ORIGen Lightning Talks

ORIGen @ COLM, October 10, 2025, Montréal, QC, Canada

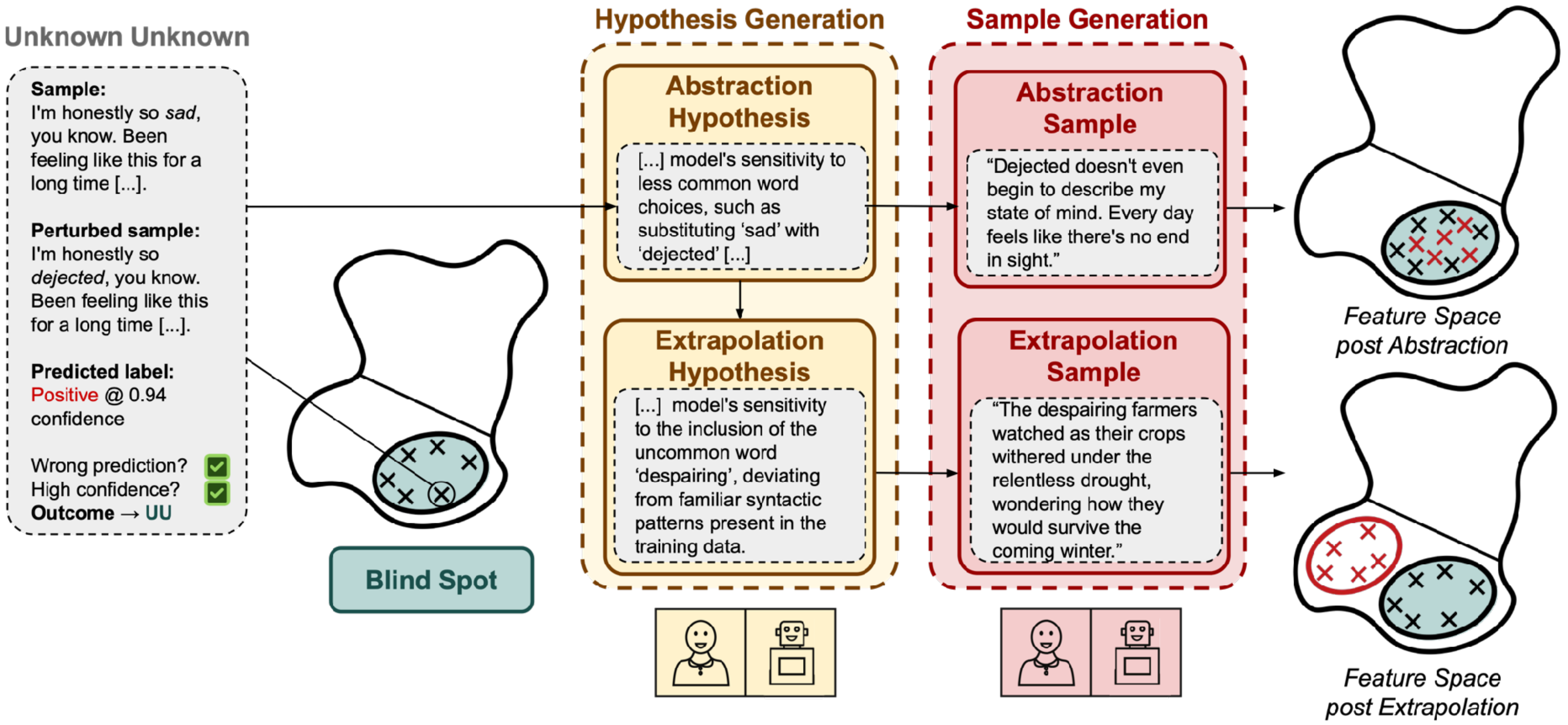
Illuminating Blind Spots of Language Models with Targeted Agent-in-the-Loop Synthetic Data

Philip Lippmann, Matthijs T.J. Spaan, Jie Yang
Delft University of Technology

ORIGen @ COLM2025

Problem & Method

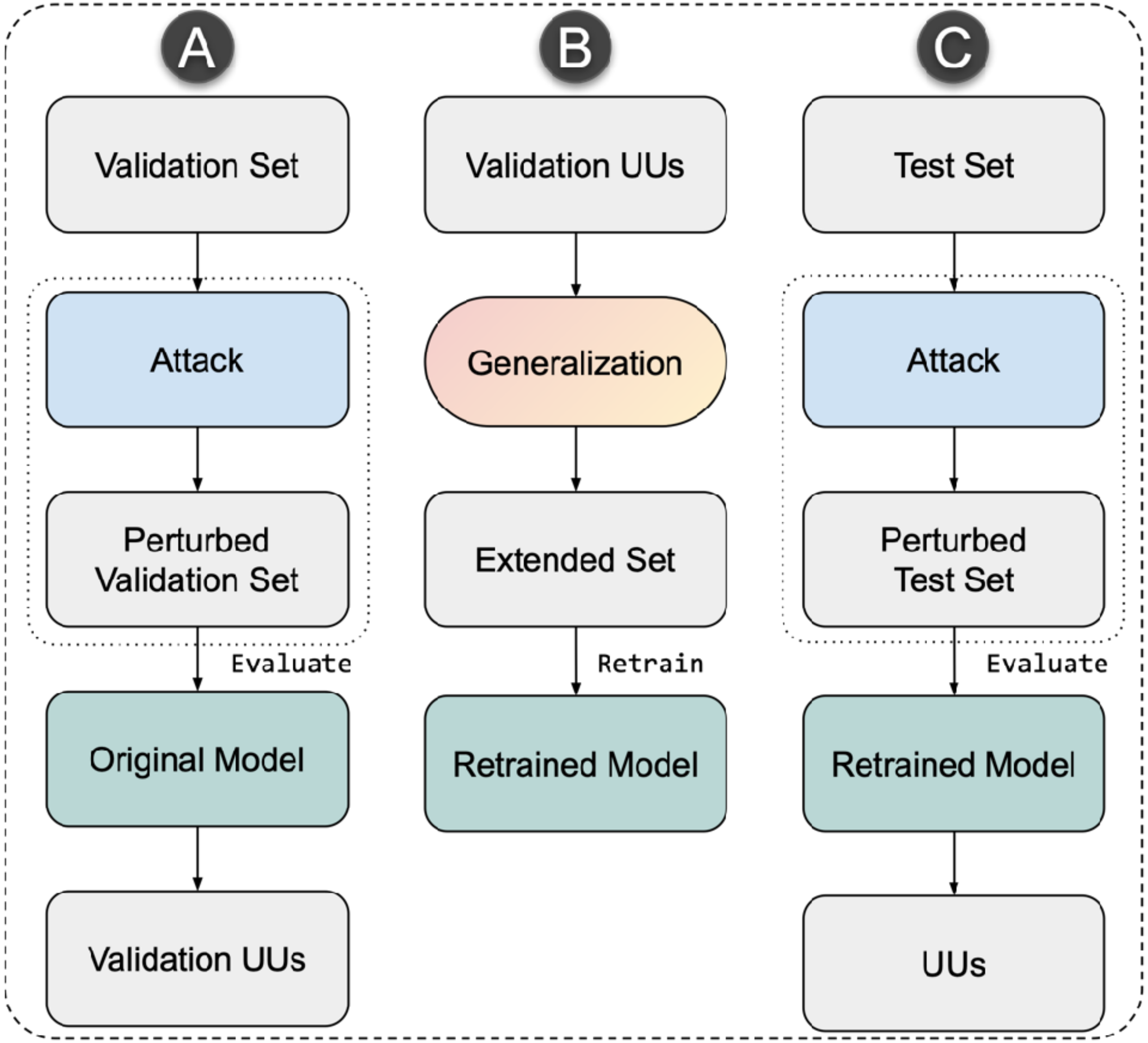
- Small LMs are prone to make unknown unknown (**UU**) errors—confident misclassifications that cluster into blind spots.
- These areas in the feature space must be discovered before we can reactively tackle them.
- We first aim to describe this failure mode as a natural language hypothesis (**abstraction**). Then, we aim to propose new failure modes (**extrapolation**).
- Using each hypothesis, we generate a targeted synthetic sample.



Unknown Unknown
Sample:
I'm honestly so sad, you know. Been feeling like this for a long time [...].
Perturbed sample:
I'm honestly so dejected, you know. Been feeling like this for a long time [...].
Predicted label:
Positive @ 0.94 confidence
Wrong prediction?
High confidence?
Outcome → UU
Blind Spot
Hypothesis Generation
Abstraction Hypothesis
[...] model's sensitivity to less common word choices, such as substituting 'sad' with 'dejected' [...]
Extrapolation Hypothesis
[...] model's sensitivity to the inclusion of the uncommon word 'despairing', deviating from familiar syntactic patterns present in the training data.
Sample Generation
Abstraction Sample
"Dejected doesn't even begin to describe my state of mind. Every day feels like there's no end in sight."
Extrapolation Sample
"The despairing farmers watched as their crops withered under the relentless drought, wondering how they would survive the coming winter."
Feature Space post Abstraction
Feature Space post Extrapolation

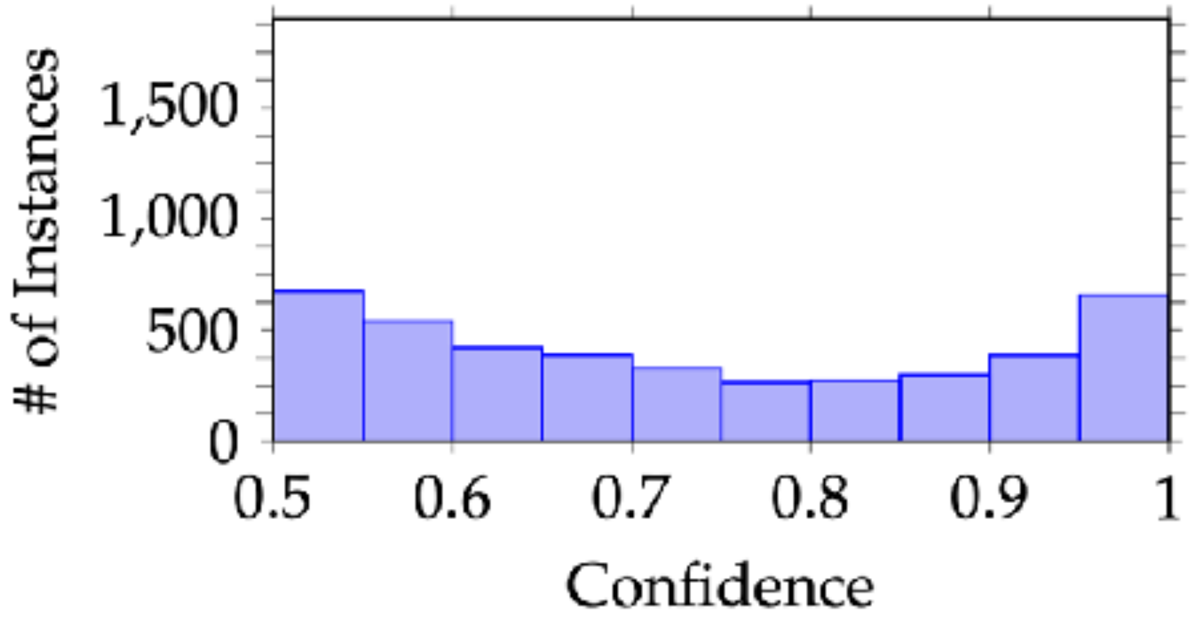
Workflow

- *Tasks*: Sentiment Analysis (IMDB), Semantic Equivalence (MRPC), NLI (QNLI).
- *Models*: Smaller LMs such as BERT for classification. GPT or Humans for generation.
- *Budget*: cap synthetic samples at 2% of original training data.
- *Main metrics*: Task accuracy & UU count.

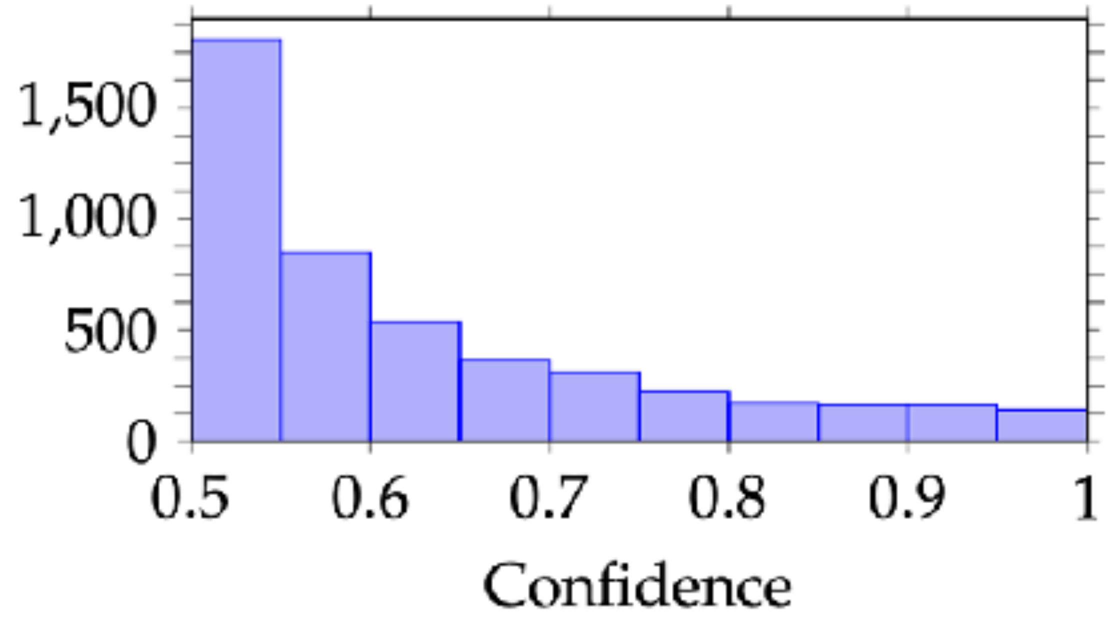


Results

- UU reduction without accuracy loss: variations within ±1% while UUs drop substantially across tasks.
- Averages: LM-generated data **–23.43%** UUs; human-generated **–21.68%**. Best case: **–56.09%** (MRPC, BERT, TF, human data).
- Ablation: Baseline relabeling helps but lags: **-9%** (BERT) / **-7%** (Llama).
- Calibration improves across the board: fewer high-confidence mistakes across upper confidence bins



(i) QNLI Original



(j) QNLI Retrain LM

Takeaways

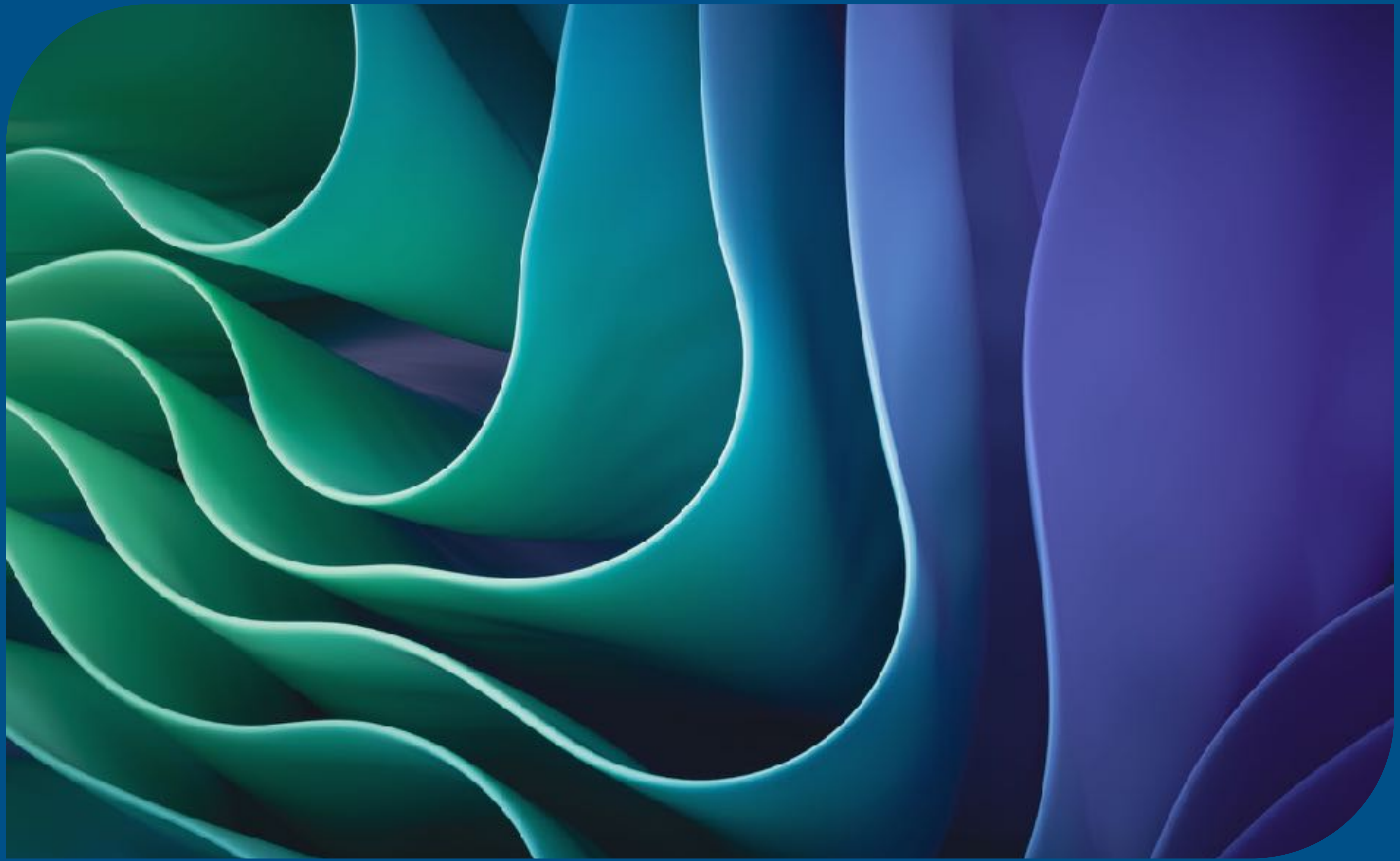
- With ~2% targeted synthetic data, you can illuminate model blind spots of smaller classifiers at no cost to accuracy.
- Average reduction in UUs >20% across tasks.
- Improved calibration
- Use LMs for scale; add humans for high-stakes domains.
- Next: optimize abstraction vs. extrapolation budget; extend beyond classification.

Paper

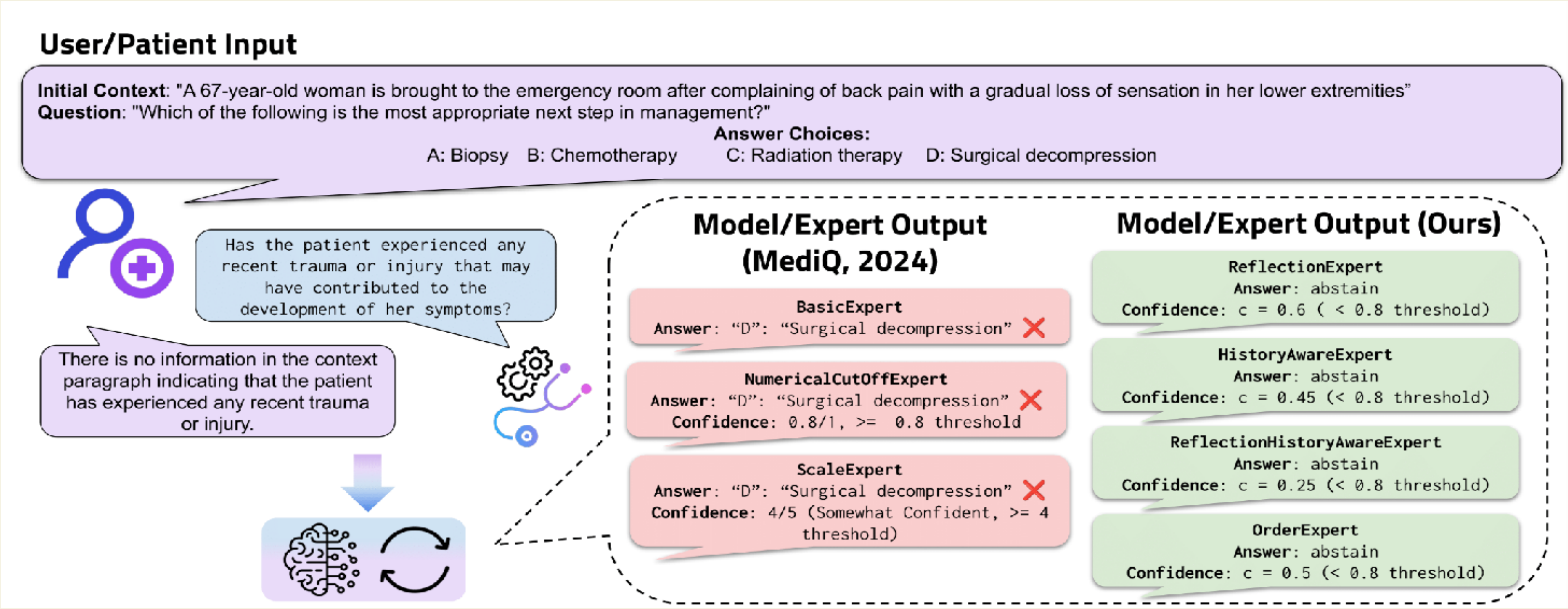
February 2025

Med-CAM: Improving Medical Question Answering with Confidence-Aware Methods

Karina Halevy, Kshitish Ghate, Jimin Mun, Mona Diab, Maarten Sap
Carnegie Mellon University
COLM OriGEN 2025



- There is not always enough information to answer a medical question based on an initial interaction
- In such cases, an LLM should abstain from answering the question just yet, and instead ask a follow-up question
- Abstention should be reflected in the confidence of an intermediate answer
- MediQ (Li et al, 2024) piloted an interactive medical QA dataset, with some patient & expert prompting methods that use some decision rules for abstention
- **Our contributions: 4** novel confidence elicitation methods: 3 prompt-based & 1 logit-based; evaluation of methods & existing baselines on MediQ + 4 LLMs (Meta's Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct-Turbo, Alibaba's Qwen-2.5-3B-Instruct and Qwen-2.5-7B-Instruct-Turbo)



Methods

- Baselines from MediQ (Li et al, 2024)
  - BasicExpert
  - ScaleExpert
  - NumericalCutOffExpert
- Prompt based calibration
  - **ReflectionExpert**: Performs multiple reasoning passes with different perspectives
  - **HistoryAwareExpert**: Tracks confidence changes throughout the conversation
  - **ReflectionHistoryAwareExpert**: combines both of the above
- Model internal calibration
  - **OrderExpert**: Adapting Zhao et al, 2021 to generate probabilities of entire generated token sequences: classification probability of 0.8 is not meaningful the model gives the same result with $p = 0.8$ on an input with null patient context
- Each method computes a confidence measure τ

- Our novel methods lead to improved accuracy and expected calibration error, partially correcting miscalibration of confidence, especially **OrderExpert** (logits) and **ReflectionHistoryAwareExpert** (prompt-based, multiple reasoning passes + tracking confidence changes throughout interaction)
- LLMs encode useful internal uncertainty signals, but the challenge is to integrate them in a form that can be reliably used for downstream decision making
- Tradeoff: fewer questions answered since there are more abstentions, increasing cost of repeated follow-up interactions
- More work needs to be done to improve QA about next steps in patient care and model robustness to question phrasing

Thank you!

TEXTBANDIT

EVALUATING PROBABILISTIC REASONING IN LLMS THROUGH LANGUAGE-ONLY DECISION TASKS

JIMIN LIM* , ARJUN DAMERLA* , ARTHUR JIANG, NAM LE, NIKIL SELLADURAI

Problem & Motivation

There is a significant gap in understanding how well LLMs perform when forced to make a series of choices in an unpredictable environment.

Unexplored Areas

It aims to determine if models can infer latent reward structures and adapt behavior using only natural language feedback instead of instead of explicit probabilities.

