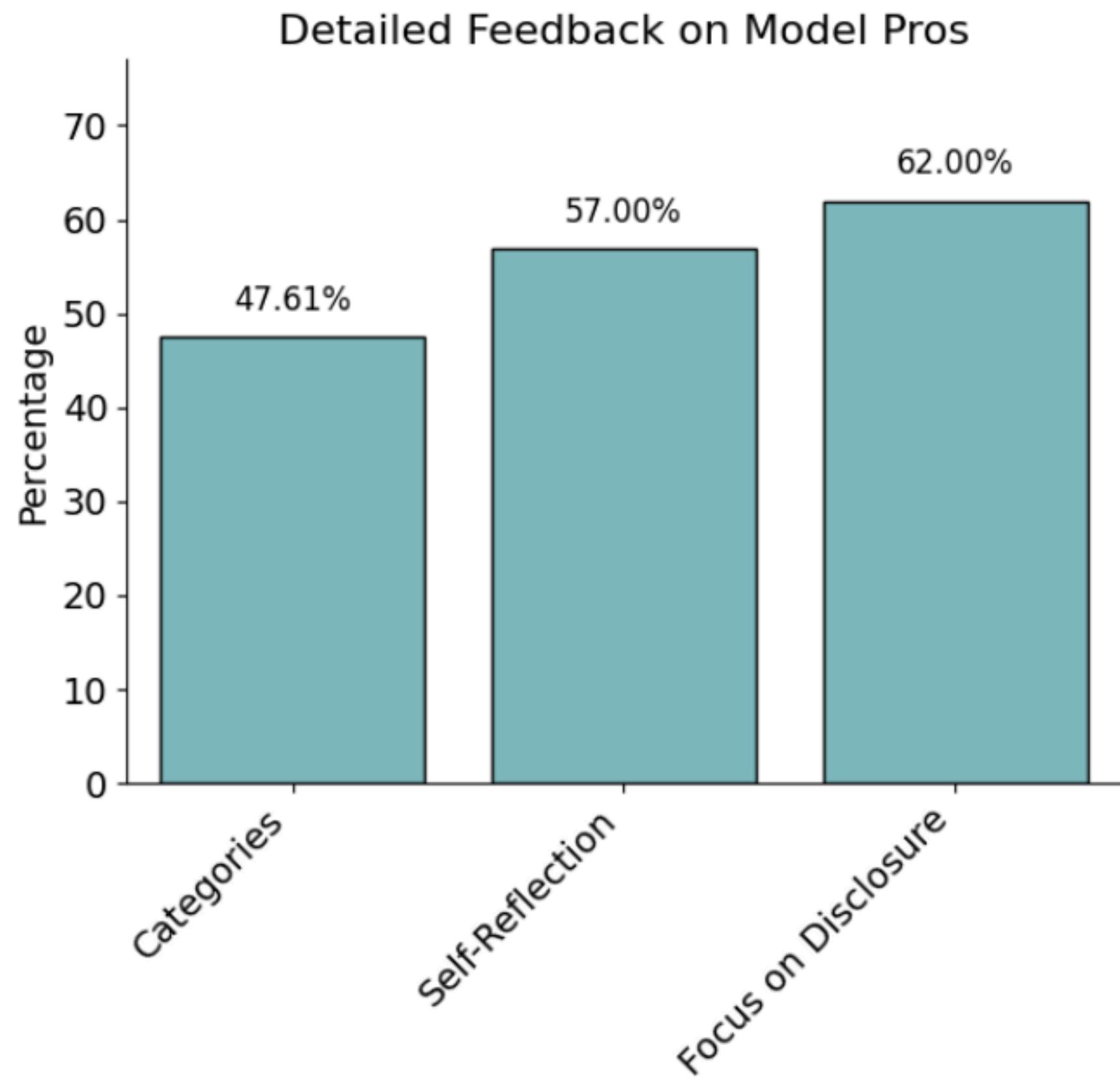