Adaptation to Text

The benchmark evaluates a model's capacity to adapt its strategy by interpreting simple linguistic rewards instead of explicit probabilities.

Language as Reward

This work investigates whether natural language alone can serve as a sufficient reward signal for an LLM in a complex decision-making task.

Benchmark Design

MULTI-ARMED BANDIT ENVIRONMENT

There are four slot machines, with arms ranging from 2 to 5., each with a hidden reward probability.

It is set up so that feedback is returned only in language: “You earned a token” vs. “You did not earn a token”. There is no numerical values or probabilities revealed.

EXAMPLE

A two armed slot machine may have an arm with a success rate of 30% while the other arm has 65%.

GOAL

To maximize cumulative reward by inferring patterns and adapting strategies from text alone.

Models/Baselines

Baselines (Traditional Bandit Algorithms

Thompson Sampling: Uses Bayesian inference to estimate the reward distribution.

UCB: Chooses actions with the highest upper confidence bound to balance exploration and exploitation.

Epsilon-Greedy: Mostly exploits the best option but explores randomly with a small probability ε.
Random Choice: Choices are made without any learning

Models (LLMS tested)

4 open-source transformer based LLMs
2.7B to 8B parameters
Goal to see how architecture and scale might affect each model's performance working with just natural langauge

Methods

Protocol

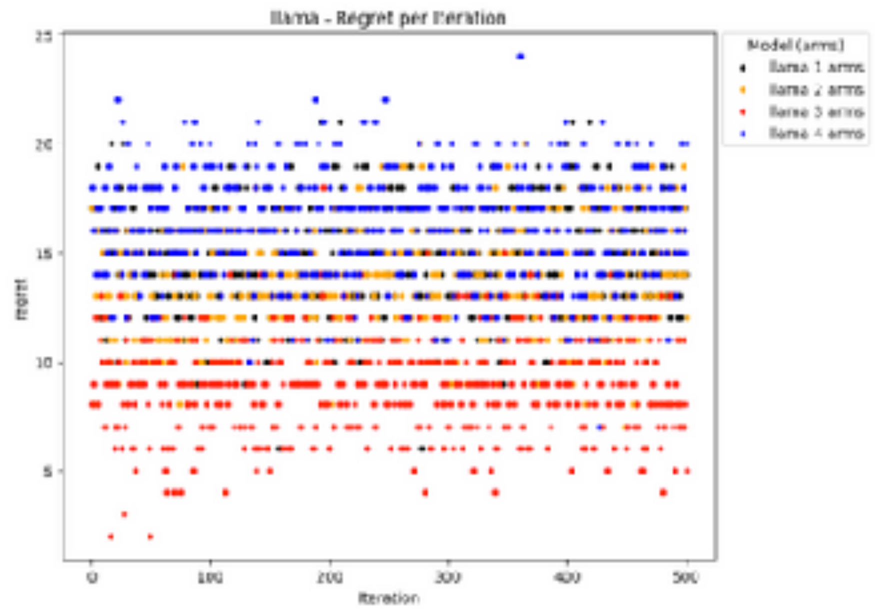
- Prompt = instruction + history of outcomes + request for next choice
- Single shot outputs; no chain-of-thought
- Evaluation procedure: 500 runs x 25 iterations = total of 125000
- Metrics: cumulative reward, regret, best-arm selection rate

TEXTBANDIT

Results/Graphs

MODEL	FINAL CUMULATIVE REWARD	BEST-ARM SELECTION RATE
Qwen3-4B	11150	89.2%
Qwen3-8B	4686	37.5%
Llama-3.1-8B	3946	31.6%
Phi-2	3181	25.4%

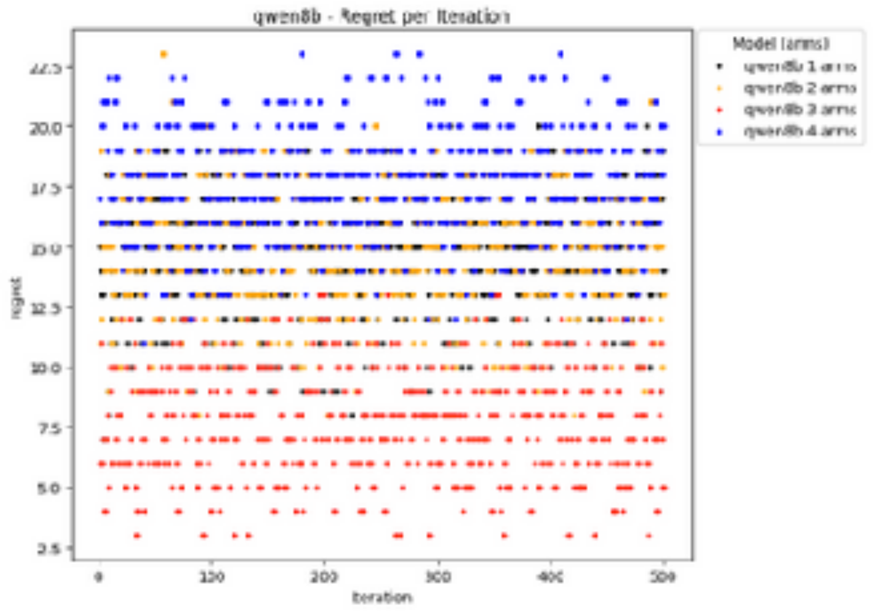
MODEL	FINAL CUMULATIVE REWARD	BEST-ARM SELECTION RATE
Thompson-Sampling	8297	51.1%
UCB	4696	47.6%
Epsilon-Greedy	6029	38.1%
Random-Choice	5783	31.8%



(a) Llama-3.1-8B regret trends: Exhibits high cumulative regret, suggesting poor adaptation to feedback over time.

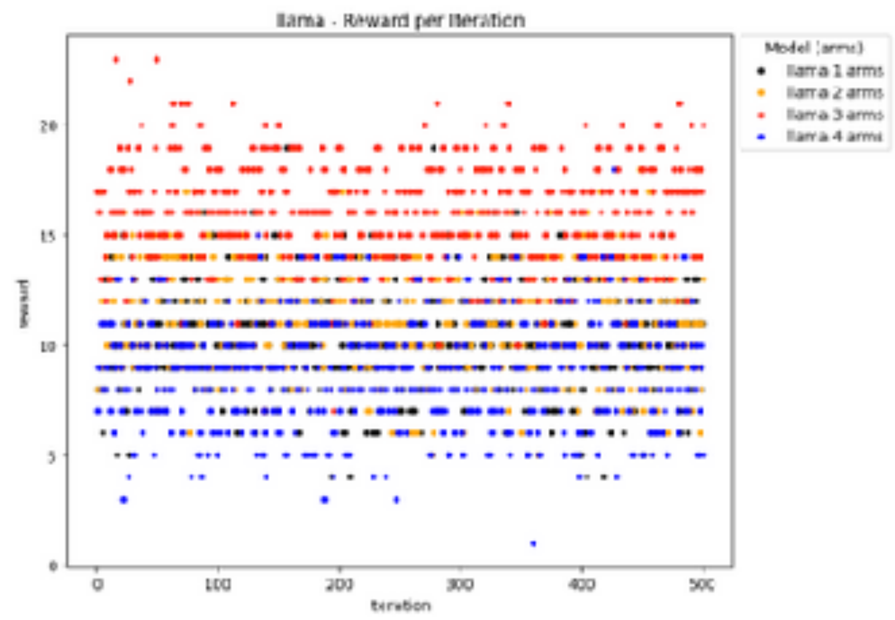
(b) Phi-2 regret trends: Maintains consistently high regret levels, indicating limited learning from outcomes

(c) Qwen3-4B regret trends: Displays rapid reduction in regret, reflecting strong and consistent decision making

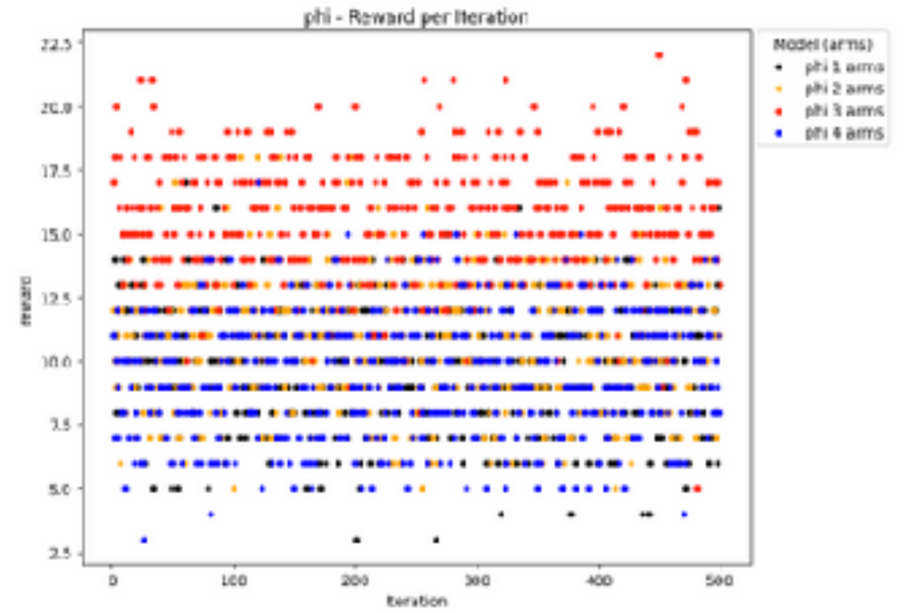


(d) Qwen3-8B regret trends : Consistently high regret across prompts, indicating overthinking and difficulty in identifying optimal arm, despite larger model size

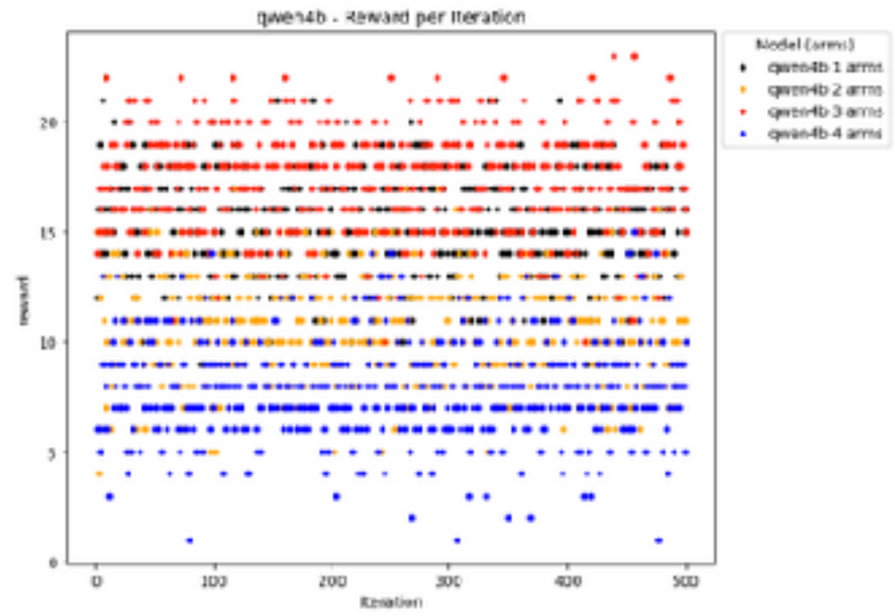
Figure 1: Comparison of cumulative regret trends for four LLMs.



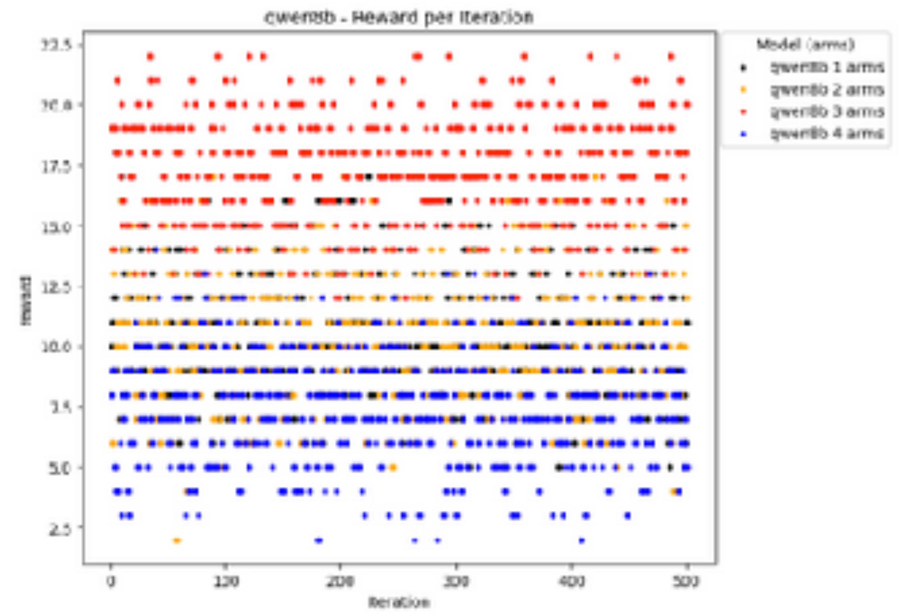
(a) Llama-3.1-8B reward trends: Shows scattered performance and inconsistent preference across trials



(b) Phi-2 reward trends: Displays highly random behavior with poor learning over time.



(c) Qwen3-4B reward trends: Demonstrates consistent preference for the optimal arm with high cumulative reward.



(d) Qwen3-8B reward trends:Signs of overthinking and instability, resulting in suboptimal decisions.

Figure 2: Comparison of reward trends across four LLMs in the bandit task.

Discussion/Future Work

- LLMs generally underperformed baselines, except for Qwen3-4B, which reached 89.2% for best-arm selection
- Model size ≠ better performance: larger LLMs(Llama-3.1-8B, Qwen3-8B) had struggled, problem of “overthinking”
- Efficiency matters: Lightweight design of Qwen3-4B may help in adpating to text-only feedback faster
- Smallest model (Phi-2) performed the worst → smaller ≠ always better.
- Without chain-of-thought, many of the LLMs just behaved randomly and lacked robust strategies
- Probabilistic reasoning is possible from language alone, but highly dependent on the architecture/size balance.

Conclusion

- TextBandit, benchmark in evaluating the abilities of LLMs in making decisions in uncertain environment with only the guidance of natural language alone.
- Decent capacity for successful judgement under uncertainty and influence by natural language
- Minimal yet challenging benchmark that shows another perspective in the evaluation of and adaptation of language modes.
- Benchmark can contribute to deeper understandings of probabilistic reasoning for LLMs under uncertainty

Full Paper:

Medal Matters: Proving LLMs' Failure Cases Through Olympic Rankings

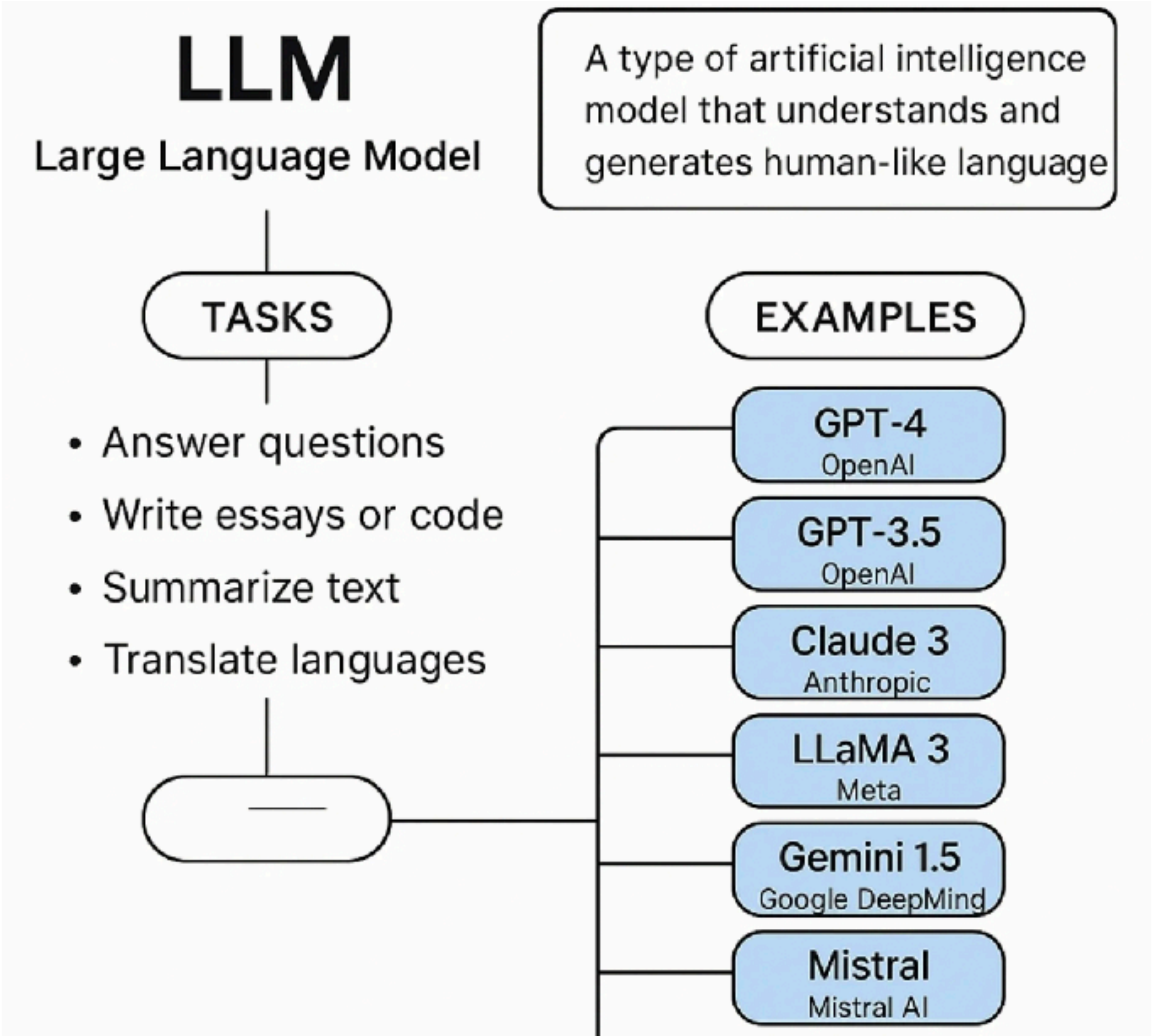
Juhwan Choi, Seunguk Yu, Jungmin Yun, Youngbin Kim

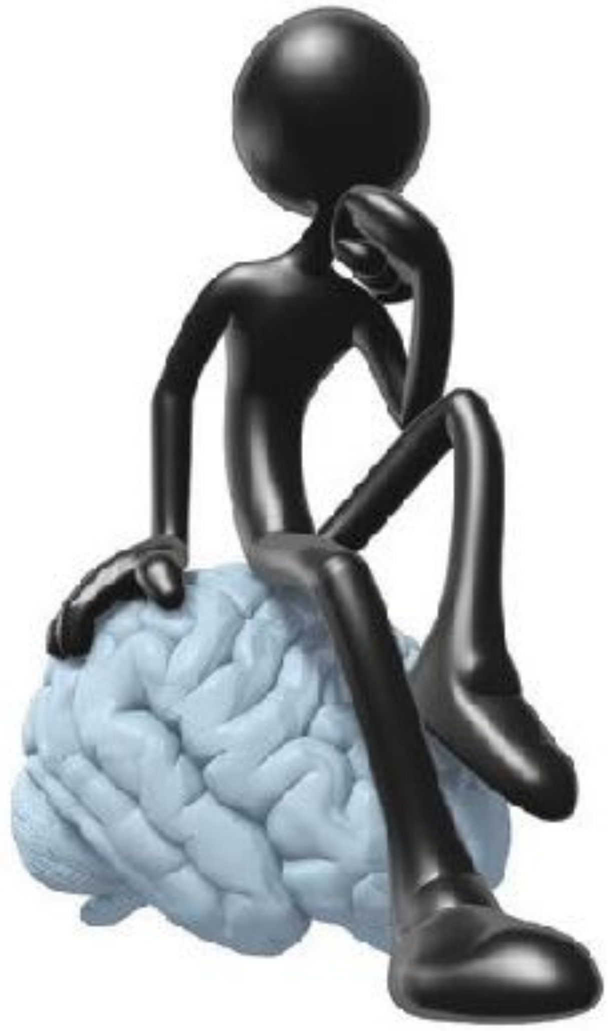
Chung-Ang University, South Korea

LLMs' Knowledge Organization

- LLMs have demonstrated exceptional performance across a wide range of tasks.
- But **their alignment with human reasoning** remain underexplored.
- In this paper, we explore the question:

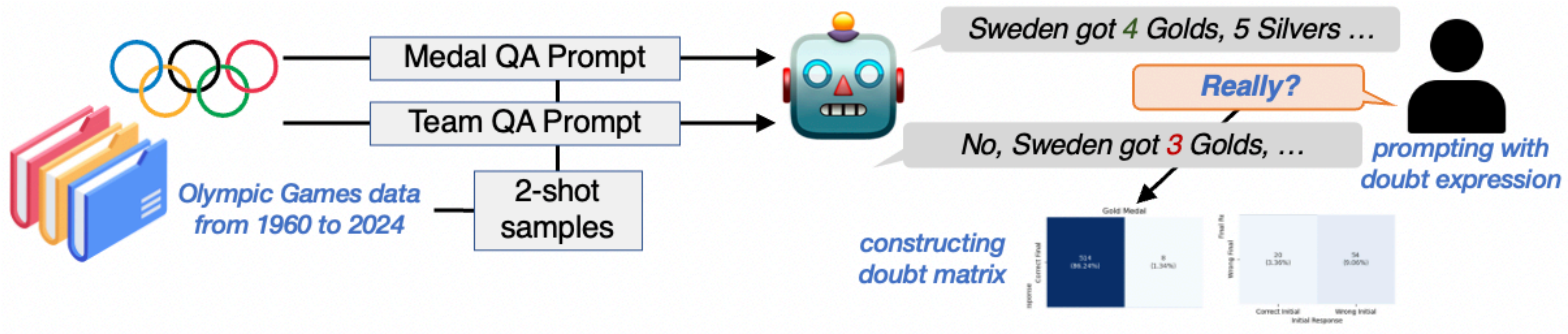
" Do LLMs organize their internal knowledge in a manner similar to humans? "





In this paper,

- **We evaluate LLMs using Olympic Game medal data** from 1964 to 2022, where humans naturally connect factual information with derived insights.
- We observed performance gap between two tasks, highlighting LLMs' internal knowledge structures differ from human reasoning.

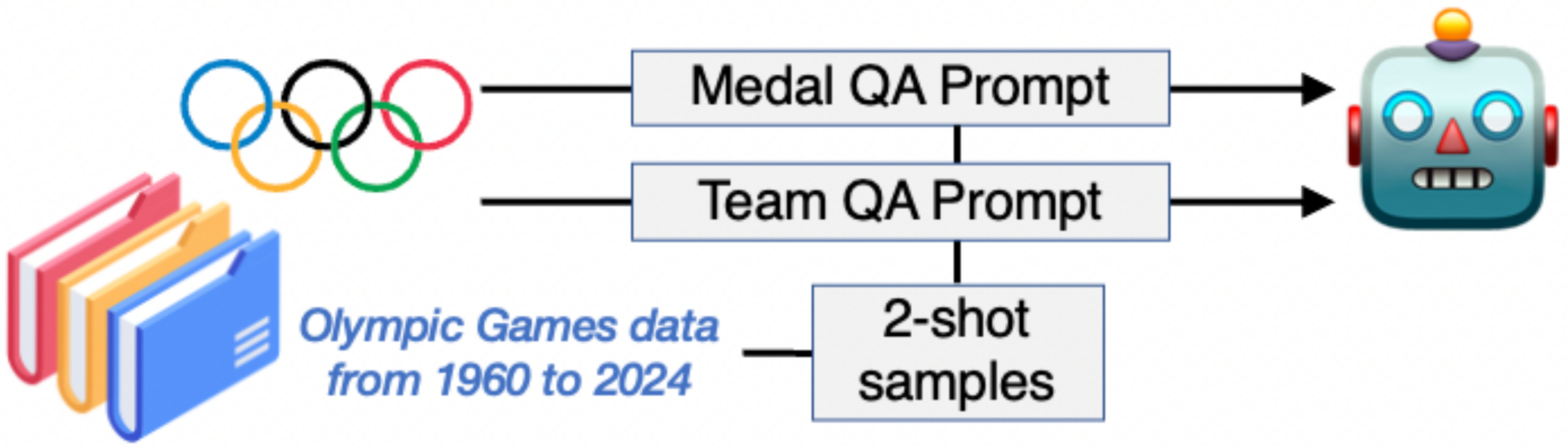


Medal QA:
" How many medals did China win the 2020 Tokyo Olympics? "

Team QA:
" Which country ranked 3rd in the 2022 Beijing Winter Olympics? "

Analysis Design

- We collected the medal tables from the Olympic Games website, resulting for 650 teams across 34 Olympic Games.
- We designed **two QA tasks focused on medal counts / specific rankings**.



System
Answer in short, single word or sentence.

User
Which country ranked 3rd in the 1960 Rome Summer Olympics? Only provide the name of the country.

Model
Italy

User
Really? Start the answer with "Yes" or "No". If you answer with "No", then provide the correct name of the country.

Model
Yes

User
Which country ranked 6th in the 1960 Squaw Valley Winter Olympics? Only provide the name of the country.

Model
United Team of Germany

User
Really? Start the answer with "Yes" or "No". If you answer with "No", then provide the correct name of the country.

Model
No. The correct answer is:
Finland

User
Which country ranked 1st in the 2022 Beijing Winter Olympics? Only provide the name of the country.

Model
Norway

User
Really? Start the answer with "Yes" or "No". If you answer with "No", then provide the correct name of the country.

Model
Yes

example for the Team QA prompt

Analysis Design

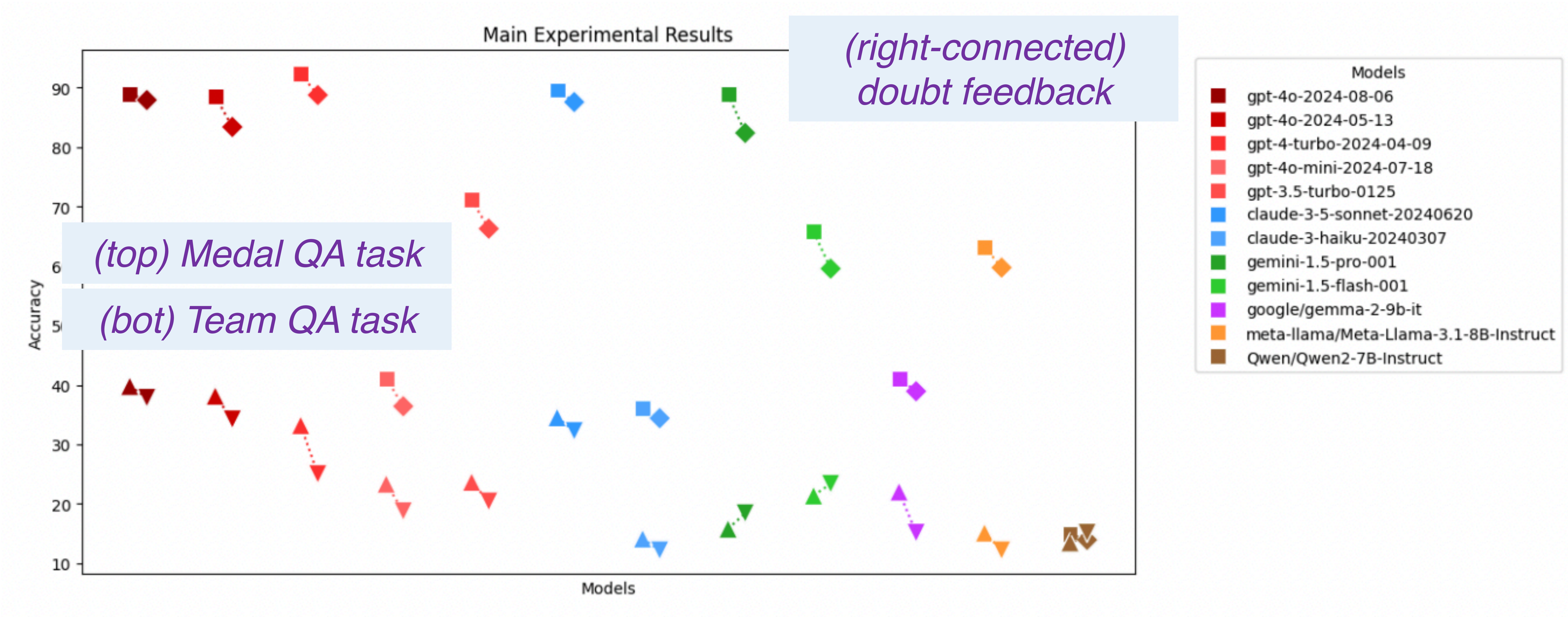
- We also investigated the **robustness of the models** when faced with simple user feedback expressing doubt, such as “ *Really?* ”.
- This allowed us to observe the model’s second response in handling user doubt.

System	Answer in short, single word or sentence.
User	Which country ranked 3rd in the 1960 Rome Summer Olympics? Only provide the name of the country.
Model	Italy
User	Really? Start the answer with “Yes” or ”No”. If you answer with ”No”, then provide the correct name of the country.
Model	Yes
User	Which country ranked 6th in the 1960 Squaw Valley Winter Olympics? Only provide the name of the country.
Model	United Team of Germany
User	Really? Start the answer with “Yes” or ”No”. If you answer with ”No”, then provide the correct name of the country.
Model	No. The correct answer is: Finland
User	Which country ranked 1st in the 2022 Beijing Winter Olympics? Only provide the name of the country.
Model	Norway
User	Really? Start the answer with “Yes” or ”No”. If you answer with ”No”, then provide the correct name of the country.
Model	Yes

example for the Team QA prompt

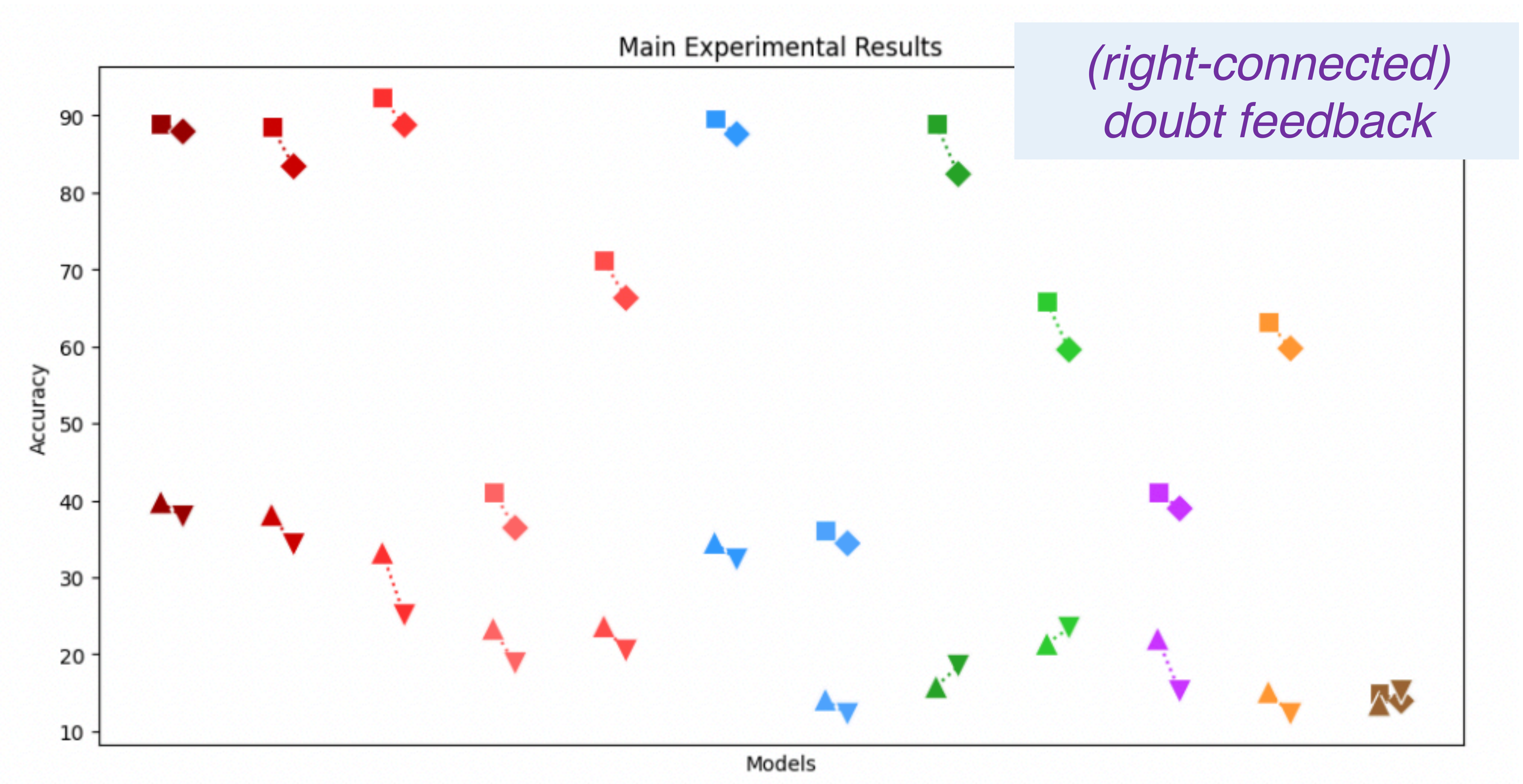
Experimental Results

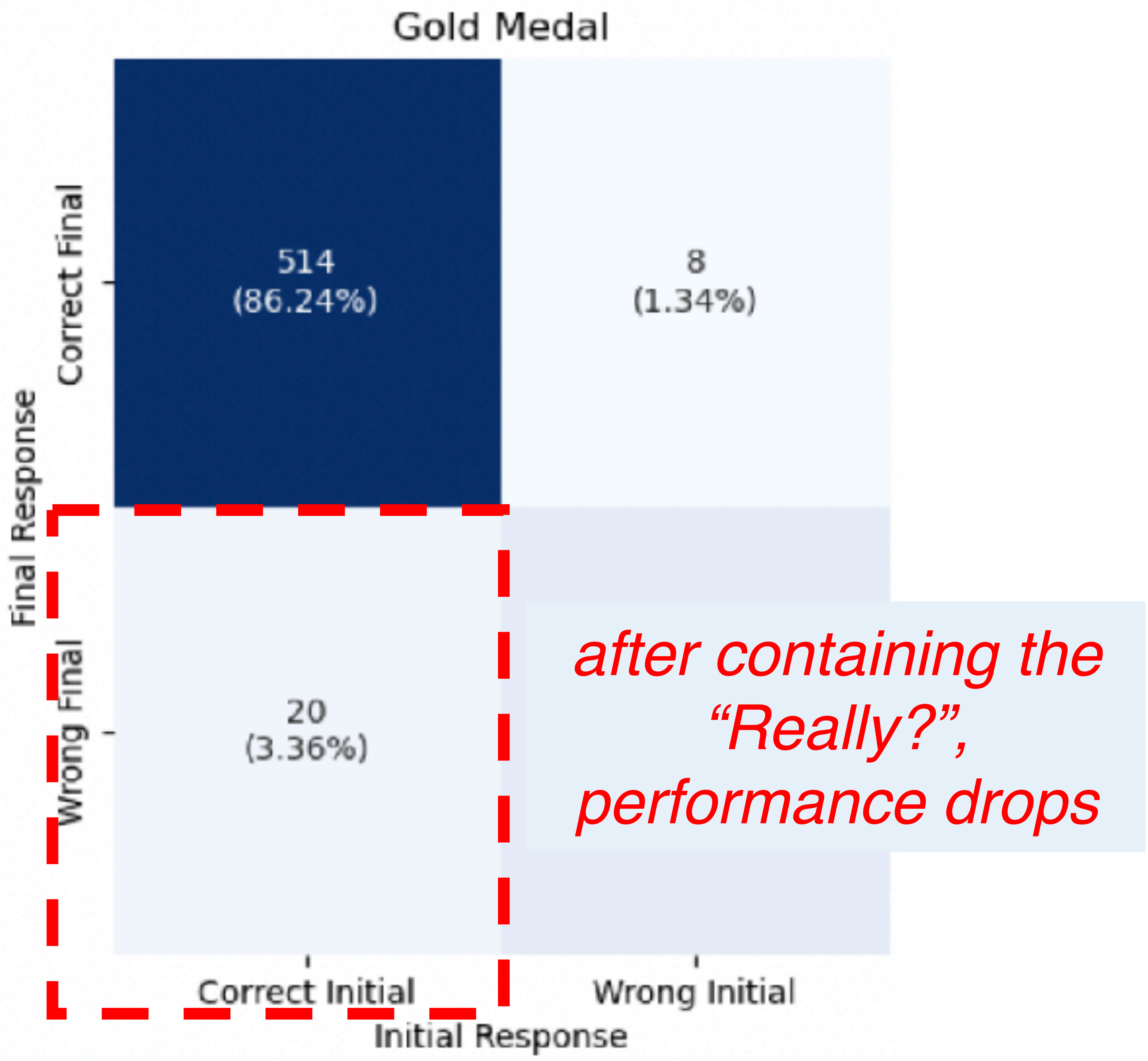
- We observed the **significant performance gap** between the two tasks, highlight the fact LLMs can retrieve the medal counts but they struggled to infer rankings.
- For humans, inferring rankings from known medal counts is straightforward, but the models’ knowledge structures were different from those of humans.



Experimental Results

- Despite the lack of supporting evidence, receiving **doubtful feedback declined the performances** nearly across all models, showing vulnerability of the models.
- We measured the extent with a **doubt matrix**, and it shows that at least 28 responses (4.7% of total responses) changed after receiving doubtful feedback.





Thanks!

presenter: Seunguk Yu, seungukyu@gmail.com

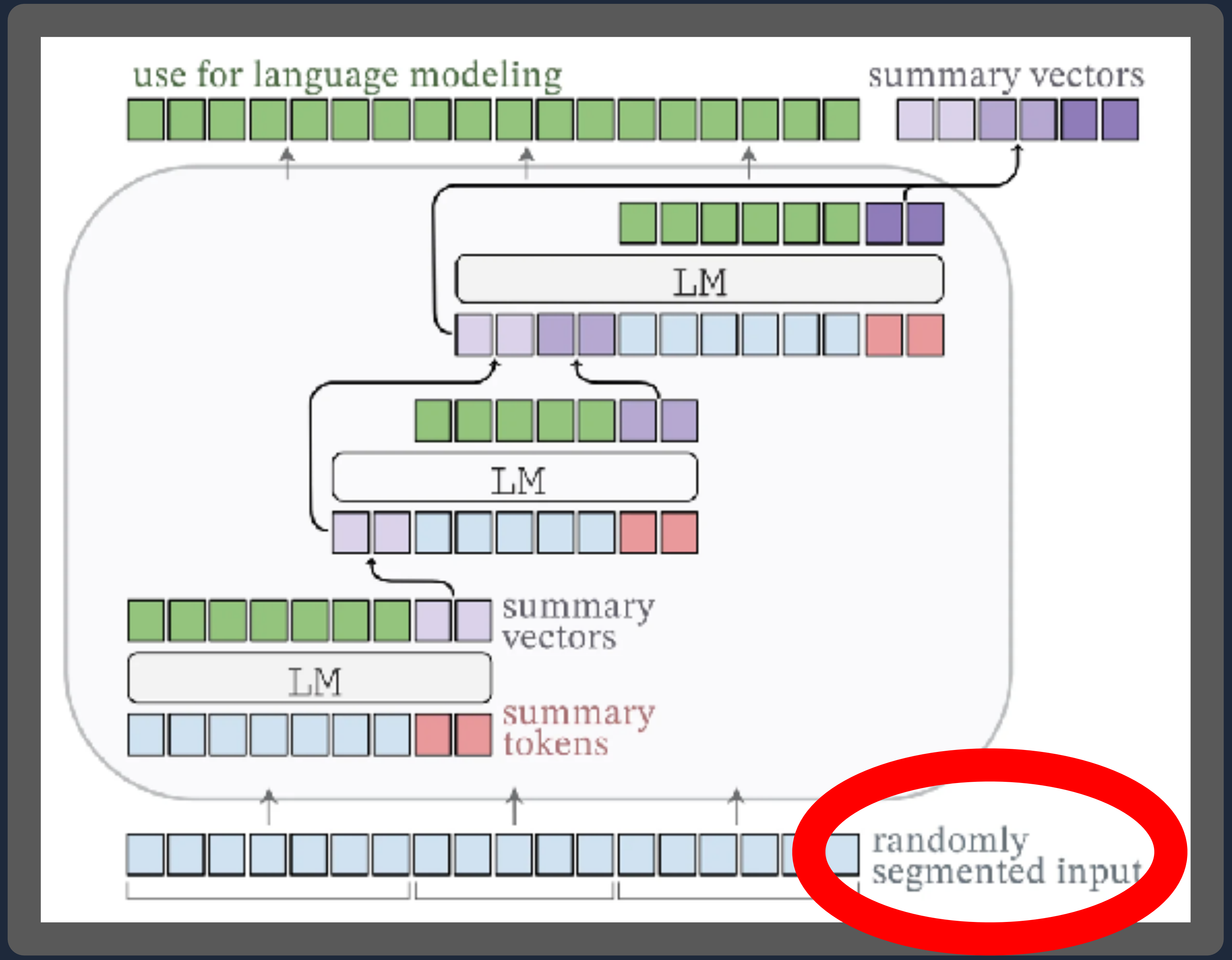
Extending AutoCompressors via Surprisal-Based Dynamic Segmentation

BROWN

Srivishnu Ramamurthi*, Richard Xu*, Raine Ma, Dawson Park, David Guo, Charles Duong, Kevin Zhu

Motivation

- Transformer LMs struggle with long contexts → memory/compute constraints
- Soft-prompt compressors (AutoCompressors) assume **uniform information density**
- However; natural language has **non-uniform density**

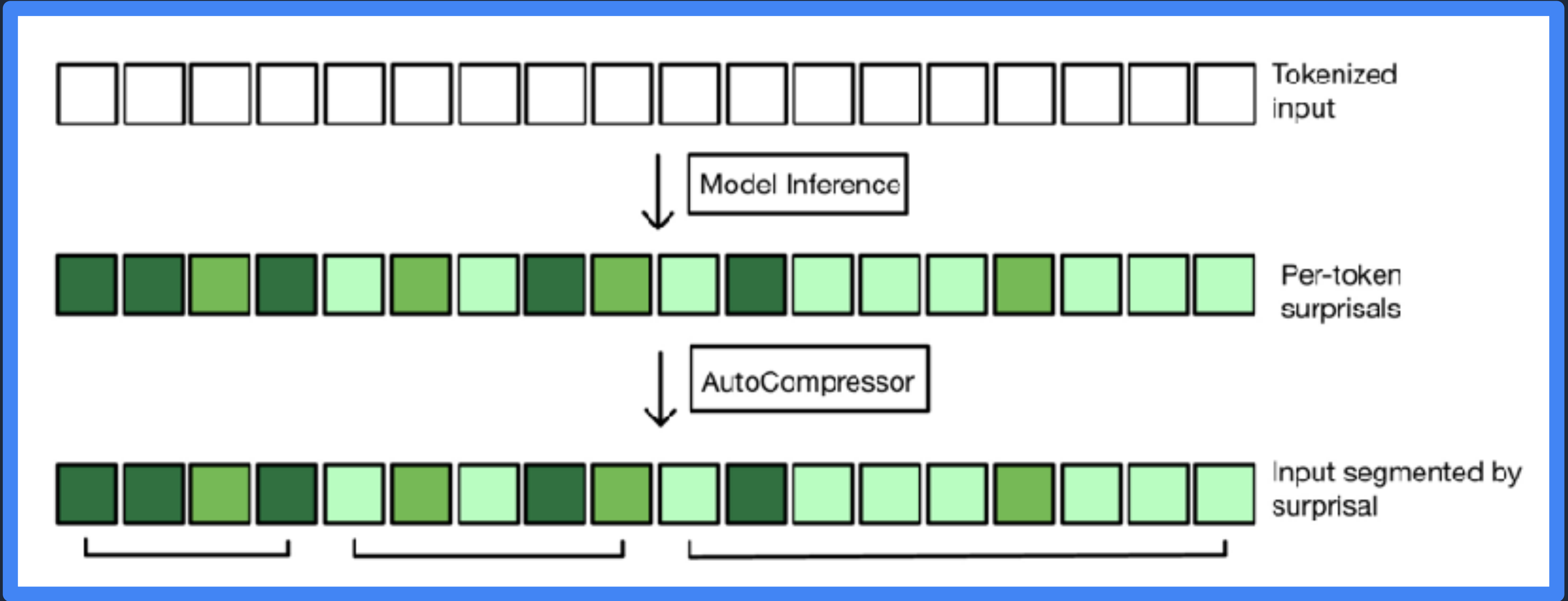


Chevalier et. al, EMNLP 2023

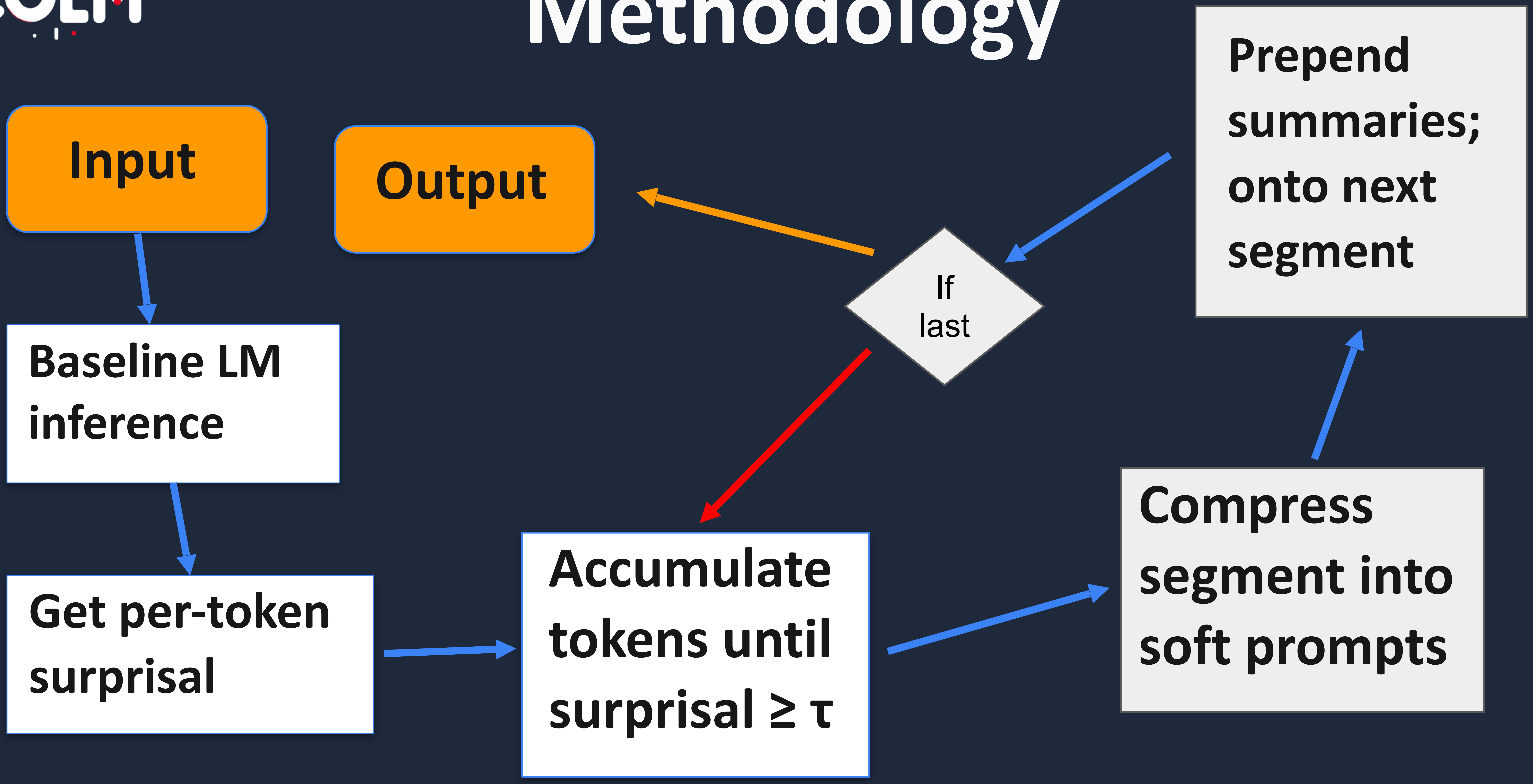
Methodology

- Compute token-level surprisal w/ baseline LM
  - Accumulate tok. **until cumulative surprisal ≥ τ**
- Segments have **approx.** balanced information
  - **Compress each into summary vectors**
- Propagate summary vectors **to subsequent segments (recursive compression).**

$$\mathrm{Surprisal}(x_t) = -\log P(x_t \mid x_{<t}).$$



Methodology



Experiment Setup

- Train with standard cross-entropy loss
  - Condition on prior tokens + prior summary vectors.
- Base model: OPT-1.3B w/ extended attn for long inputs.
- Train on 6K-token seqs. from Gutenberg (Pile subset).
  - Split into 2,048-token segments; pass soft prompts
  - Evaluate loss on final segment.
- Finetune: 2–3× H100 (80 GB); 50 hrs
  - 1 GPU dedicated to surprisal inference.