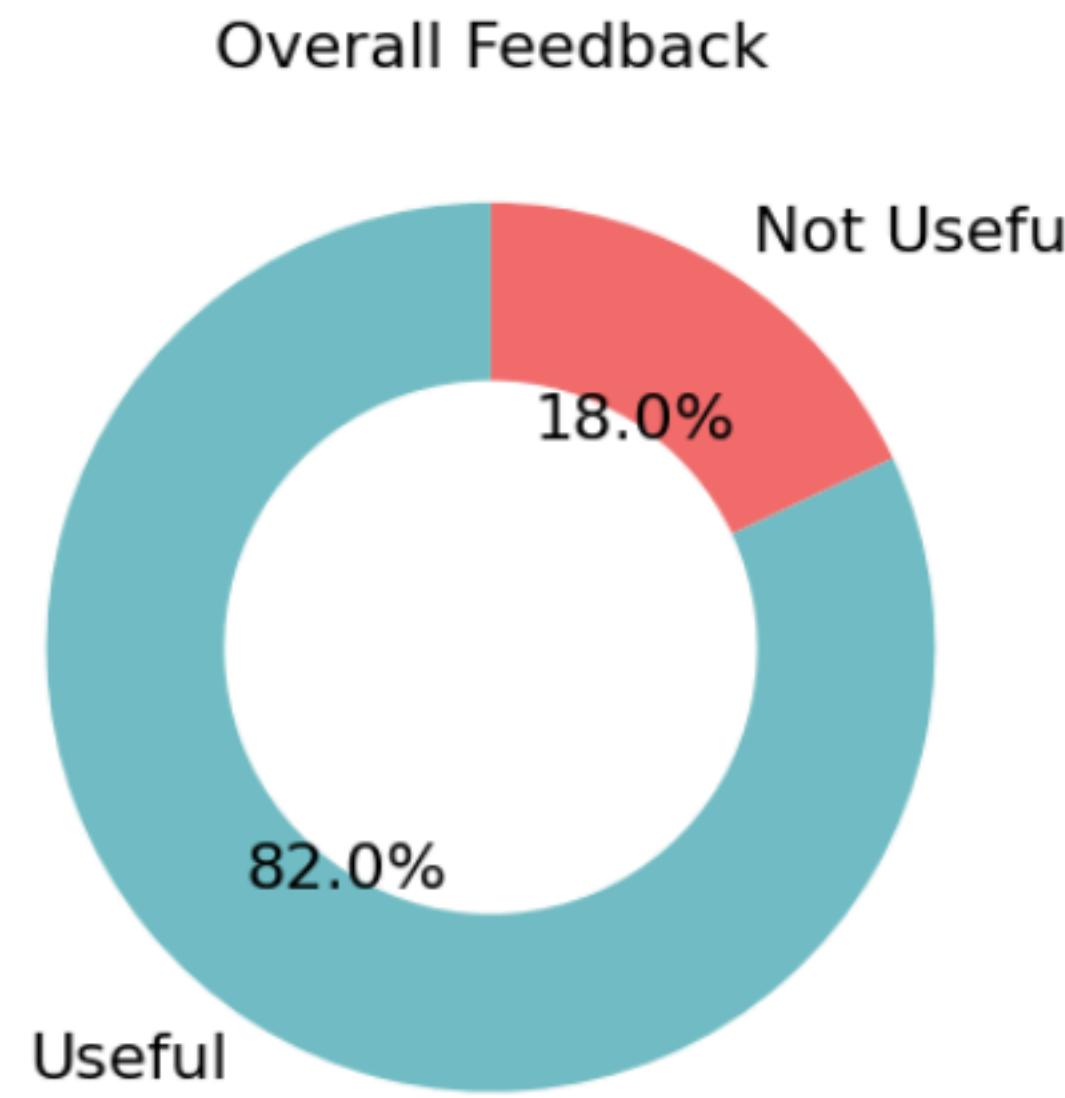
Mental Health