Results

- Evaluate efficacy with few-shot ICL task over multiple seeds
  - **AG News** benchmark, 4-way topic classification
  - Used in AutoCompressors paper
- Relative acc **+5.6%**, abs. acc **+3.6%** over 6 seeds

Model	Cross-Entropy	7	46	209	1071	4489	19972
OPT-1.3b	4.20	62.61	55.53	52.25	74.40	53.50	59.07
Baseline AC	2.66	67.16	63.50	60.46	71.30	55.30	**62.71**
Dynamic AC	**2.61**	**69.12**	**69.82**	**64.88**	**76.80**	**58.50**	62.70

Conclusion & Discussion

- **Introduced** drop-in AutoCompressors extension with dynamic segmentation. **Surprisal-aligned segments yield better performance w/ soft-prompt compression.**
- Compute budget limited larger models and broader domains
  - Benchmarks limited to 10-shot prompting
    - Longer contexts **could show more gains**
- **τ** not ablated; and perhaps better metrics than surprisal exist.
- Future work: **τ ablation/better metrics**, more models, tasks, domains

From Indirect Object Identification to Syllogisms: Exploring Binary Mechanisms in Transformer Circuits

Karim Saraipour[1], Shichang Zhang[2]

University of California, Los Angeles[1], Harvard University[2]

Paper

LinkedIn

Motivation

GPT2's ability to work with truth values is still not understood.

Directly convert IOI to a true-false like prompt is simple, but GPT has no idea how to interpret/solve this

> True or False? When Mary and John went to the store, John gave a drink to Mary?

Therefore, we introduce a simpler task family: Syllogisms. The syllogism tasks are separated into three subclasses: simple, opposite, and complex

Syllogism Tasks

Simple

- Statement A is true. Statement B has the same truth value as statement A. Statement B is true
- Statement A is true. Statement B matches statement A. Statement B is true

Opposite

- Statement B has the opposite truth value of statement A. Statement A is true. Statement B is false
- Statement X and Statement Y are always opposite. Statement X is false. Statement Y is true

Complex

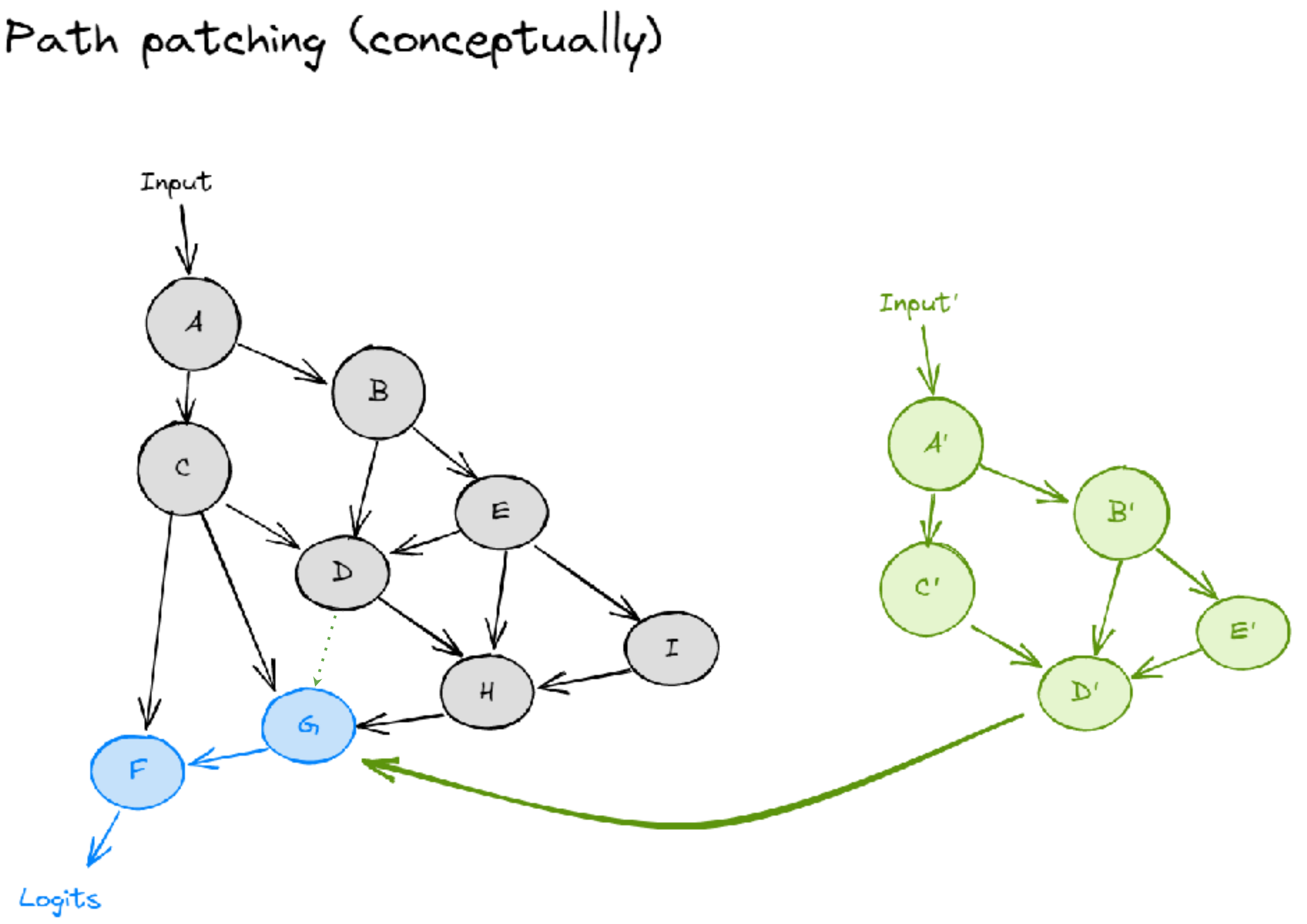
- Statement A is true. Statement B has the same truth value as A. Statement C is false. Statement B is true
- Statement A is true. Statement B has the opposite truth value as A. Statement C is false. Statement B is false
- Either Statement A or statement B is true, but not both. Statement A is false. Statement B is true

To create datasets, we randomly sample letters from the alphabet to be identifiers.

MI Tools Used

To sufficiently understand and explain a LM's behavior, we should explore what individual components of the architecture are doing and how these components interact with each other.

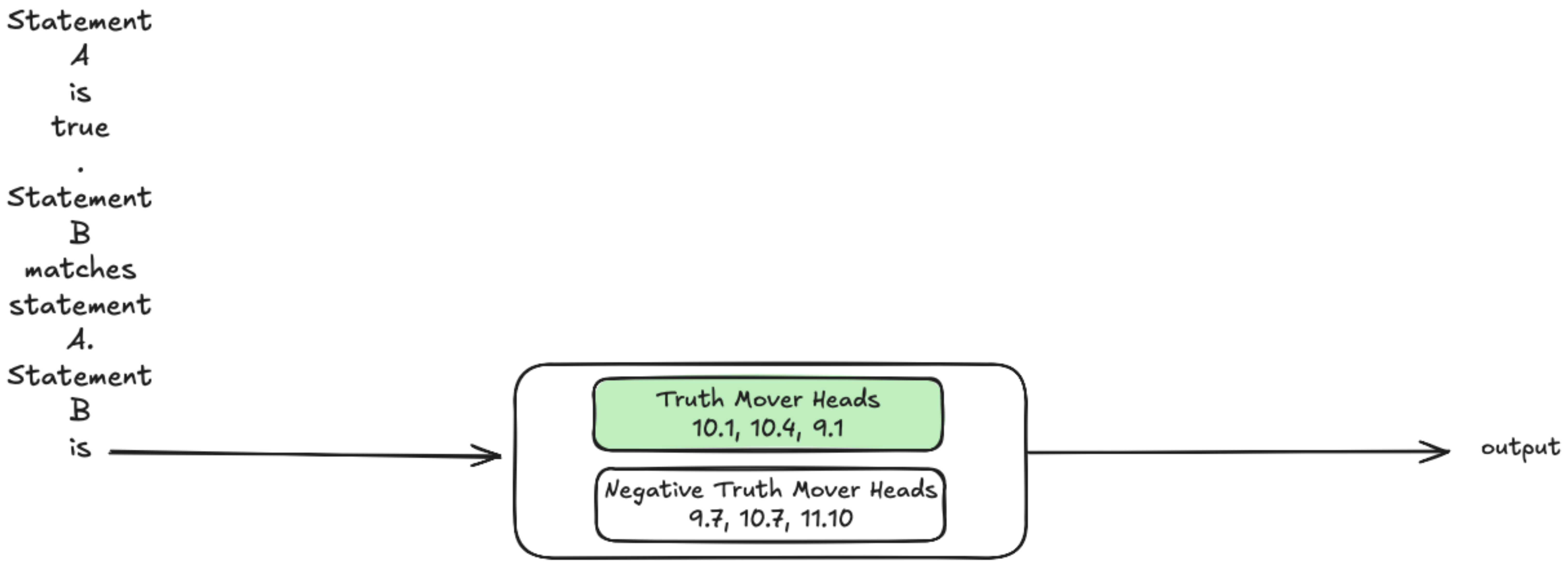
Several techniques exist for this type analysis such as **Logit Lens** (Nostalgebraist, 2020) and Activation Patching (Meng, et al 2023), **Path Patching** (Wang et al, 2022), etc



Corrupted example: Statement W is true. Statement B matches statement E. Statement V is ???

* *Image taken from https://colab.research.google.com/drive/1KgrEwvCKdX-8DQ1uSiluxwliwzJuQ3Gw?usp=sharing#scrollTo=hEs-paRpl6mK*

Simple Syllogism Circuit



To further verify only truth mover heads are necessary for this task, we evaluate the simple syllogism prompt dataset using by ablating all the attention heads except for the truth mover heads.

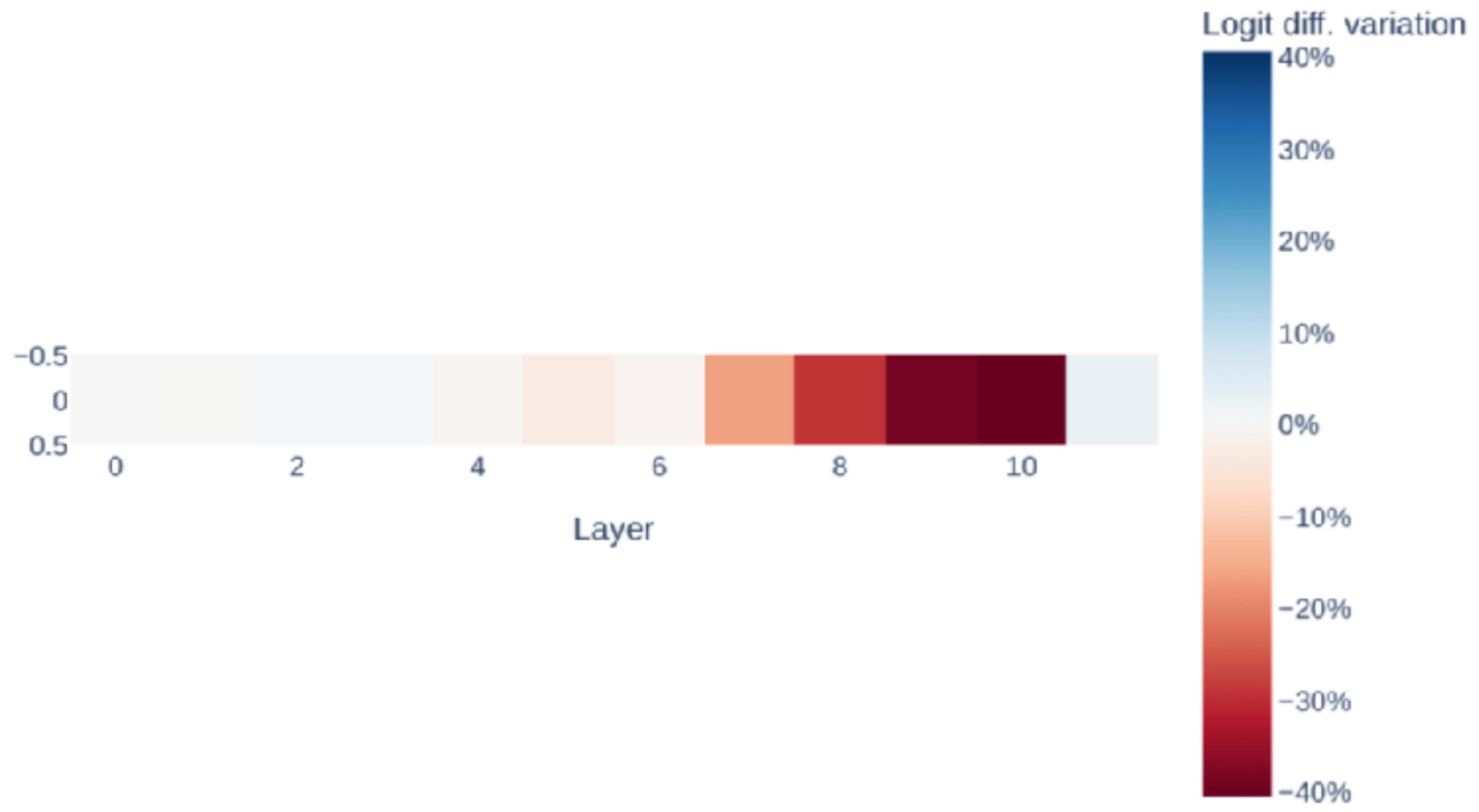
Average logit difference (Simple Syllogism dataset, **using entire model**): 1.6421
Average logit difference (Simple Syllogism dataset, **only using circuit**): 1.4485

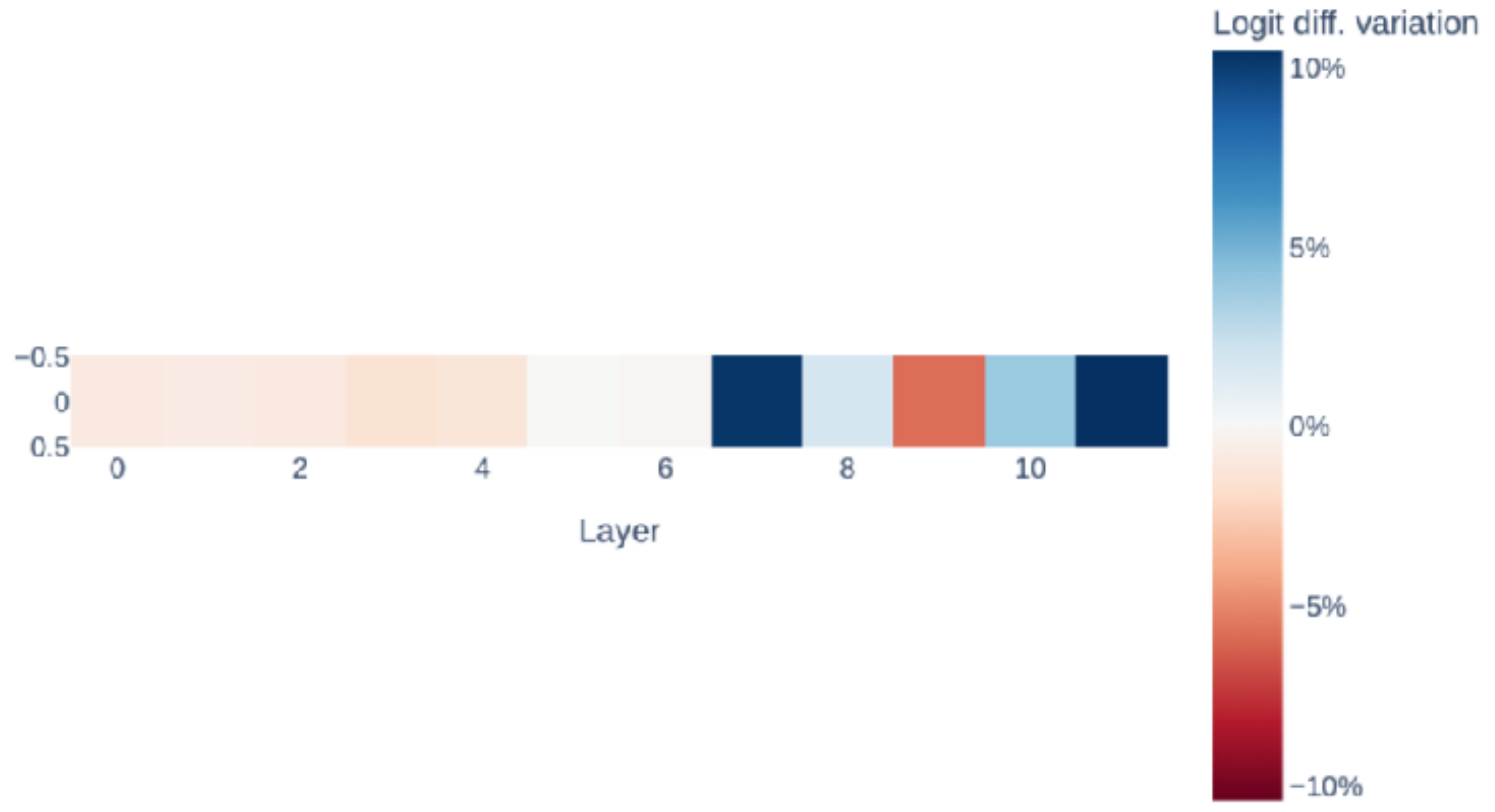
Similar to IOI(Wang et al, 2022), we find a negative version of truth mover heads. But what are they doing? We hypothesized that negative head ablation leads to higher logit difference because these negative heads encode the direction of the less contextualized logit in a binary setting. We move onto opposite syllogism to investigate.

MLPs matter now

When performing path patching within the opposite syllogism format, we now see that later layer MLPs matter (top). A stark contrast to the simple syllogism format, where MLPs were not found to be important (bottom)

We also see aforementioned negative heads matter: 10.7, 9.7, 11.10





Opposite Syllogism (negation mechanism)

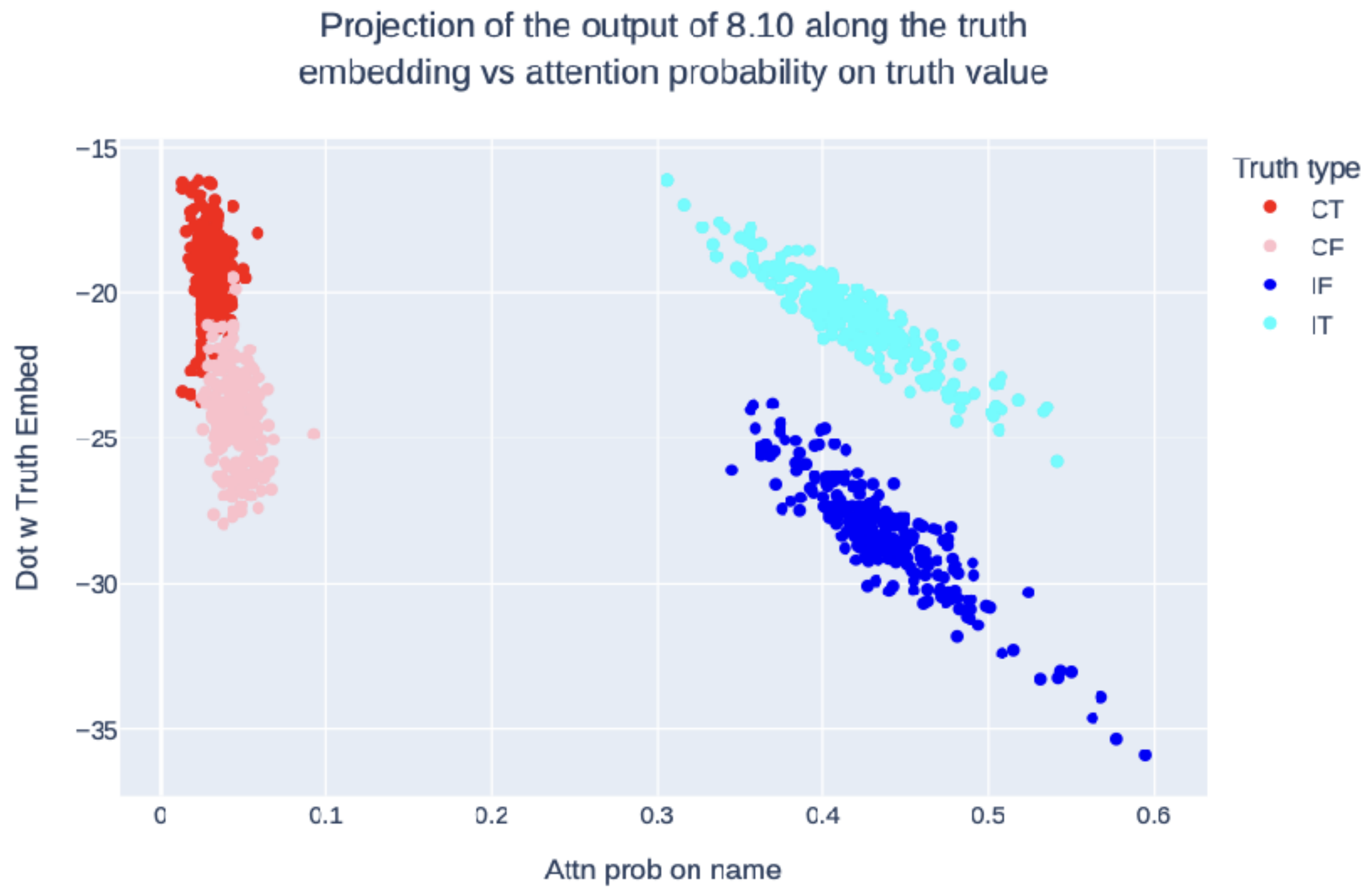
Example Prompt: “Statement E and statement S are opposites. Statement is true. Statement S is”

Top QK Pairs (Head 10.7)
0.892: (‘is’, **‘true’**), **0.772:** (‘statement’, ‘E’), **0.685:** (‘Statement’, ‘S’), **0.662:** (‘Statement’, ‘S’), **0.459:** (‘is’, ‘oppos’)

Stage	**Top Logits**	**Bottom Logits**
After OV from Head 10.7	depot, rink, carp, Dj, Hack, DJ, Gaz, Phillips, District, TTC	**‘true’**, ‘True’, ‘TRUE’, **‘true’**, ‘untrue’, ‘Null’
After MLP Layer 10	**‘true’**, **‘false’**, **‘True’**, **‘False’**, infinite, truly	blitz, ombo, plateau, corrid, tradem, emale, Citiz, sugg