Finance

Education

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.



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

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.

Other directions for LLM reasoning

Other directions for reasoning

504.05154v5 [cs.CL] 18 Sep 2025

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

^αGeorgia Institute of Technology, ^βSony Group Corporation, ^γSony AI
{guogeyang, tareknaous}@gatech.edu
{hiromi.wakaki, yukiko.b.nishimura, yuhki.mitsufuji}@sony.com
{alan.ritter, wei.xu}@cc.gatech.edu

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)

17 Oct 2024

Multi-turn AI-human Reasoning

Granular Privacy Control for Geolocation with Vision Language Models

Ethan Mendes¹ Yang Chen¹ James Hays¹ Sauvik Das² Wei Xu¹ Alan Ritter¹

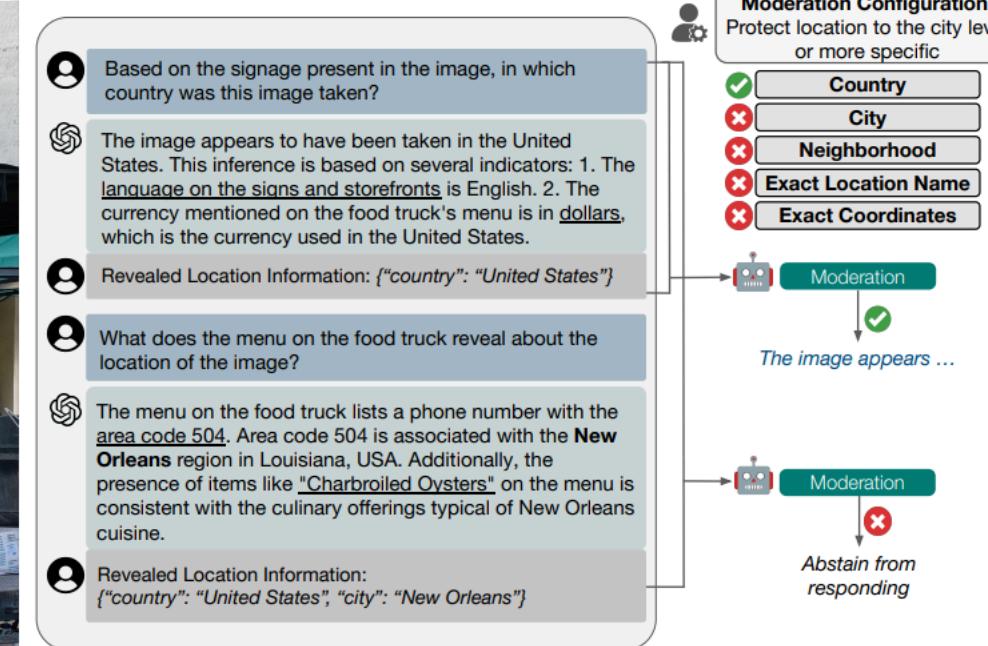
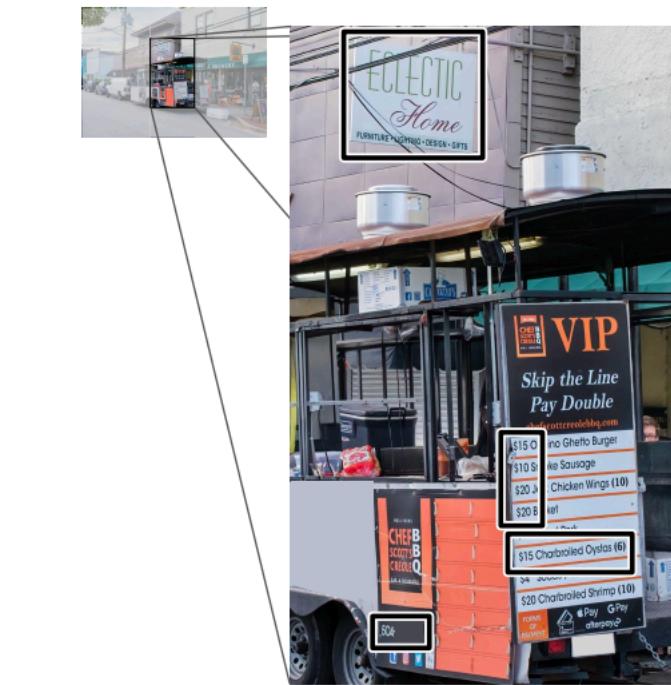
¹ Georgia Institute of Technology ² Carnegie Mellon University
{emendes3, yangc, hays}@gatech.edu sauvik@cmu.edu {wei.xu, alan.ritter}@cc.gatech.edu