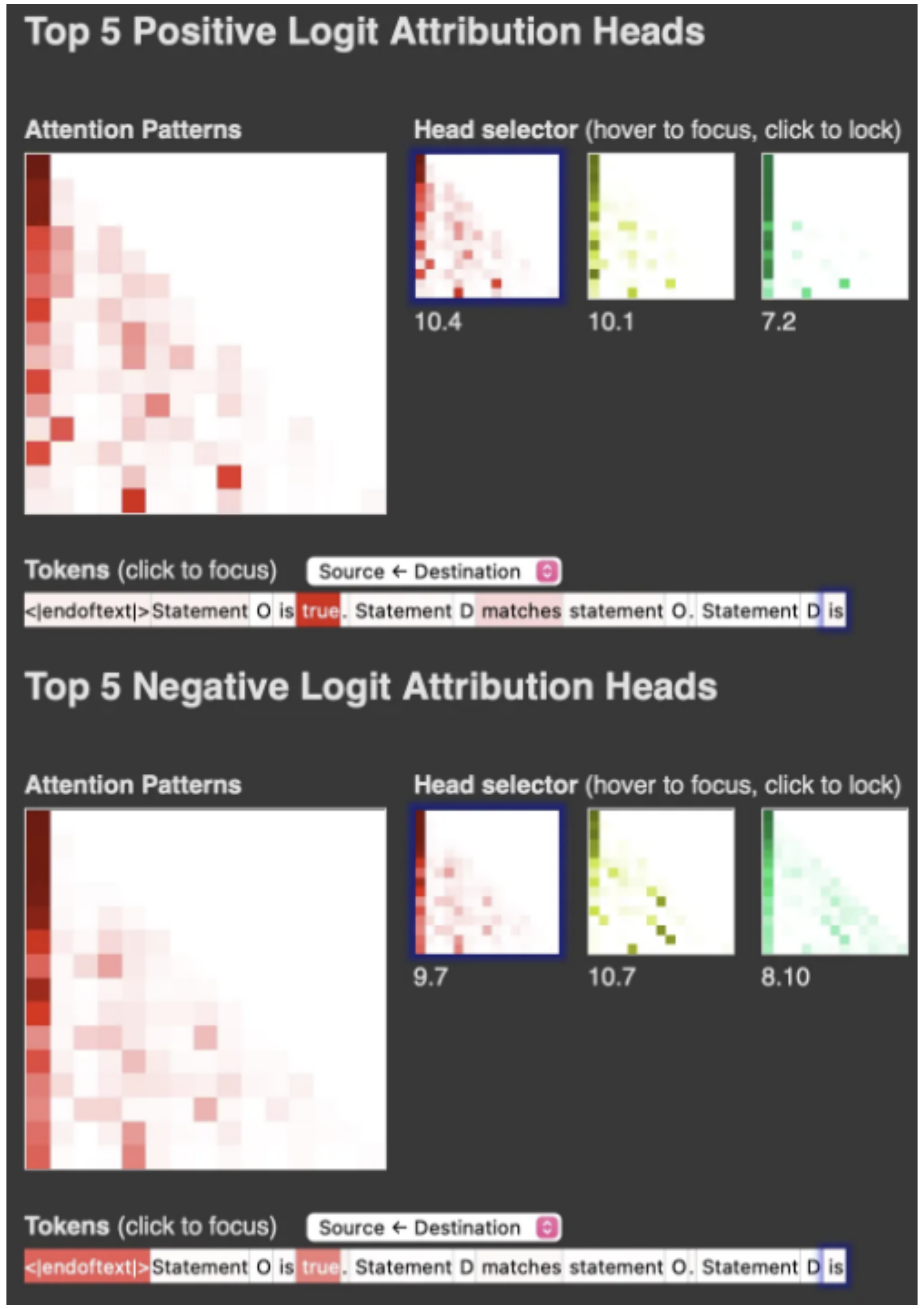
Complex Syllogism Findings

- **Emergence of Modulation Heads:** In complex syllogisms, new attention heads appear that identify which truth token is relevant and suppress distracting ones.
- **Disentangled Functions:** The model separates roles — *Modulation Heads* decide *what* to focus on, while *Negative Heads* and MLPs handle *how* to negate it.
- **MLP Negation under Guidance:** The same MLP mechanism that flips truth values in opposite syllogisms still operates here, but it's now directed by the modulation heads' selection of the correct token.

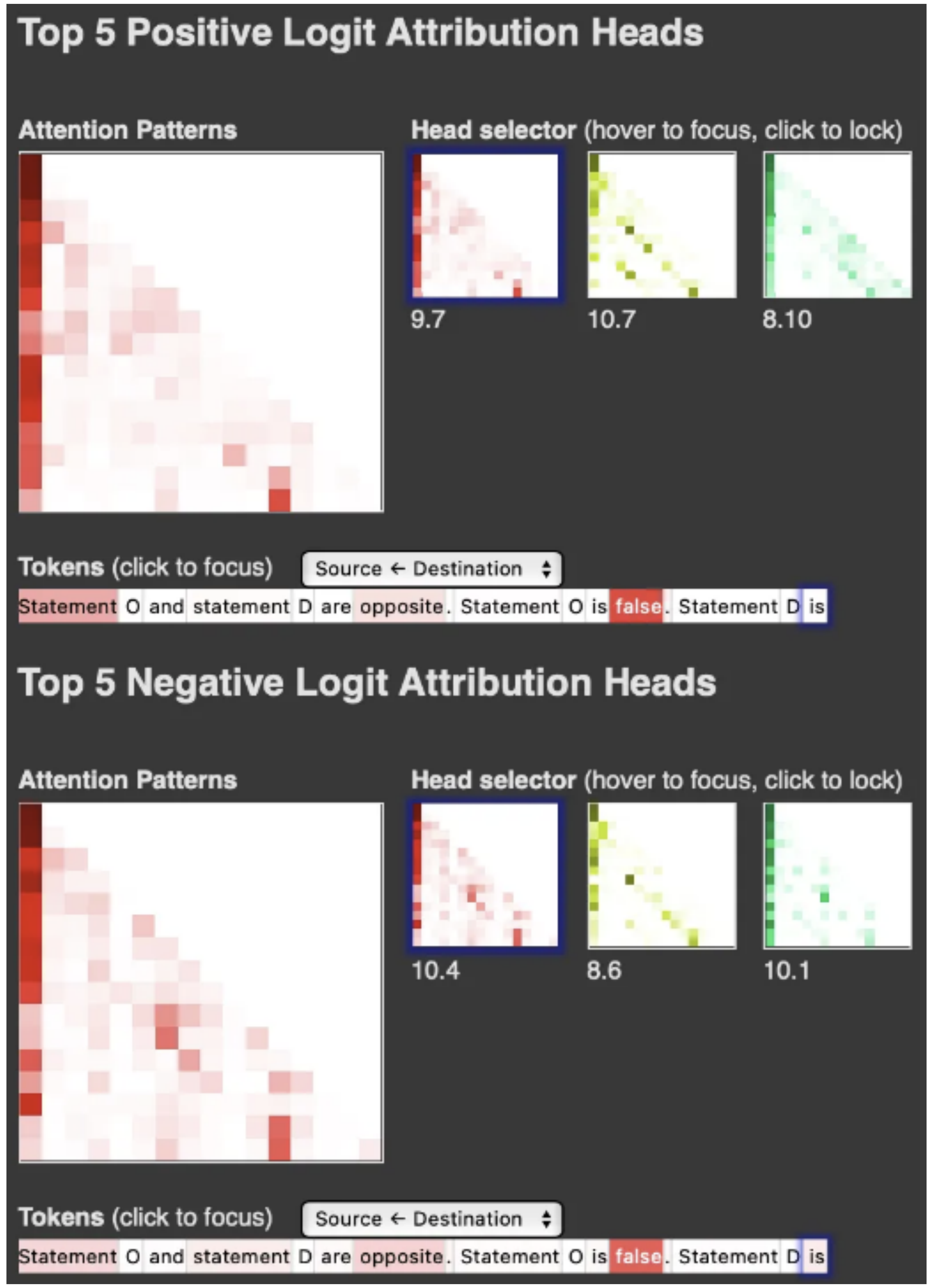


Binary Nature in Attention Heads

Syllogism



Opposite Syllogism



The same heads are playing the opposite role in the opposite task

Binary Nature extends beyond T/F

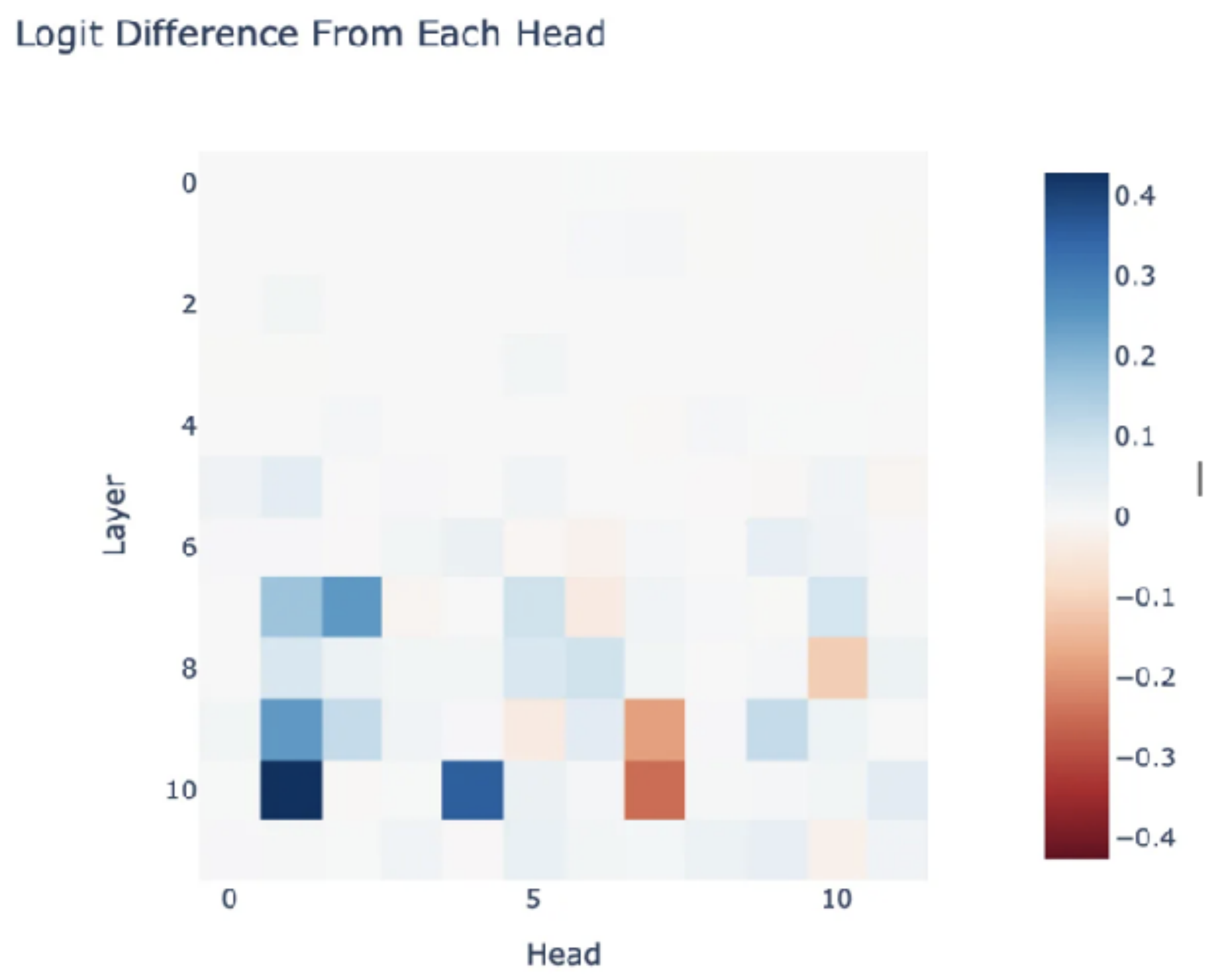
In order to further verify this phenomenon, we replicate the same experiments with different binary pairs

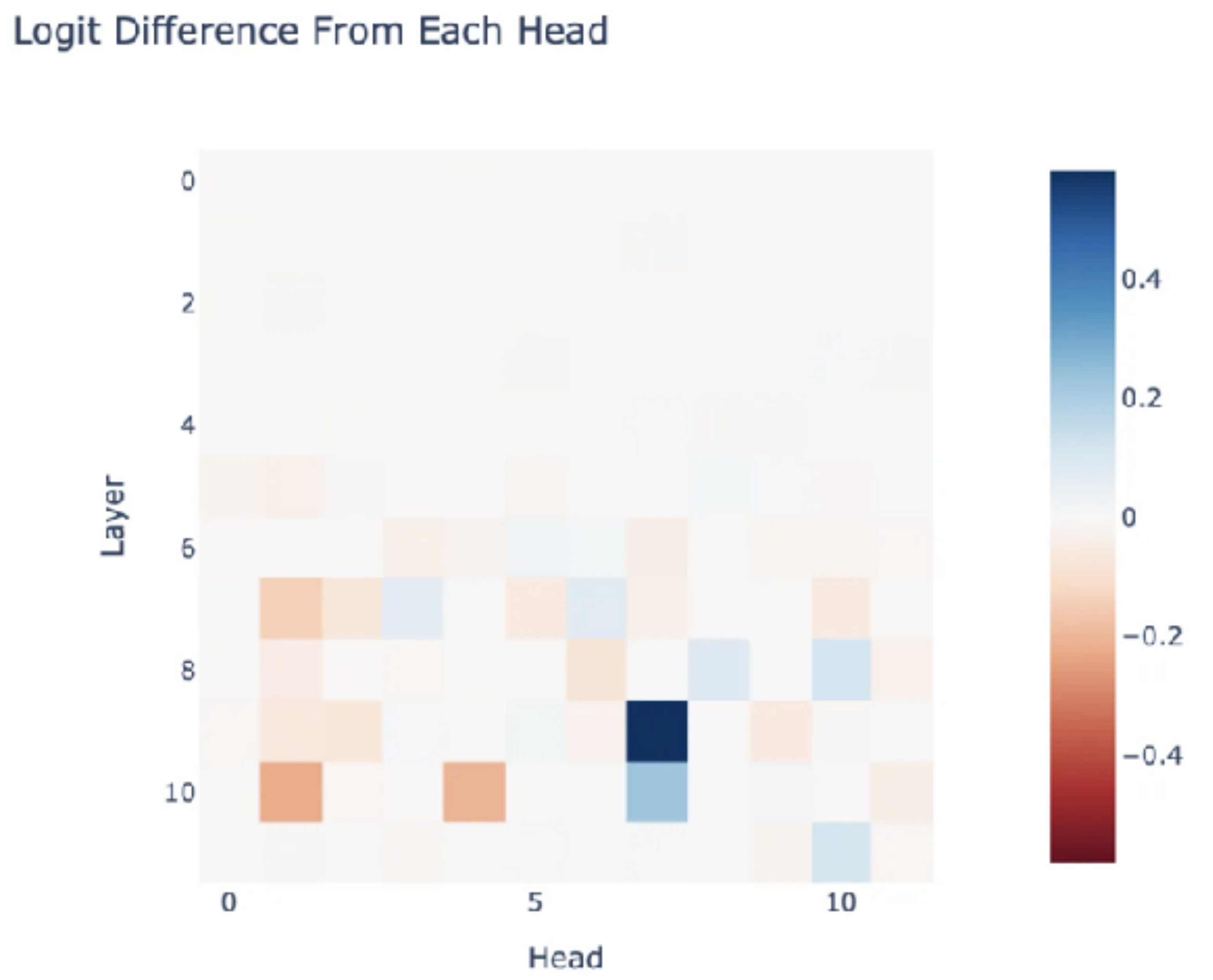
- right vs wrong
- good vs bad
- positive vs negative
- correct vs incorrect
- All binary cases combined

All pairs verify the binary nature extends beyond true/false. Visually, the colors flip with logit difference between the two tasks

	Original	Good/Bad	Pos/Neg	Correct/Incorrect	Right/Wrong
GPT-2 Small	1.8399	1.7738	0.6958	2.1221	2.0309
C_{SS}	1.9234	1.9940	1.1584	1.6785	2.1599

	Original	Good/Bad	Pos/Neg	Correct/Incorrect	Right/Wrong
GPT-2 Small	1.2632	2.1163	3.0032	0.7986	1.3469
C_{OS}	1.3136	1.7136	1.0113	0.8142	1.2481





Thank you!

Paper

LinkedIn

UCLA Samueli School of Engineering

How Post-Training Reshapes LLMs: A Mechanistic View on Knowledge, Truthfulness, Refusal, and Confidence

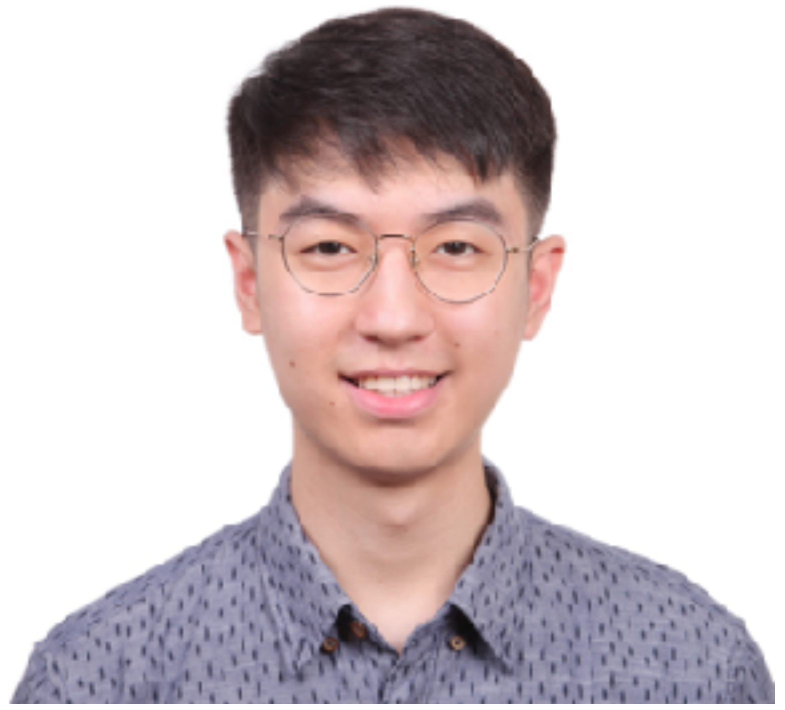
Hongzhe Du

Thanks to My Amazing Collaborators

Weikai Li
UCLA

Min Cai
University of Alberta

Karim Saraipour
UCLA



Zimin Zhang
UIUC

Himabindu Lakkaraju
Harvard

Yizhou Sun
UCLA

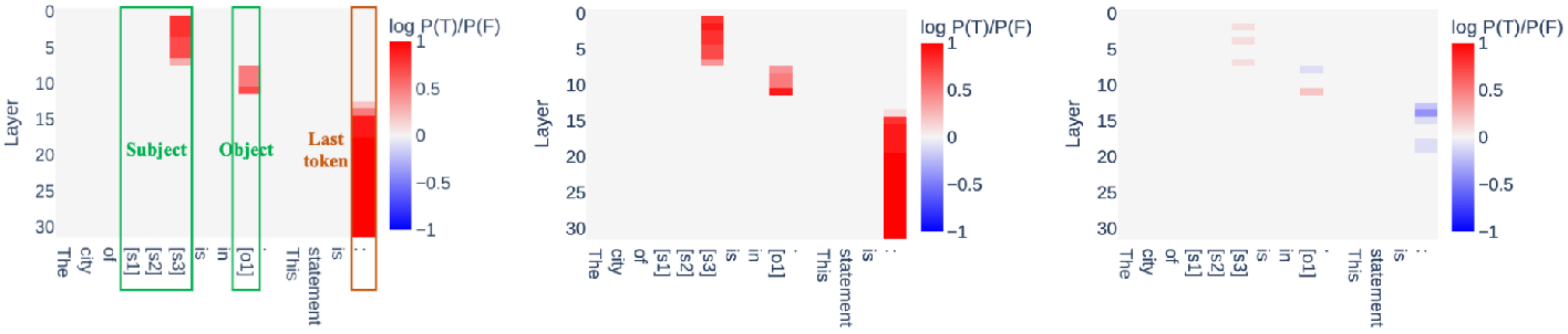
Shichang Zhang
Harvard

Post-training effects are usually evaluated externally through the model output.

HOW ABOUT INTERNALLY?

Knowledge Perspective (Causal Tracing)