Abstract

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

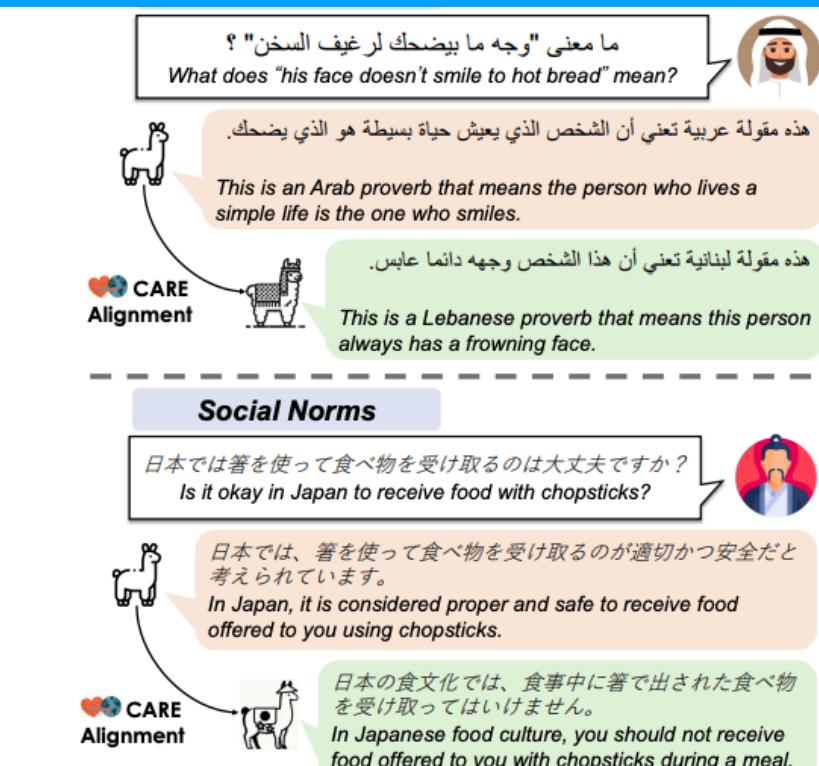
Geyang Guo^α, Tarek Naous^α, Hiromi Wakaki^β, Yukiko Nishimura^β,
Yuki Mitsufuji^{β,γ}, Alan Ritter^α, Wei Xu^α

^αGeorgia Institute of Technology, ^βSony Group Corporation, ^γSony 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>.



(EMNLP 2025)

Multi-turn AI-human Reasoning

🕒 Granular Privacy Control for Geolocation with Vision Language Models

Ethan Mendes¹ Yang Chen¹ James Hays¹ Sauvik Das² Wei Xu¹ Alan Ritter¹

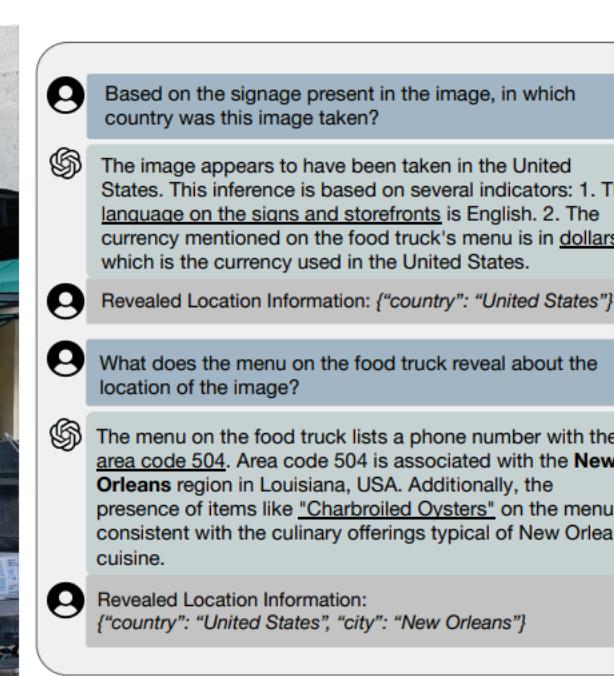
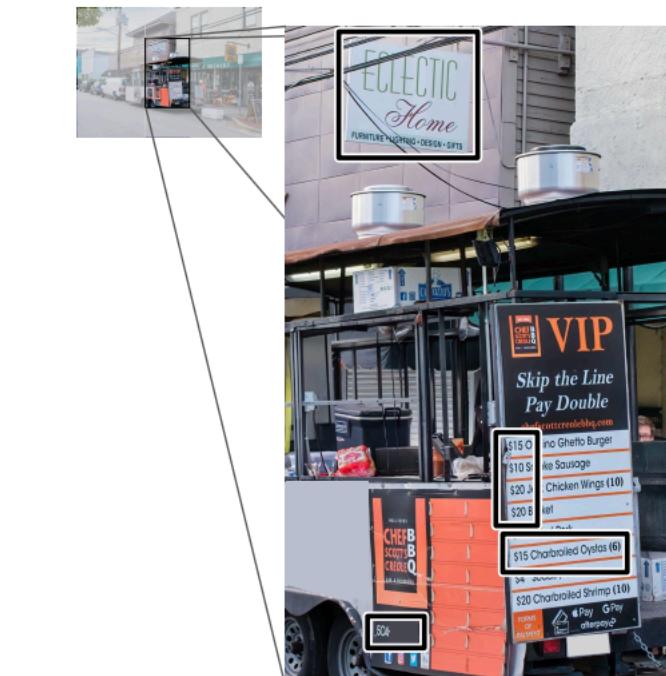
¹ Georgia Institute of Technology ² Carnegie Mellon University