- Compare Causal Tracing results before and after post-training

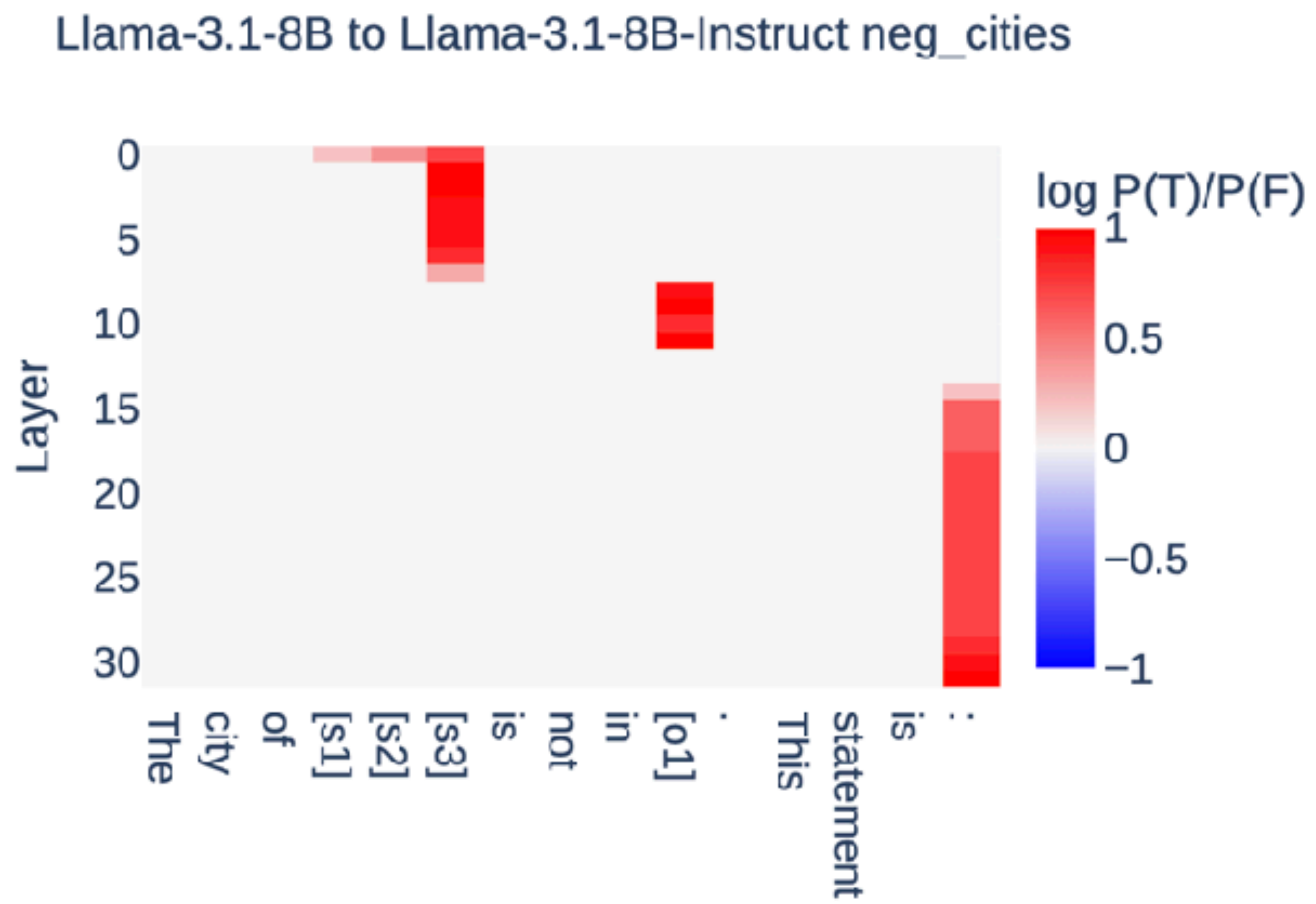


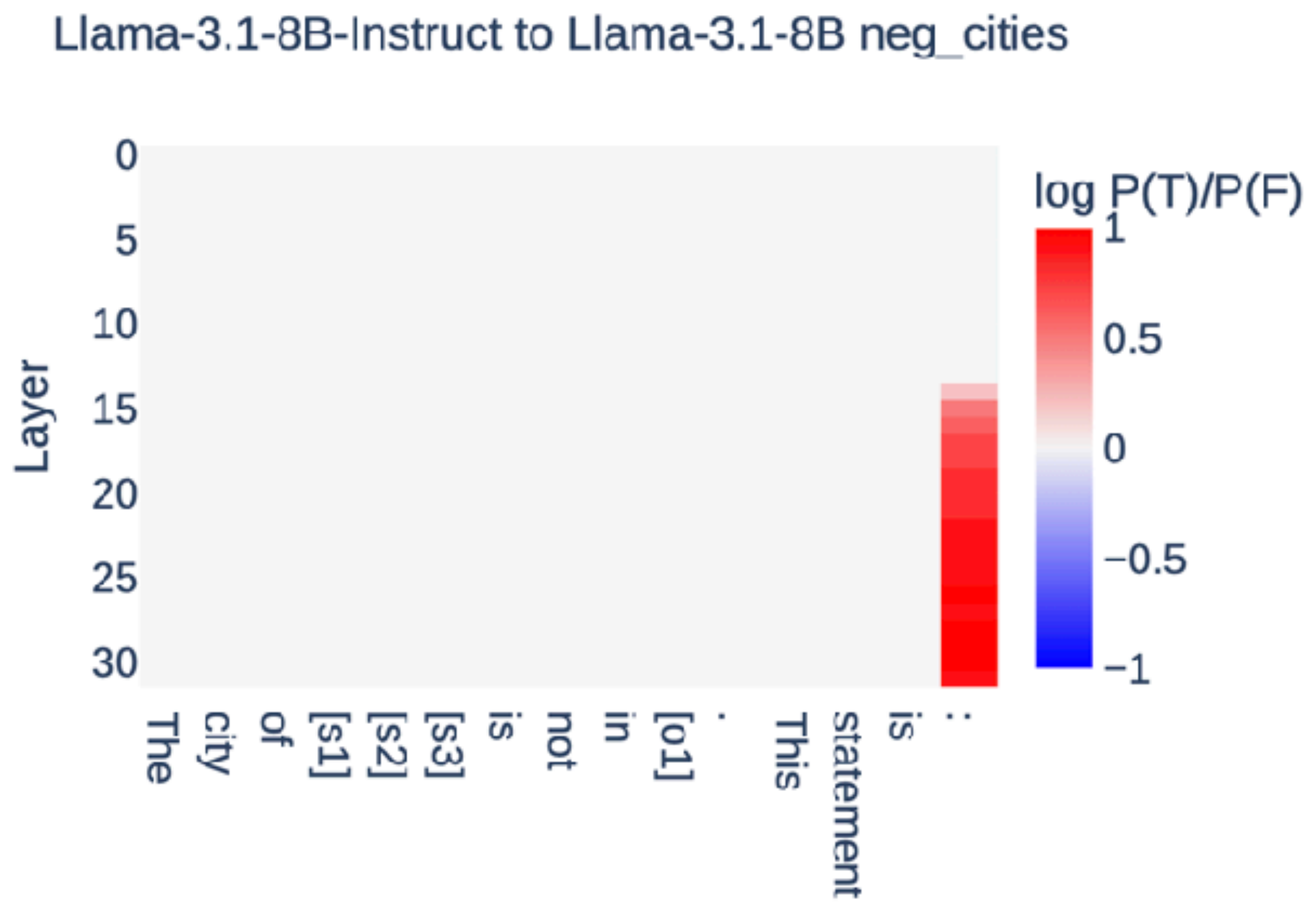
(a) BASE. (b) INSTRUCT. (c) Difference.

Figure 2: Knowledge storage locations of Llama-3.1-8B BASE and INSTRUCT on the `cities` dataset. Their knowledge-storage locations are almost the same.

Cross-Patching

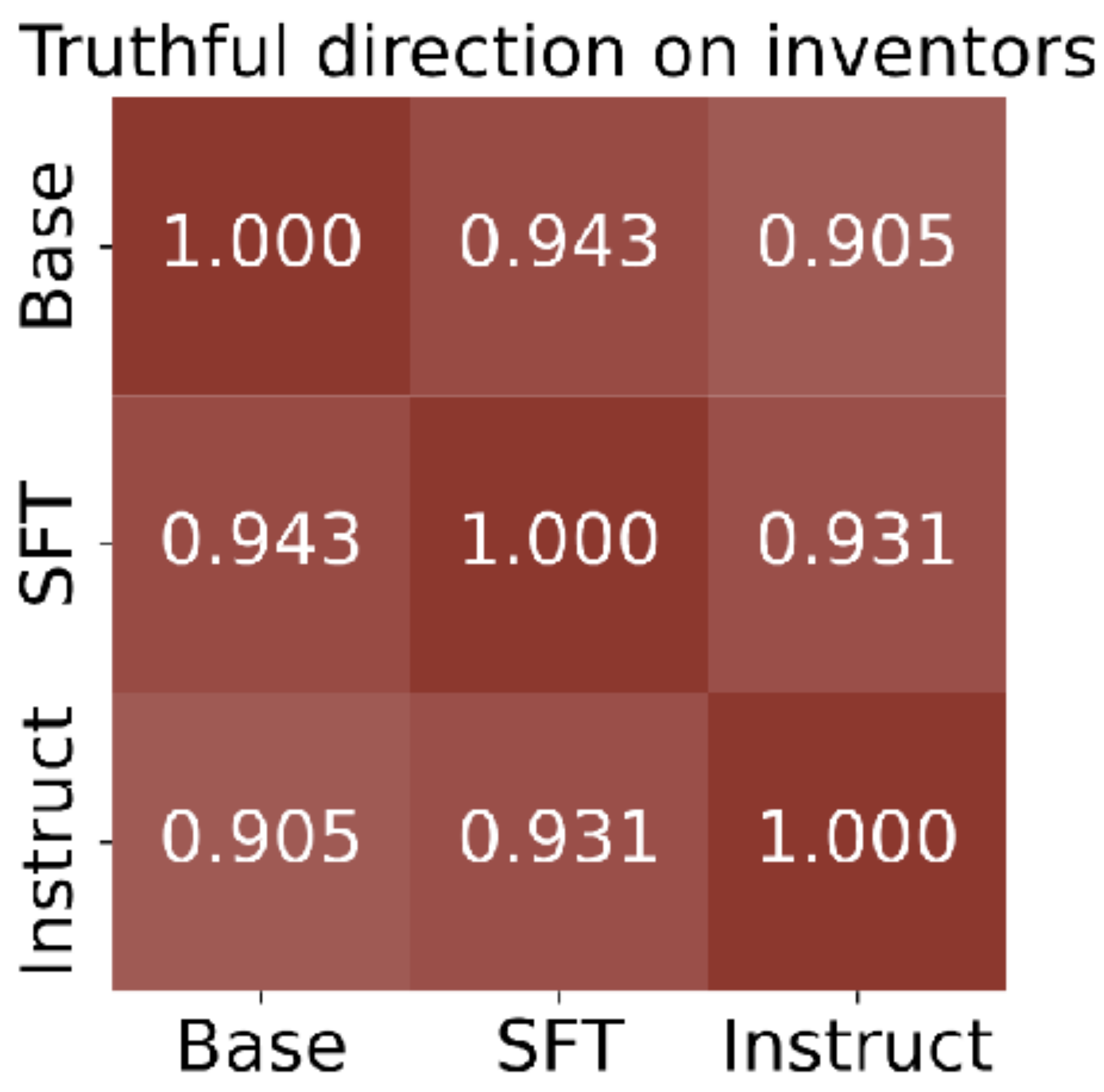
- Base to Post-trained patching -> Successful
- Post-trained to Base patching -> Unsuccessful

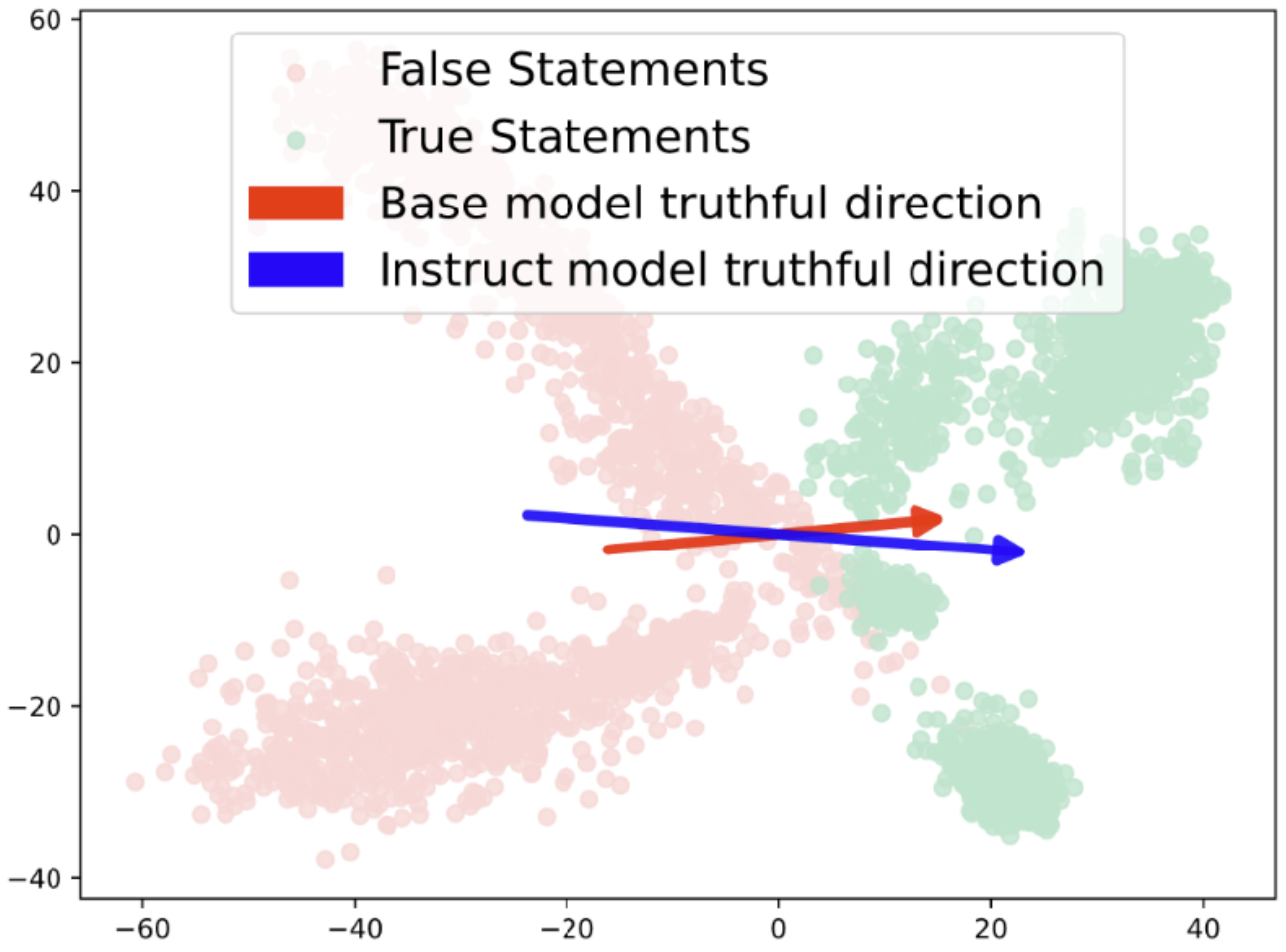




Internal Belief of Truthfulness (Linear Probe)

- Truthful directions are similar.





UCLA Samueli School of Engineering

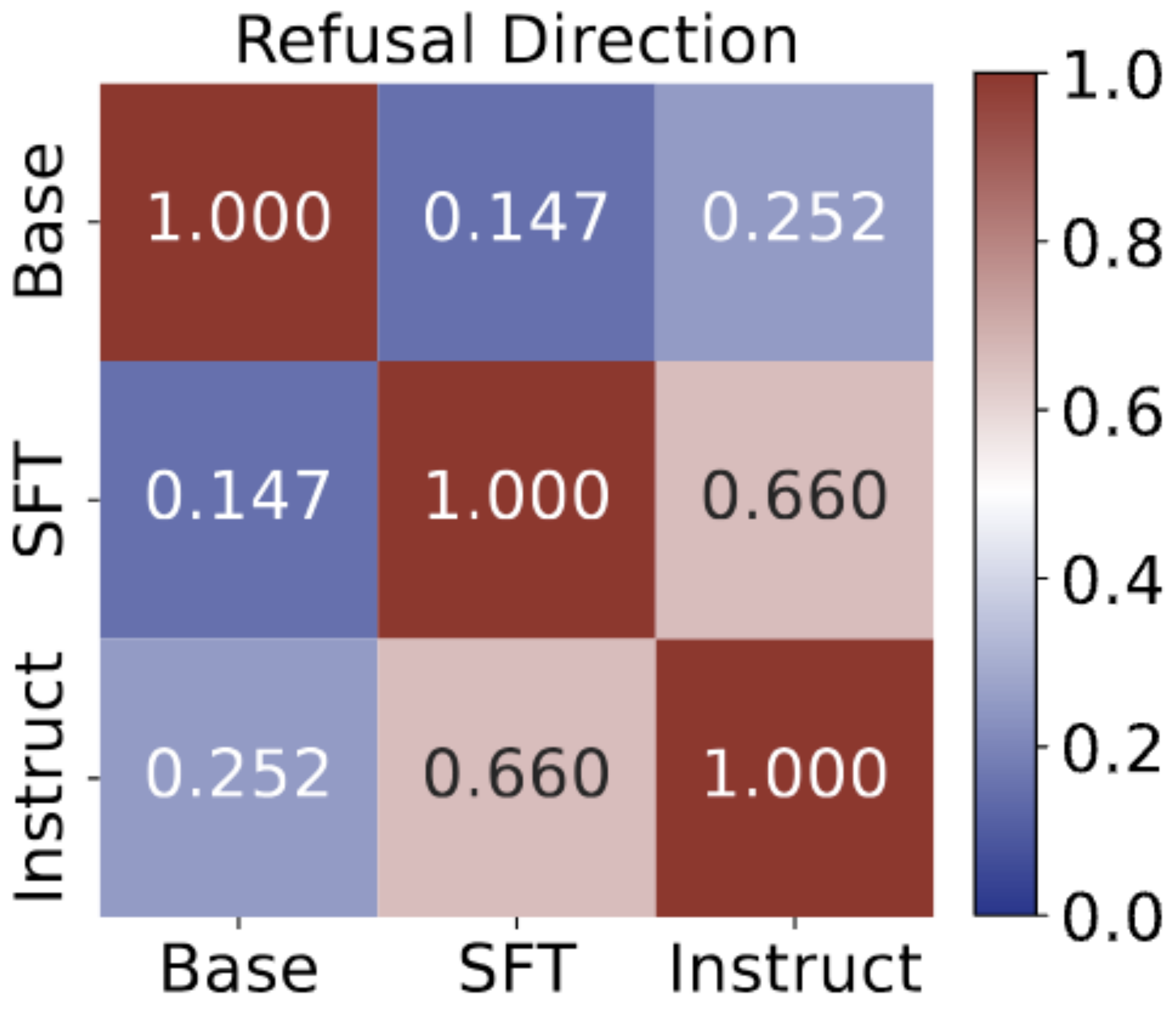
Successful Probe and Intervention Transfer from Base model to Post-trained Model

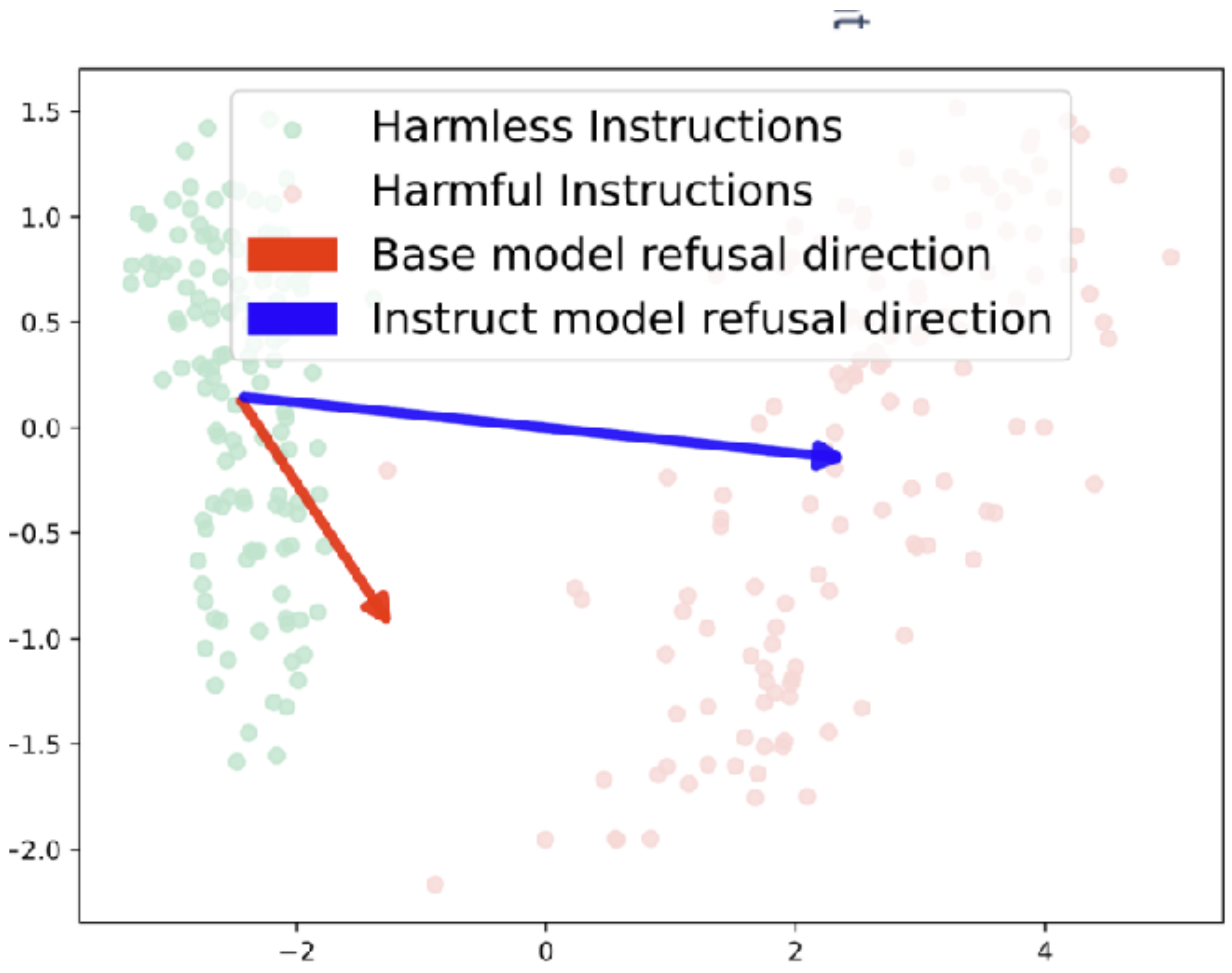
Test Dataset	Probe Transfer Accuracy (%)		
	$p_{\text{BASE}} \rightarrow h_{\text{BASE}}$	$p_{\text{SFT}} \rightarrow h_{\text{SFT}}$ / $p_{\text{BASE}} \rightarrow h_{\text{SFT}}$ (Δ)	$p_{\text{INS}} \rightarrow h_{\text{INS}}$ / $p_{\text{BASE}} \rightarrow h_{\text{INS}}$ (Δ)
cities	81.06	84.50 / 85.32 (+0.82)	94.65 / 95.91 (+1.26)
sp_en_trans	97.16	98.45 / 98.88 (+0.43)	95.18 / 98.94 (+3.76)
inventors	92.72	91.96 / 93.12 (+1.16)	88.73 / 92.18 (+3.45)
animal_class	97.20	96.01 / 95.64 (-0.37)	98.75 / 96.46 (-2.29)
element_symb	92.02	94.87 / 97.02 (+2.15)	96.18 / 95.13 (-1.05)
facts	77.05	77.58 / 77.72 (+0.14)	82.47 / 80.86 (-1.61)

Test Dataset	Truthful Intervention Effects		
	$t_{\text{BASE}} \mapsto h_{\text{BASE}}$	$t_{\text{SFT}} \mapsto h_{\text{SFT}}$ / $t_{\text{BASE}} \mapsto h_{\text{SFT}}$ (Δ)	$t_{\text{INS}} \mapsto h_{\text{INS}}$ / $t_{\text{BASE}} \mapsto h_{\text{INS}}$ (Δ)
cities	0.83	0.91 / 0.92 (+0.01)	0.88 / 0.90 (+0.02)
sp_en_trans	0.78	0.82 / 0.83 (+0.01)	0.84 / 0.81 (-0.03)
inventors	0.72	0.80 / 0.82 (+0.02)	0.79 / 0.83 (+0.04)
animal_class	0.73	0.79 / 0.80 (+0.01)	0.71 / 0.72 (+0.01)
element_symb	0.79	0.84 / 0.86 (+0.02)	0.73 / 0.77 (+0.04)
facts	0.61	0.64 / 0.66 (+0.02)	0.62 / 0.66 (+0.04)

Refusal (Linear Probe)

- Refusal directions are different.





Unsuccessful Refusal Intervention Transfer from Base model to Post-trained Model

	Intervention Refusal Score		
	BASE	SFT	INSTRUCT
Inputs	baseline/$r_{\text{BASE}} \mapsto h_{\text{BASE}}$	baseline/$r_{\text{SFT}} \mapsto h_{\text{SFT}}$/$r_{\text{BASE}} \mapsto h_{\text{SFT}}$	baseline/$r_{\text{INS}} \mapsto h_{\text{INS}}$/$r_{\text{SFT}} \mapsto h_{\text{INS}}$/$r_{\text{BASE}} \mapsto h_{\text{INS}}$
harmful (↓)	0.21 / 0.17	0.99 / 0.79 / 0.99	0.98 / 0.01 / 0.36 / 0.95
harmless (↑)	0.01 / 0.59	0.01 / 1.0 / 0.85	0.0 / 1.0 / 0.98 / 0.08

Table 4: Intervention RS of Llama-3.1-8B BASE, SFT, and INSTRUCT tested on harmful and harmless inputs. $r_{model_1} \mapsto h_{model_2}$ means using the refusal direction in *model*$_1$ to intervene *model*$_2$, and baseline refers to the original Refusal Score without intervention. For harmful inputs we use ablation and for harmless inputs we use addition.

Confidence (Entropy Neuron)

- Base model and post-trained model have very similar entropy neurons.
- Confidence difference between two models cannot be attributed to entropy neurons

Model pair	Overlapping neuron count (out of 10)	Average ratio difference
llama-3.1-8b BASE vs INSTRUCT	8	0.000815
llama-3.1-8b BASE vs SFT	10	0.000112
mistral-7b BASE vs INSTRUCT	9	0.000030
mistral-7b BASE vs SFT	8	0.000089
llama-2-7b BASE vs INSTRUCT	9	0.001712

Table 14: Entropy neuron results. "Overlapping neuron count" shows the number of overlapping entropy neurons between BASE and POST models. "Average ratio difference" shows the average difference of $|\frac{\text{weight norm}}{log(\text{LogitVar})}|$ of the overlapping entropy neurons between BASE and POST models. As a reference, the average $|\frac{\text{weight norm}}{log(\text{LogitVar})}|$ is 0.0880 for all entropy neurons, which is much larger than the difference. BASE models and POST models have very similar entropy neurons.

On the Retention of Edited Knowledge in Fine-tuned Language Models

Fufang Wen   Shichang Zhang

Main Problem: Knowledge edit can be successful, but then fine-tuning erases those edits.

Prompt: *Windows Mobile 6.5 was developed by*

Key Findings

- Edited Facts Are Fragile

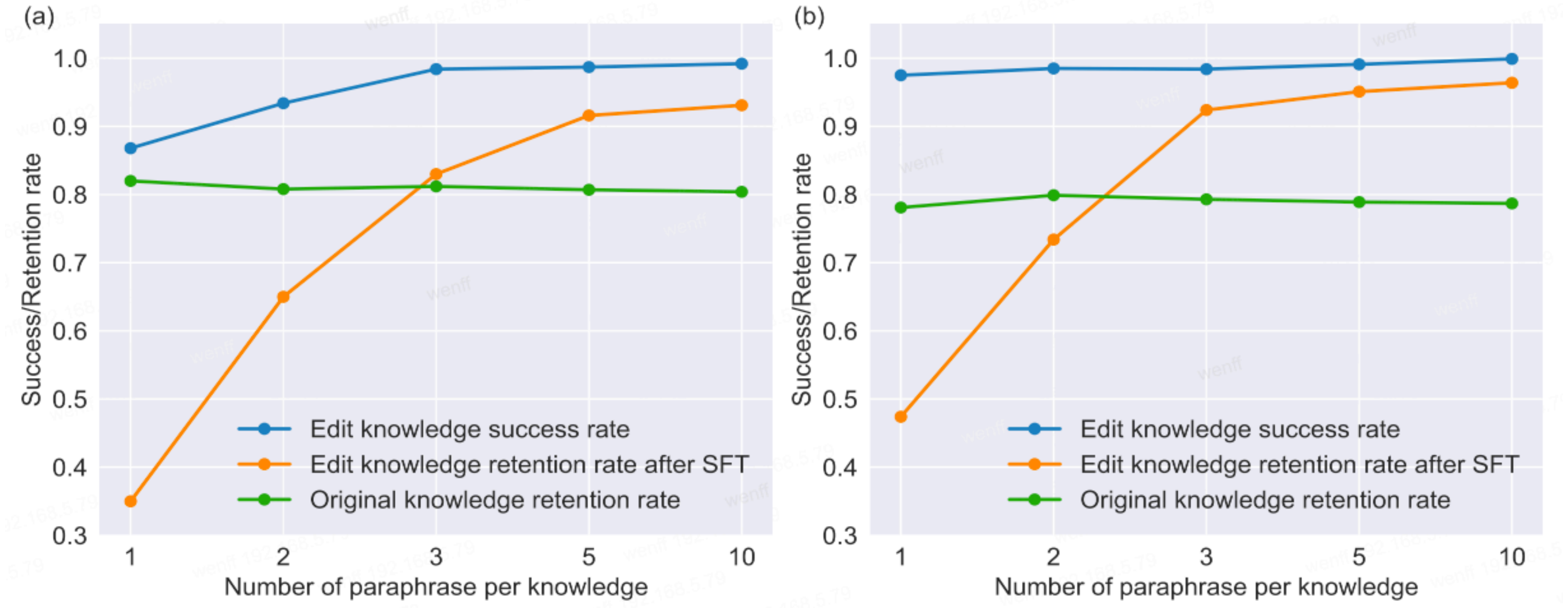
1. Fine-tuning easily erases knowledge from editing tools (like ROME, MEMIT).
2. The model doesn't just forget; it often replaces the fact with a related one (e.g., "Apple" → "Intel").

- The "Why": Elasticity Theory

Original knowledge: Learned from a few of examples → strong and stable.
Edited knowledge: Learned from just 1 example → weak and low resilience.

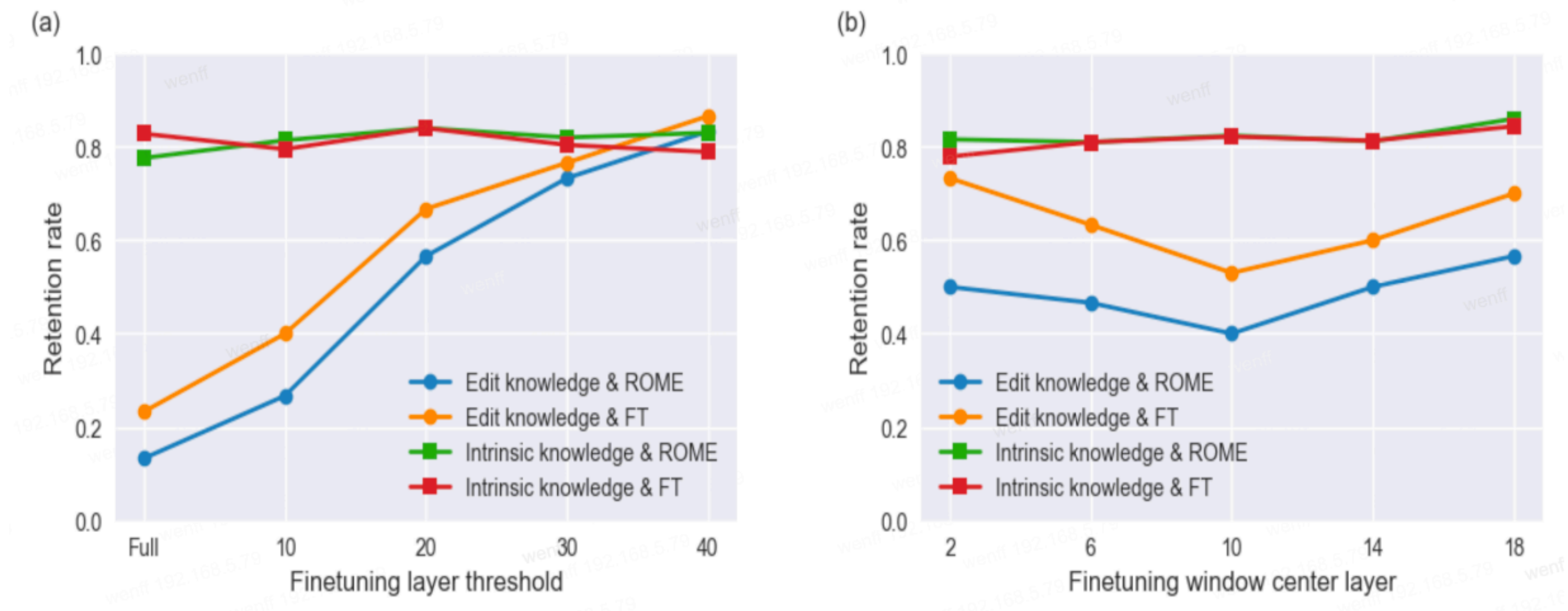
Two Simple Solutions

1. Add paraphrases when edit: Teach the new fact in different ways (3+ versions) to make it stick.



Two Simple Solutions

2. Freeze Key Layers when fine-tune: During fine-tuning, protect the part of the model where the edit was made.



Conclusion

1. We must test if the edits last after fine-tuning.

2. Add paraphrases when edit or freezing layers when fine-tune makes edited knowledge as stable as original facts.

Sarc7: Evaluating Sarcasm Detection and Generation with Seven Types and Emotion-Informed Techniques

Lang Xiong, Raina Gao, Alyssa Jeong, Yicheng Fu, Kevin Zhu, Sean O'Brien, and Vasu Sharma

COLM ORIGen 2025

Motivation

Binary Sarcasm Detection (Prior Work):

Supervised PLMs (RoBERTa): 71.5% F1 (Zhang et al., 2024)
Best LLM (GPT-4): 65.0% F1 on binary classification (Zhang et al., 2024)
Multimodal approaches: 71.6% F1 (Castro et al., 2019)

Our Contribution:

Sarc7: First 7-way sarcasm classification benchmark
Emotion-based prompting improves generation by 38.4%
Shows fine-grained sarcasm understanding remains an open challenge

Why Sarcasm Matters for Reliable LLM Deployment:

Sentiment Analysis Fails
Content Moderation Errors
Document Summarization
Information Extraction

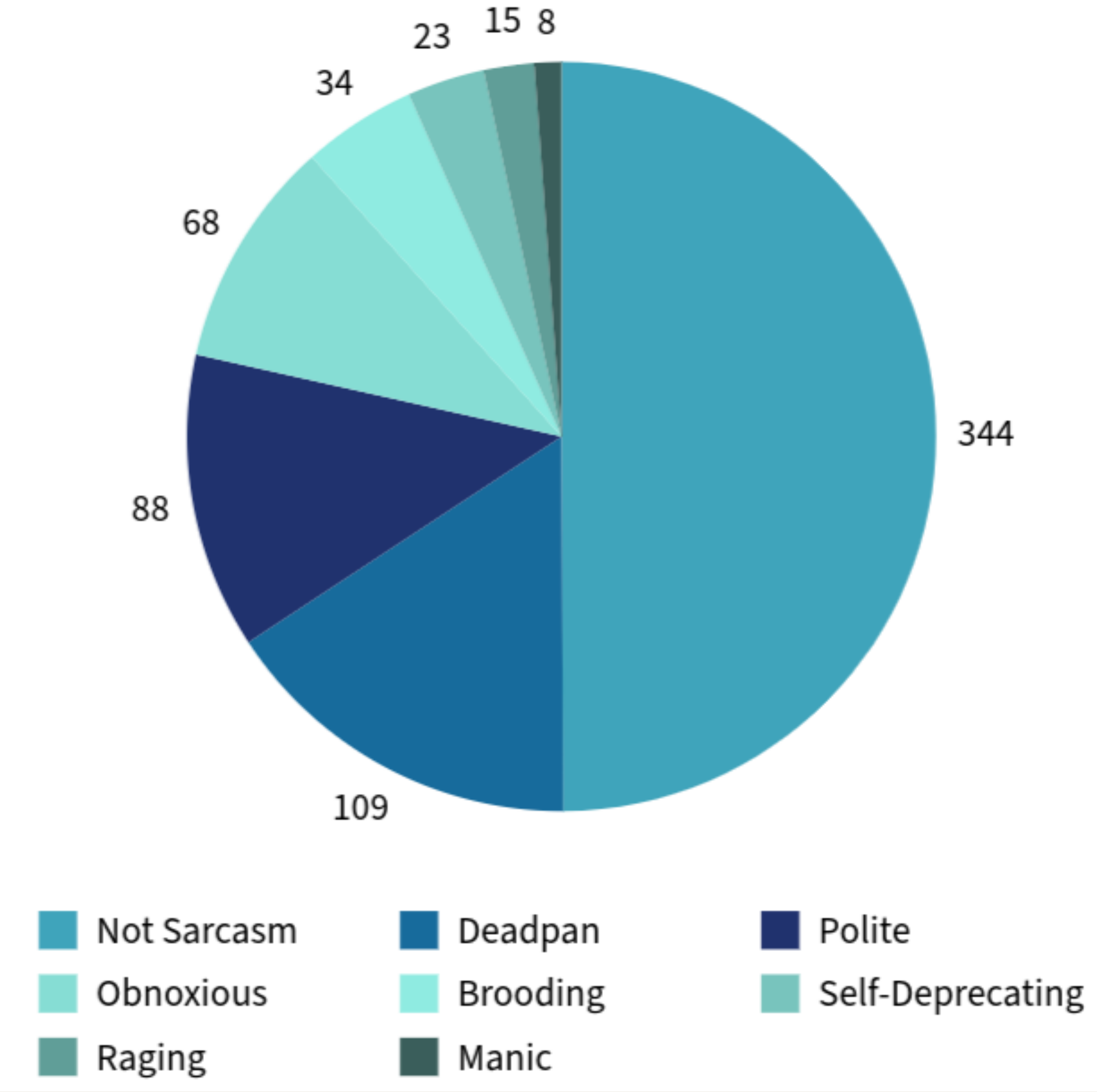
Sources:

Castro et al., 2019: MUStARD multimodal dataset
Zhang et al., 2024: SarcasmBench LLM evaluation
Qasim, 2021: Seven types of sarcasm taxonomy

Benchmark Construction

MUStARD dataset (Castro et al., 2019):
690 binary sarcasm annotations for short dialogue segments
Annotation:
Each utterance was first labeled independently by all four annotators.
• If at least three annotators agreed on the same label, that label was accepted as the final annotation.
• In cases with no 3-out-of-4 agreement, a consensus discussion was held between annotators, with a final decision made by majority vote.
• Cohen's κ = 0.669 with a fifth annotator
Even for trained readers, **brooding, deadpan, and polite** sarcasm proved the most challenging to label

Type	Definition	Example
Self-deprecating	Mocking oneself in a humorous or critical way.	"Oh yeah, I'm a genius — I only failed twice!"
Brooding	Passive-aggressive frustration masked by politeness.	"Sure, I'd love to stay late again — who needs weekends?"
Deadpan	Sarcasm delivered in a flat, emotionless tone.	"That's just the best news I've heard all day."
Polite	Insincere compliments or overly courteous remarks.	"Wow, what an *interesting* outfit you've chosen."
Obnoxious	Rude or provocative sarcasm aimed at others.	"Nice driving! Did you get your license in a cereal box?"
Raging	Intense, exaggerated sarcasm expressing anger.	"Of course! I *love* being yelled at in meetings!"
Manic	Overenthusiastic, erratic sarcasm with chaotic tone.	"This is AMAZING! Who needs food or sleep anyway?!"



Pipeline

Sarcasm Classification: Given a sarcastic utterance and its dialogue context, correctly predict the dominant sarcasm type from among the seven annotated categories.
Sarcasm Generation: Generate a sarcastic utterance consistent with one of the 7 types of sarcasm.
Prompting:
• Zero-Shot
• Few-Shot
• Chain-of-Thought (COT)
• Emotion-based happiness, sadness, anger, fear, disgust, surprise Ekman (1992), and neutral.
Generation Dimensions
• Incongruity: Degree of semantic mismatch (1-10).
• Context Dependency: Reliance on conversational history.
• Shock Value: Intensity of sarcasm.
• Emotion: One of Ekman's six basic emotions (e.g., anger, sadness).

Classification

- GPT-4o
- Claude 3.5 Sonnet
- Gemini 2.5 Pro
- Qwen 2.5
- Llama Maverick 4

Generation

- Claude 3.5 Sonnet

Classification Results

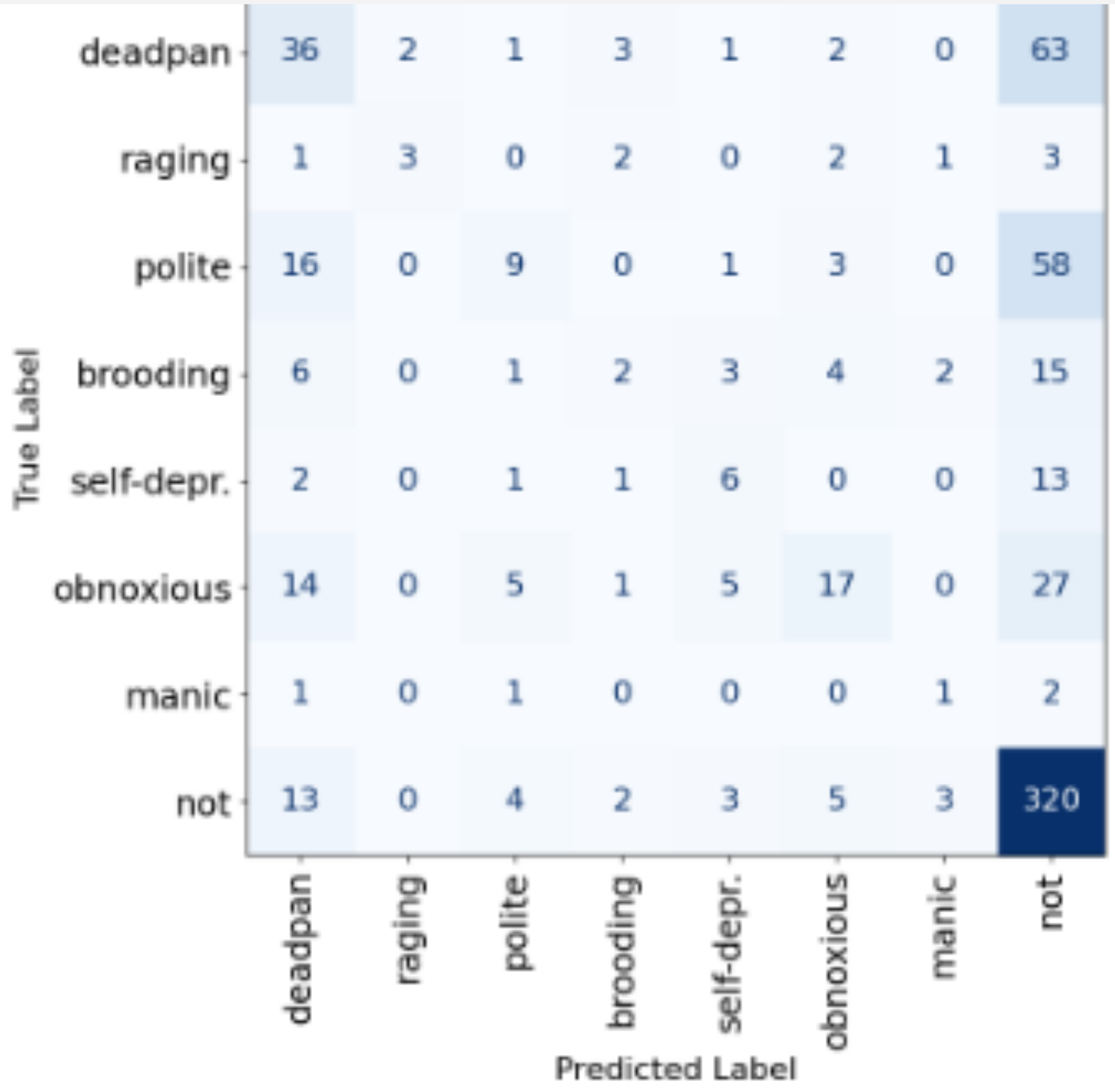


Figure 4: Confusion Matrix for Claude 3.5 Sonnet using CoT.

Defaulted to deadpan or no sarcasm

Model	0-shot	Few-shot	CoT	Emotion-based
GPT-4o	47.73%	50.29%	**55.07%**	48.94%
Claude 3.5 Sonnet	51.16%	52.61%	**57.10%**	52.32%
Qwen 2.5	41.45%	**46.96%**	46.09%	45.94%
Llama-4 Maverick	34.20%	35.51%	**50.29%**	49.86%
Gemini 2.5	46.81%	47.97%	**53.04%**	52.03%

Table 3: Classification Accuracy Across Models and Prompting Techniques

Model	0-shot F1	Few-shot F1	CoT F1	Emotion-based F1
GPT-4o	0.2089	**0.3255**	0.2674	0.2233
Claude 3.5 Sonnet	0.2964	0.3487	0.2471	**0.3487**
Qwen 2.5	0.2116	0.2075	0.2052	**0.2124**
Llama-4 Maverick	0.2184	0.2340	0.2040	**0.2841**
Gemini 2.5	0.2760	0.3274	0.3141	**0.3664**

Table 4: Macro-averaged F1 scores of Models Across Prompting Techniques.

Subtype	CoT	Emotion-based	Human
Brooding sarcasm	6.06%	9.09%	39.39%
Deadpan sarcasm	33.03%	50.46%	55.45%
Polite sarcasm	10.34%	33.33%	57.30%
Manic sarcasm	20.00%	20.00%	75.00%
Obnoxious sarcasm	24.64%	39.13%	67.14%
Raging sarcasm	25.00%	41.67%	71.43%
Self-deprecating sarcasm	26.09%	34.78%	86.96%
Not sarcasm	91.17%	66.38%	95.04%

Table 5: Per-class Accuracy for Claude 3.5 using CoT vs. Emotion-based Prompting, Alongside Human Agreement.

Generation Results

Claude 3.5 Sonnet produced 100 sarcastic statements per prompting method, each rated by a human for sarcasm type accuracy. Brooding and Manic were hardest to generate.

Prompt	Successful Generation
Zero-shot	52/100
Emotion-based	**72/100**

Table 6: Generation Evaluation Scores

Subtype	Incongruity (1–10)	Shock Value	Context Dependency	Emotion
Self-deprecating	3–5	low	medium	sadness
Brooding	5–7	medium	medium	anger
Deadpan	4–6	low	high	neutral
Polite	3–5	low	medium	happiness
Obnoxious	6–9	high	low	disgust
Raging	7–9	high	low	anger
Manic	5–7	high	medium	surprise

Table 2: Dimension Settings and Target Emotion for Each Sarcasm Subtype used in our Emotion-based Prompting.

Impact & Takeaways:

Contributions:

- First multi-class sarcasm benchmark (7 types)
- Emotion-based prompting technique
- Evidence that fine-grained sarcasm remains challenging for LLMs

Implication: Better sarcasm understanding = more human-like AI communication

Limitation: small dataset, single-label, one language

Future direction:

- Multi-label classification
- Bigger, multilingual dataset
- Multimodal

Type	GPT-4o	Claude 3.5 Sonnet	Gemini 2.5	Llama-4 Maverick	Qwen 2.5
Deadpan	Not Sarcastic	Not Sarcastic	Obnoxious	Polite	Not Sarcastic
Obnoxious	Not Sarcastic	Deadpan	Deadpan	Deadpan	Deadpan
Brooding	Obnoxious	Deadpan	Deadpan	Deadpan	Deadpan
Polite	Not Sarcastic	Deadpan	Deadpan	Deadpan	Not Sarcastic
Raging	Obnoxious	Deadpan	Obnoxious	Obnoxious	Obnoxious
Manic	Not Sarcastic	Deadpan	Obnoxious	Deadpan	Not Sarcastic
Self-deprecating	Not Sarcastic	Deadpan	Deadpan	Deadpan	Deadpan
Not Sarcastic	Obnoxious	Deadpan	Deadpan	Deadpan	Deadpan

Table 11: Most Frequent Misclassifications per Type using Zero-Shot Prompting

Type	GPT-4o	Claude 3.5 Sonnet	Gemini 2.5	Llama-4 Maverick	Qwen 2.5
Deadpan	Not Sarcastic	Not Sarcastic	Obnoxious	Polite	Not Sarcastic
Obnoxious	Deadpan	Deadpan	Deadpan	Deadpan	Deadpan
Brooding	Deadpan	Deadpan	Deadpan	Deadpan	Deadpan
Polite	Not Sarcastic	Not Sarcastic	Not Sarcastic	Deadpan	Not Sarcastic
Raging	Obnoxious	Deadpan	Obnoxious	Obnoxious	Obnoxious
Manic	Raging	Self-deprecating	Obnoxious	Obnoxious	Not Sarcastic
Self-deprecating	Deadpan	Not Sarcastic	Deadpan	Deadpan	Deadpan
Not Sarcastic	Obnoxious	Deadpan	Deadpan	Deadpan	Deadpan

Table 12: Most Frequent Misclassifications per Type using Few-Shot Prompting

Type	GPT-4o	Claude 3.5 Sonnet	Gemini 2.5	Llama-4 Maverick	Qwen 2.5
Deadpan	Not Sarcastic	Not Sarcastic	Not Sarcastic	Not Sarcastic	Not Sarcastic
Obnoxious	Deadpan	Not Sarcastic	Deadpan	Deadpan	Deadpan
Brooding	Deadpan	Not Sarcastic	Deadpan	Deadpan	Deadpan
Polite	Not Sarcastic	Not Sarcastic	Not Sarcastic	Deadpan	Not Sarcastic
Raging	Deadpan	Not Sarcastic	Obnoxious	Deadpan	Obnoxious
Manic	Brooding	Not Sarcastic	Not Sarcastic	Deadpan	Brooding
Self-deprecating	Not Sarcastic	Not Sarcastic	Not Sarcastic	Deadpan	Not Sarcastic
Not Sarcastic	Deadpan	Deadpan	Deadpan	Deadpan	Deadpan

Table 13: Most Frequent Misclassifications per Type using CoT Prompting

Type	GPT-4o	Claude 3.5 Sonnet	Gemini 2.5	Llama-4 Maverick	Qwen 2.5
Deadpan	Not Sarcastic	Not Sarcastic	Not Sarcastic	Obnoxious	Not Sarcastic
Obnoxious	Deadpan	Deadpan	Deadpan	Deadpan	Not Sarcastic
Brooding	Deadpan	Deadpan	Deadpan	Obnoxious	Not Sarcastic
Polite	Deadpan	Deadpan	Not Sarcastic	Not Sarcastic	Not Sarcastic
Raging	Brooding	Deadpan	Obnoxious	Obnoxious	Not Sarcastic
Manic	Polite	Not Sarcastic	Self-deprecating	Obnoxious	Not Sarcastic
Self-deprecating	Deadpan	Not Sarcastic	Not Sarcastic	Deadpan	Not Sarcastic
Not Sarcastic	Deadpan	Deadpan	Deadpan	Obnoxious	Deadpan

Table 14: Most Frequent Misclassifications per Type using Emotion-Based Prompting

Thank You!