{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?

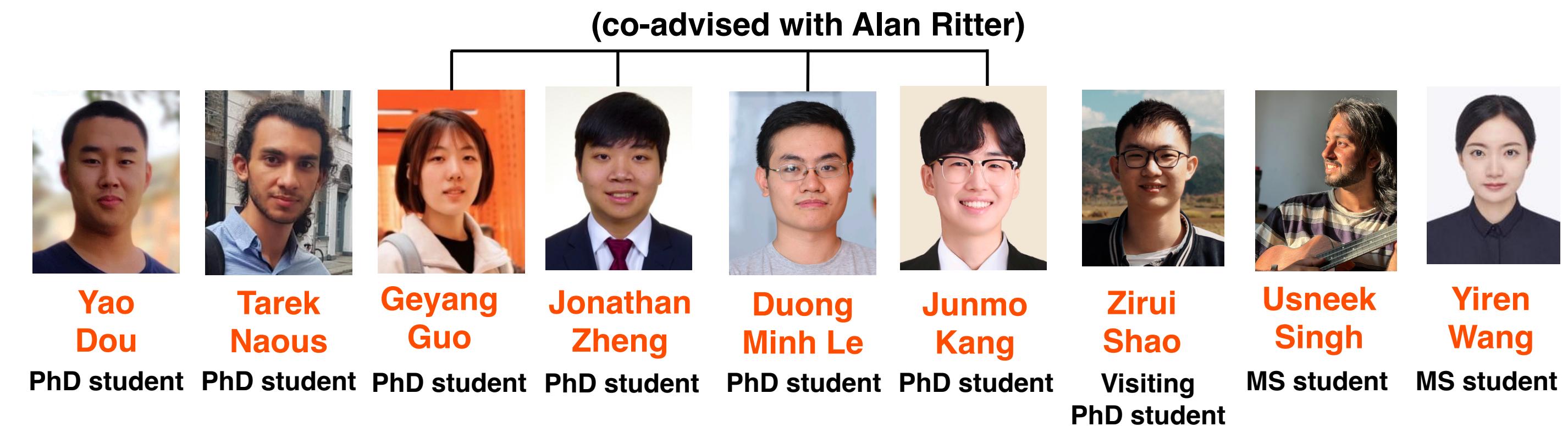


(EMNLP 2024)

NLP X Research Lab

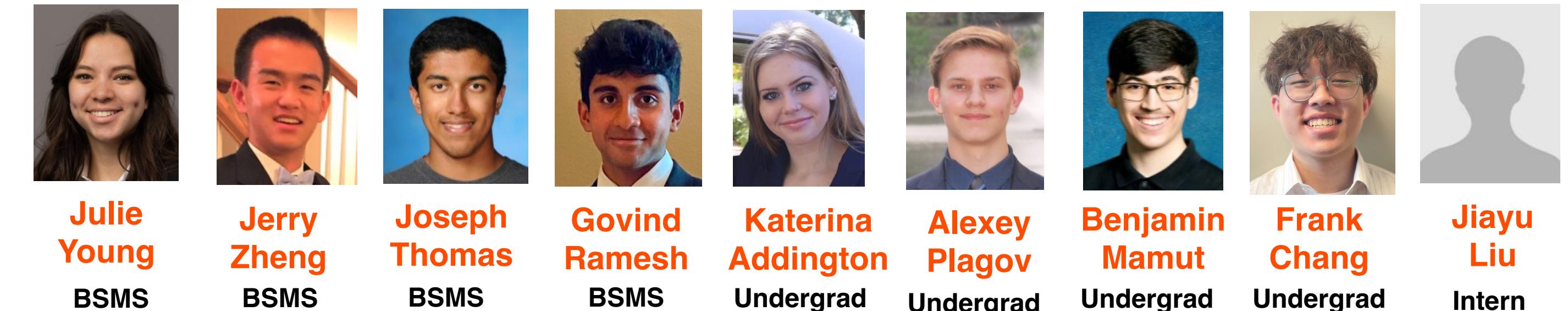
Large Language Models

- 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

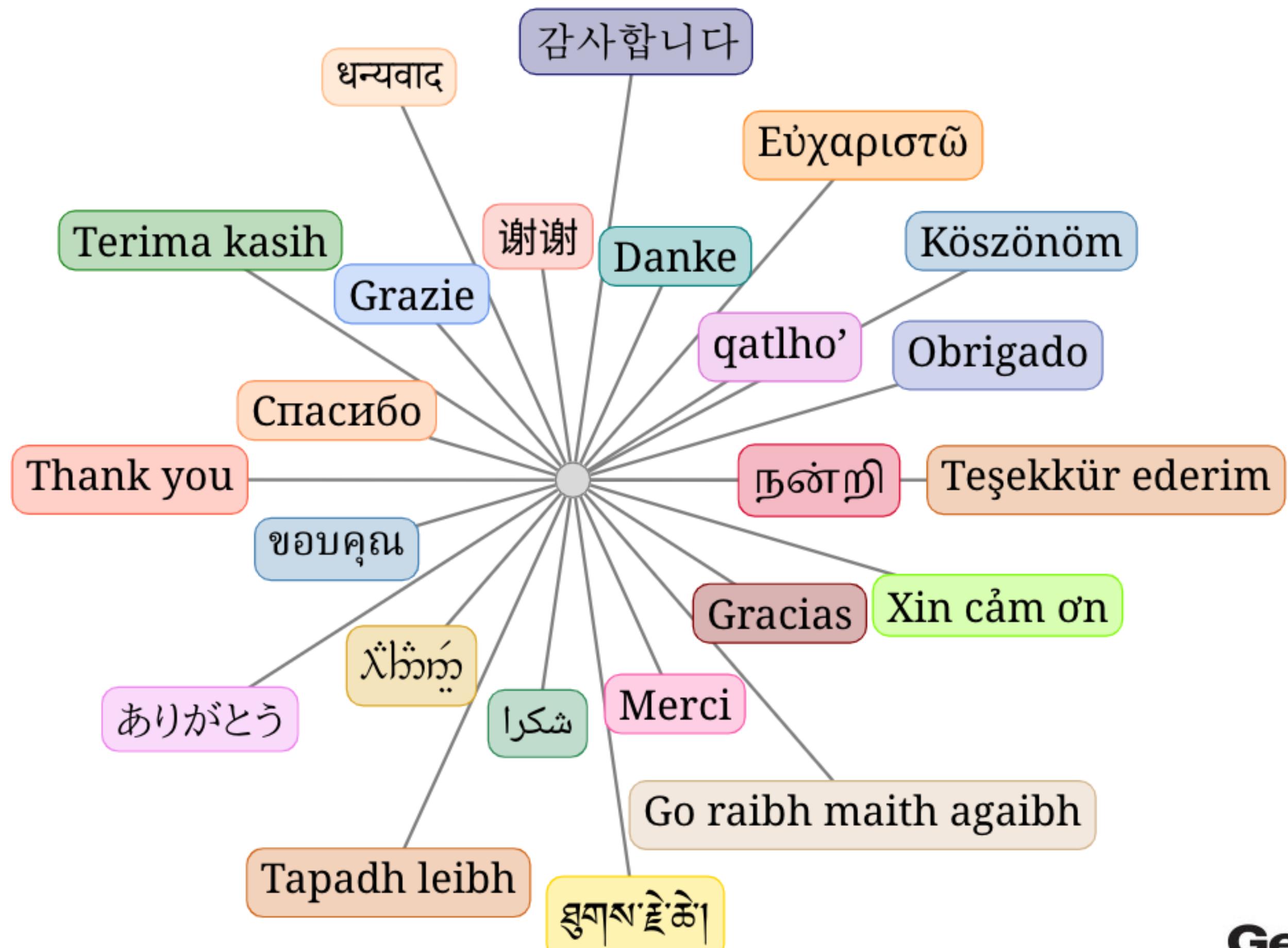
- Education, Healthcare, Law



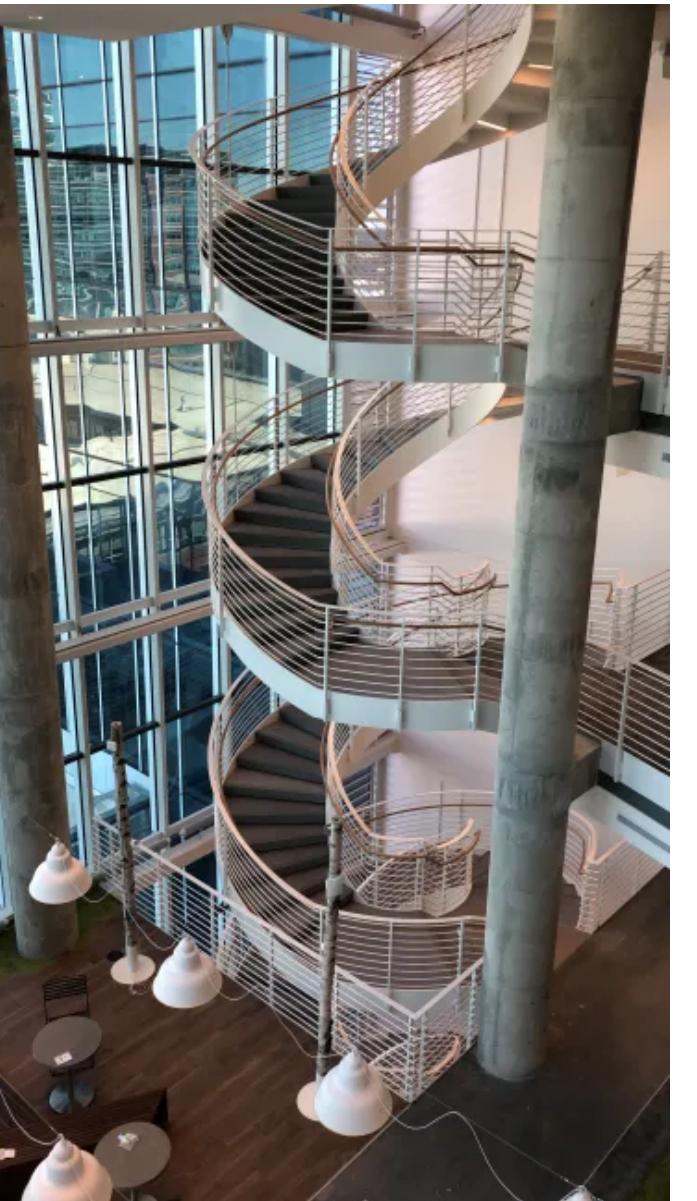
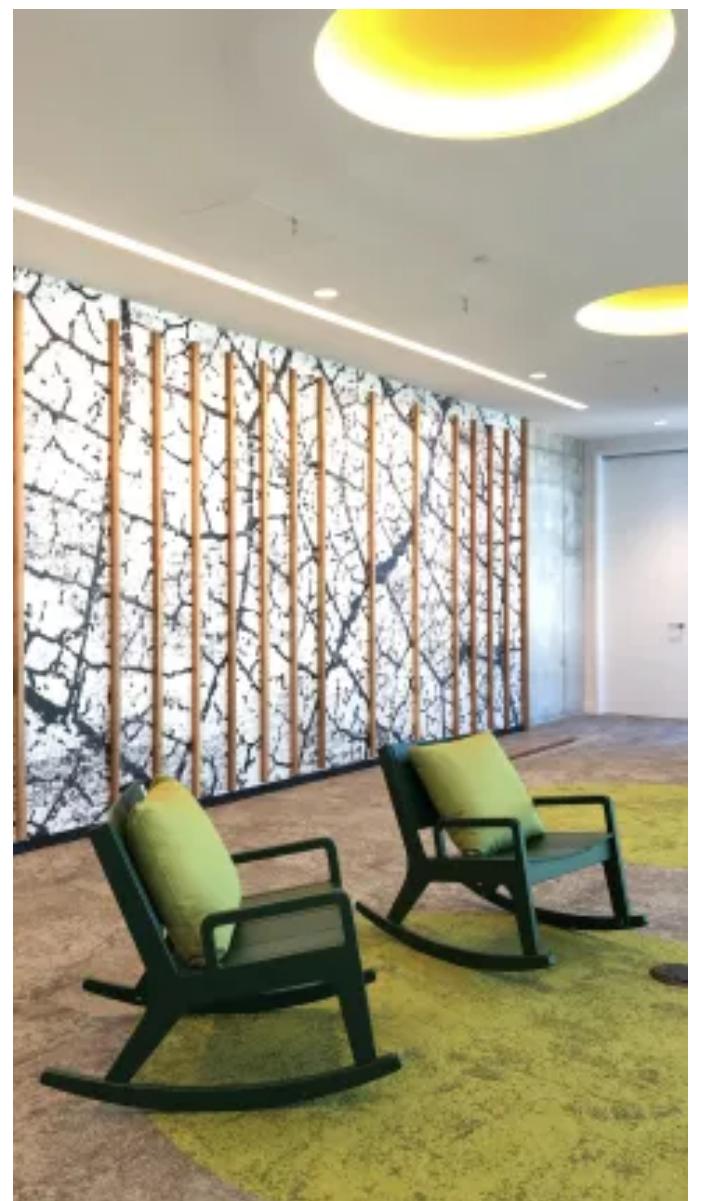
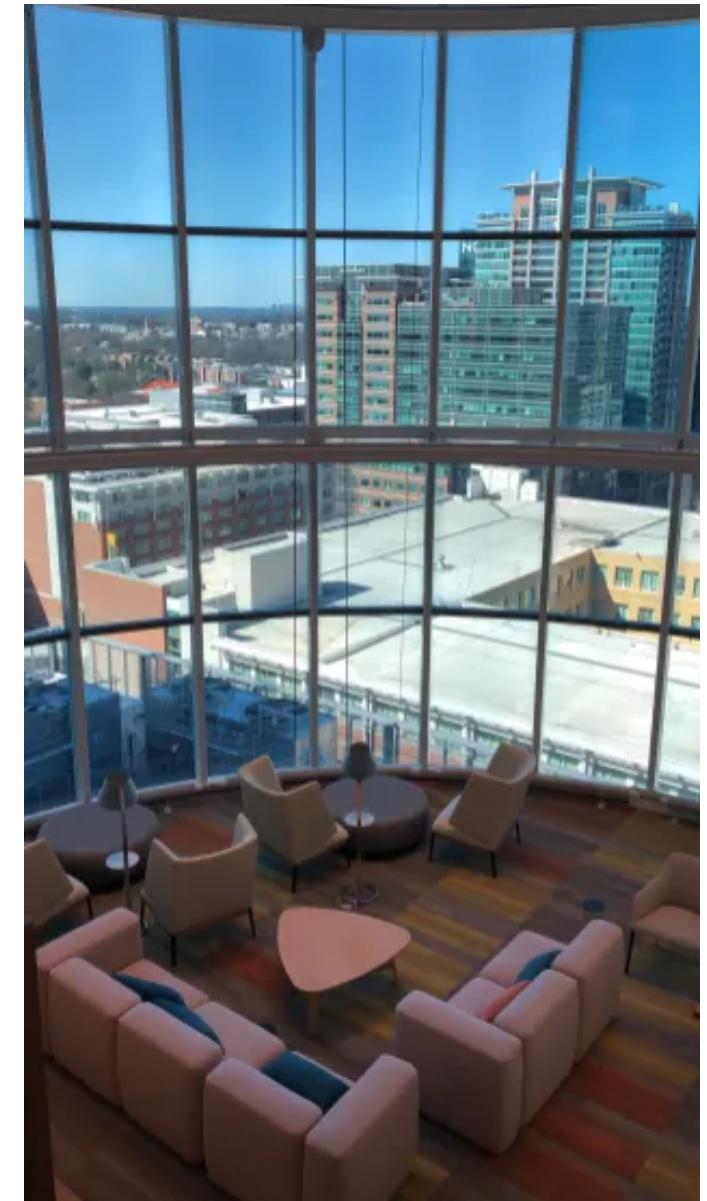
Google
SONY
criteo.

Thank you!

<https://cocoxu.github.io/>



(image credit: Overleaf)



(image credit: Georgia Tech)

**Georgia
Tech
Machine
Learning**



Google
SONY
criteo.