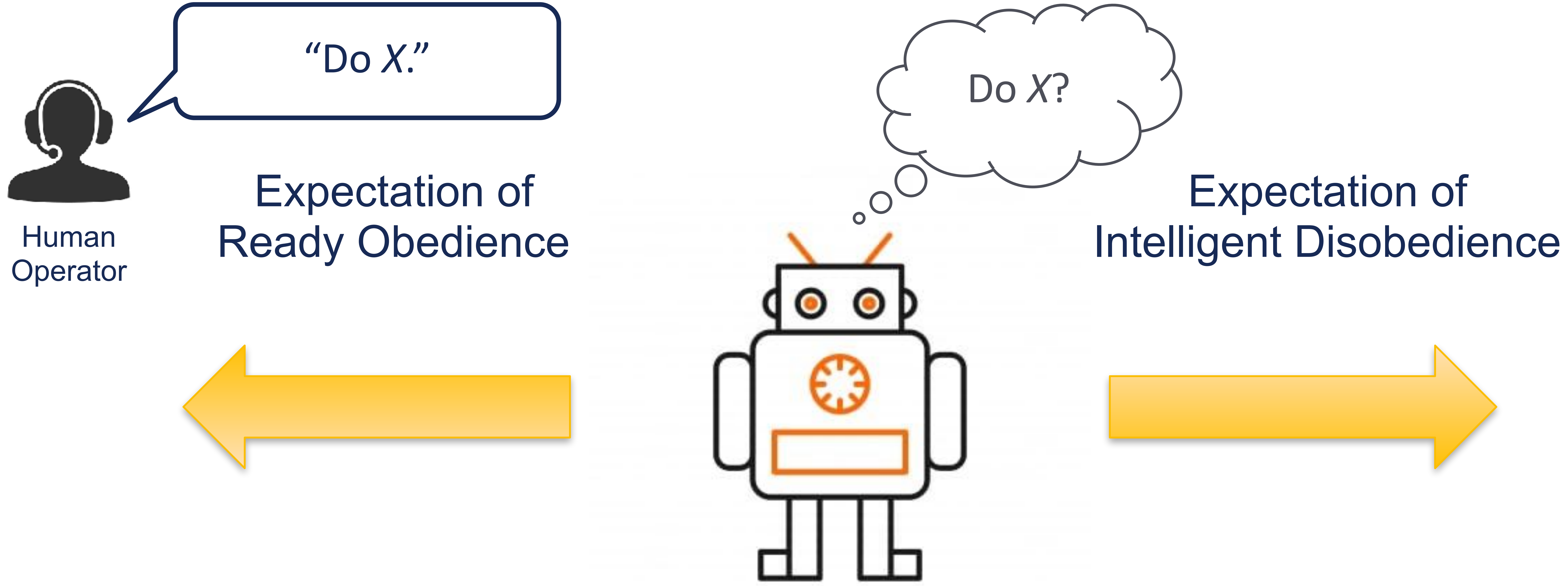
Constructive Disobedience and Trust in Human-Agent Interaction: A Multi-Scale Study

Gordon Briggs[1] and Christina Wasylyshyn[1]

[1] Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory, Washington, DC 20375 USA

Workshop on Optimal Reliance and Accountability in Interactions with Generative Language Models (ORIGen)
@COLM2025
Oct 10, 2025
Montreal, Canada

Introduction: Inherent Tension of AI/Robot Autonomy



Constructive disobedience: lower-level *disobedience* in service of higher-level *alignment* to superseding goals or norms

Trust and Constructive Disobedience

Trustworthy Disobedience Hypothesis:

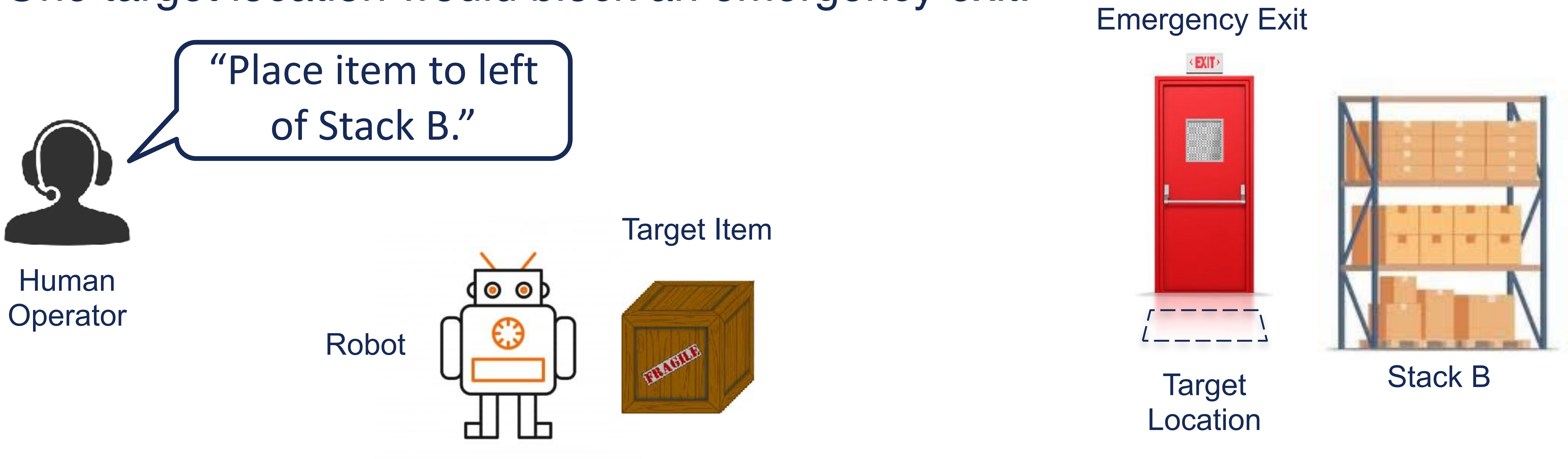
Autonomous agents that exhibit *constructive disobedience* are trusted more than robots that exhibit *strict obedience*

We conduct an experiment to test this hypothesis

Trust and Constructive Disobedience: Vignette Experiment

Warehouse Safety Vignette Scenario

Robots are being used to move loads to specific locations in a warehouse. One target location would block an emergency exit.



Vignette Experiment (cont.)

Two Conditions

Strict Obedience

Constructive Disobedience

Participants (N=60) assess robots in each condition (within-subjects design) using three trust scales:

- Multi-Dimensional Model of Trust (MDMT) (Malle and Ullman, 2021)
- Trust Perception Scale (TPS-HRI) (Schaeffer, 2016)
- Reliance Intention Scale (RIS) (Lyons and Guznov, 2019)

Results

Results **support** the Trustworthy Disobedience Hypothesis

Results consistent across all trust scales and measures

However, not all trust related items improve with constructive disobedience

Future work is needed to understand additional factors in that effect when constructive disobedience improves trust

Trustworthy Disobedience Hypothesis:
Robots that exhibit *constructive disobedience* are trusted more than robots that exhibit *strict obedience*

Check out the paper for more details!

COLORADO STATE UNIVERSITY

Let’s Roleplay: Examining LLM Alignment in Collaborative Dialogues

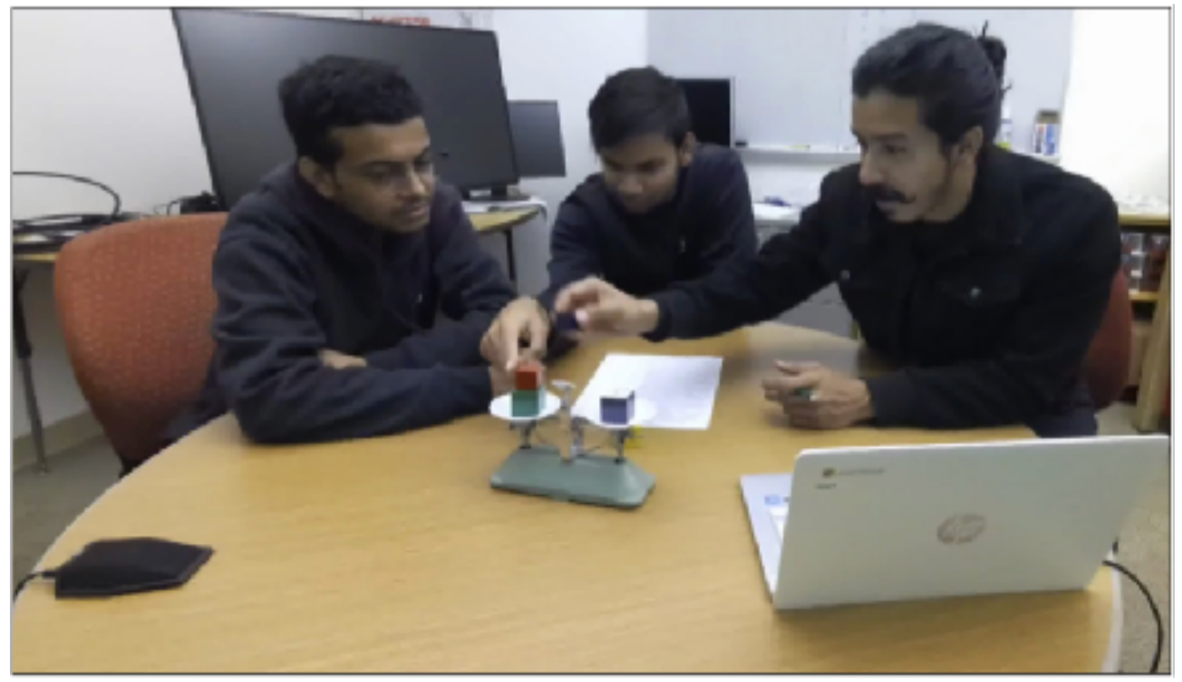
Abhijnan Nath, Carine Graff, and Nikhil Krishnaswamy

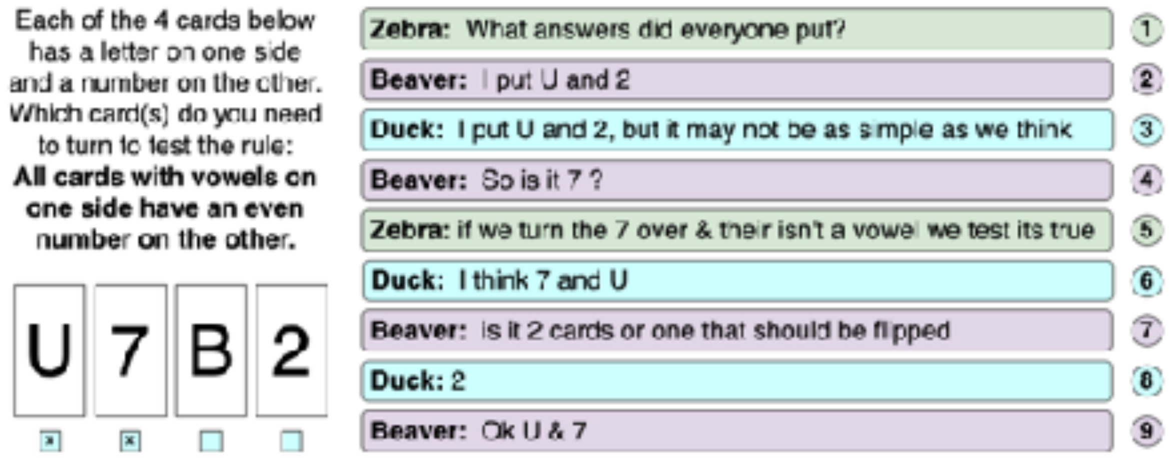
Situated Grounding and Natural Language (SIGNAL) Lab, Colorado State University

ORIGen @ COLM 2025

Introduction

- Collaborative groups of agents (human or AI) frequently succumb to belief misalignment and breakdown in common ground
  - This challenges optimality assumptions of LLM alignment algorithms
- **RQ**: *How do different LLM alignment methods perform in collaborative settings?*
- We explore this problem through the lens of ***friction interventions***
  - Prompting the dialogue participants to slow down, reflect and deliberate on their existing assumptions
- Use a *roleplay* methodology to examine LLM behavior in multiparty collaborative settings
- Examine how LLM alignment techniques contribute to construction of common ground and task success

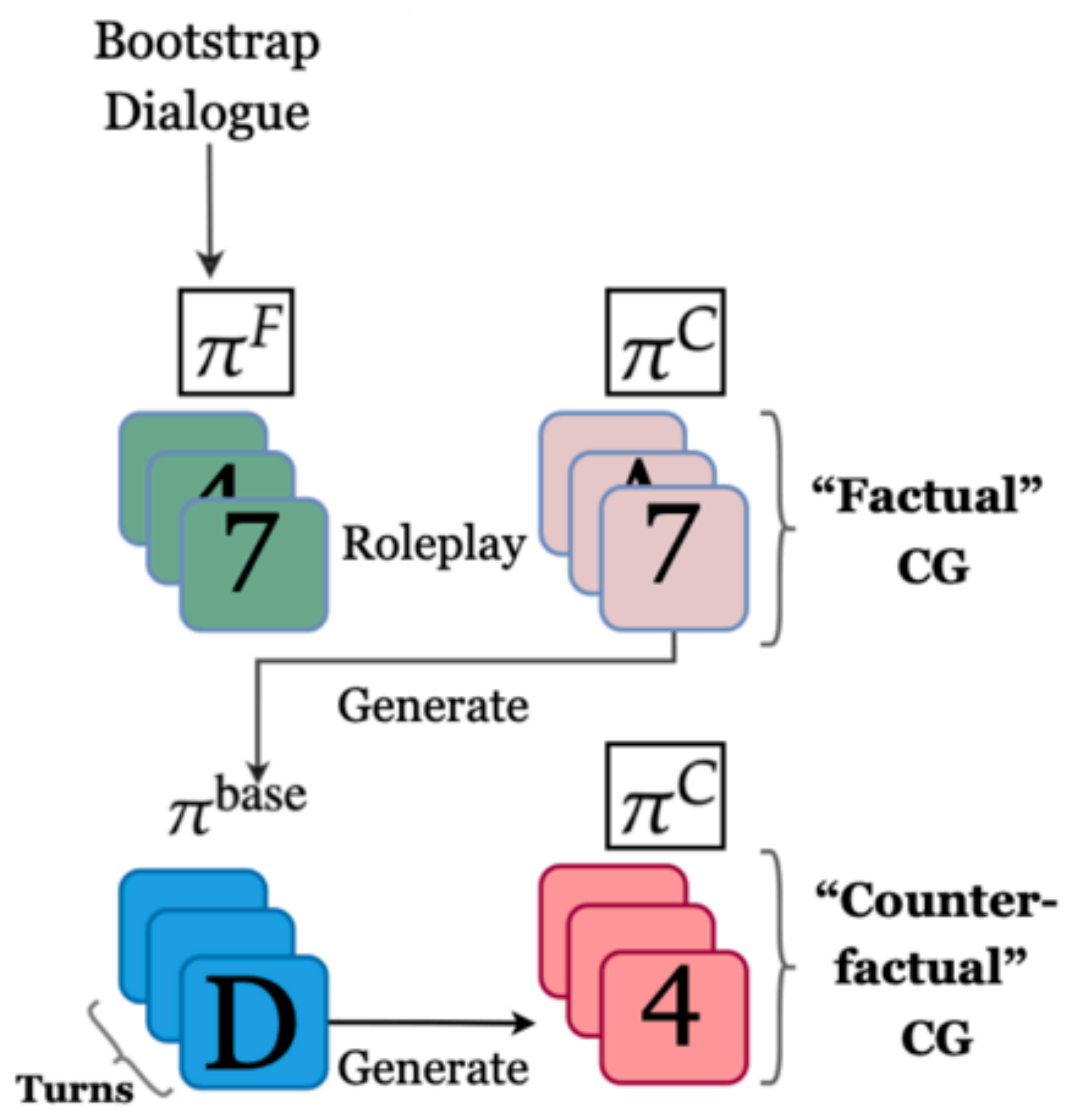




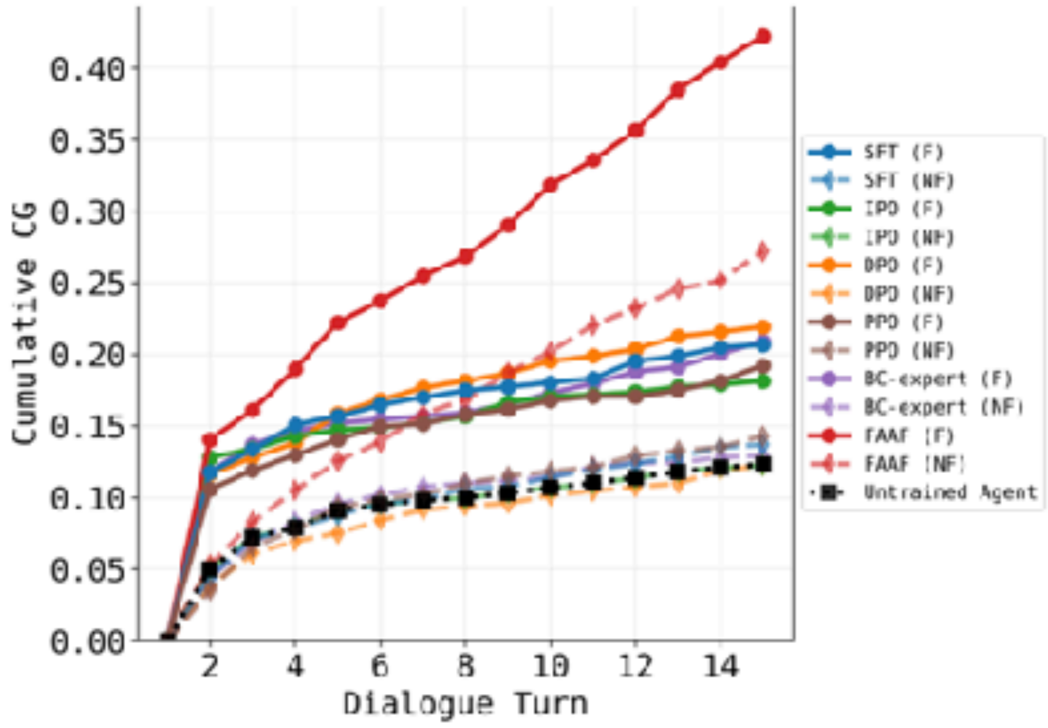
2 collaborative tasks:
Wason Card Task and Weights Task

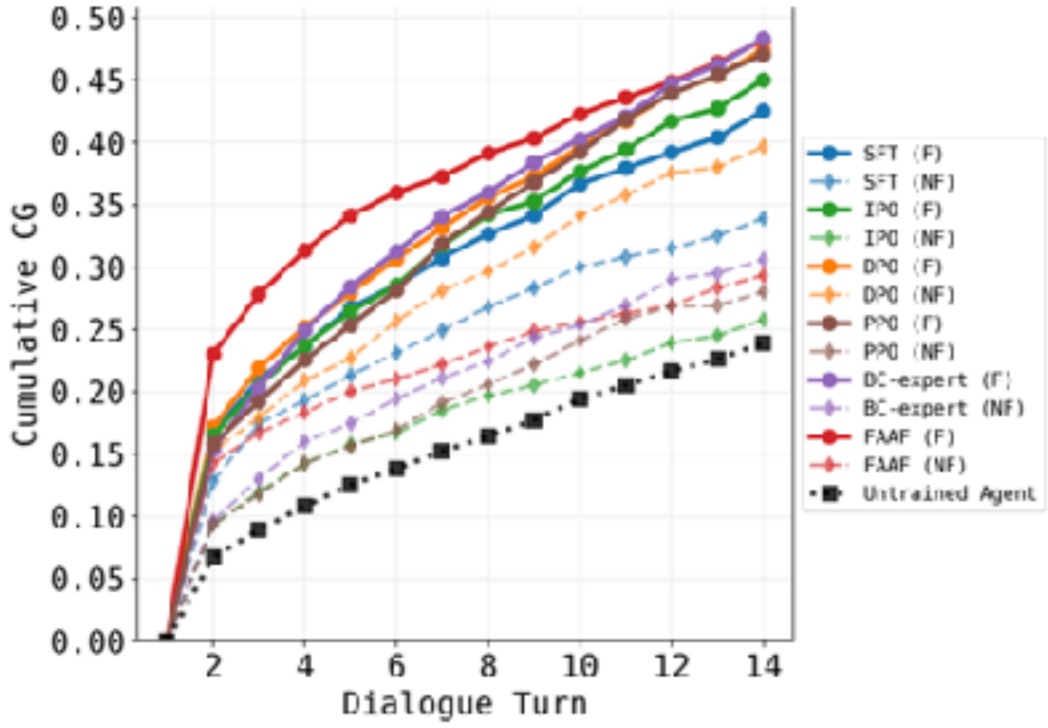
Methodology

- Counterfactual role-play methodology:
  - **Step 1**: Collect *factual* trajectories where aligned intervention agent (SFT, PPO, DPO, IPO, BC, FAAF) interacts with roleplayed collaborators (GPT-4o)
  - **Step 2**: Use *untrained* instruction-tuned agent to generate alternate interventions in response to collaborator outputs, resulting in *counterfactual* trajectory
  - **Step 3**: Run new dialogue loop to collect fresh collaborator responses to cached untrained responses, resulting in *counterfactual* trajectory in which collaborators receive interventions from untrained agent
  - Compare the group common ground (set of shared beliefs) and task success of factual vs. counterfactual trajectories



Results





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	$14.82_{\pm 0.13}$	$10.10_{\pm 0.11}$	$0.54_{\pm 0.05}$	$0.80_{\pm 0.02}$	$0.37_{\pm 0.04}$	$0.72_{\pm 0.02}$
$\text{FAAF}_{\Delta R'}$	$9.03_{\pm 0.10}$	$7.56_{\pm 0.08}$	$0.39_{\pm 0.05}$	$0.79_{\pm 0.02}$	$0.30_{\pm 0.05}$	$0.62_{\pm 0.02}$
FAAF	$\mathbf{14.91}_{\pm 0.14}$	$\mathbf{14.16}_{\pm 0.13}$	$\mathbf{0.60}_{\pm 0.05}$	$\mathbf{0.87}_{\pm 0.02}$	$\mathbf{0.45}_{\pm 0.05}$	$\mathbf{0.80}_{\pm 0.02}$

What do these numbers mean?

Come see our poster to find out!

Thank you!

Uncertainty Quantification in Retrieval Augmented Question Answering

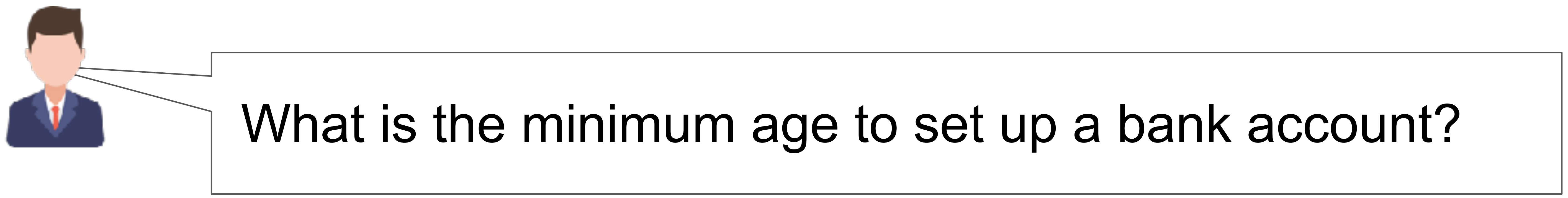
Laura Perez-Beltrachini and Mirella Lapata

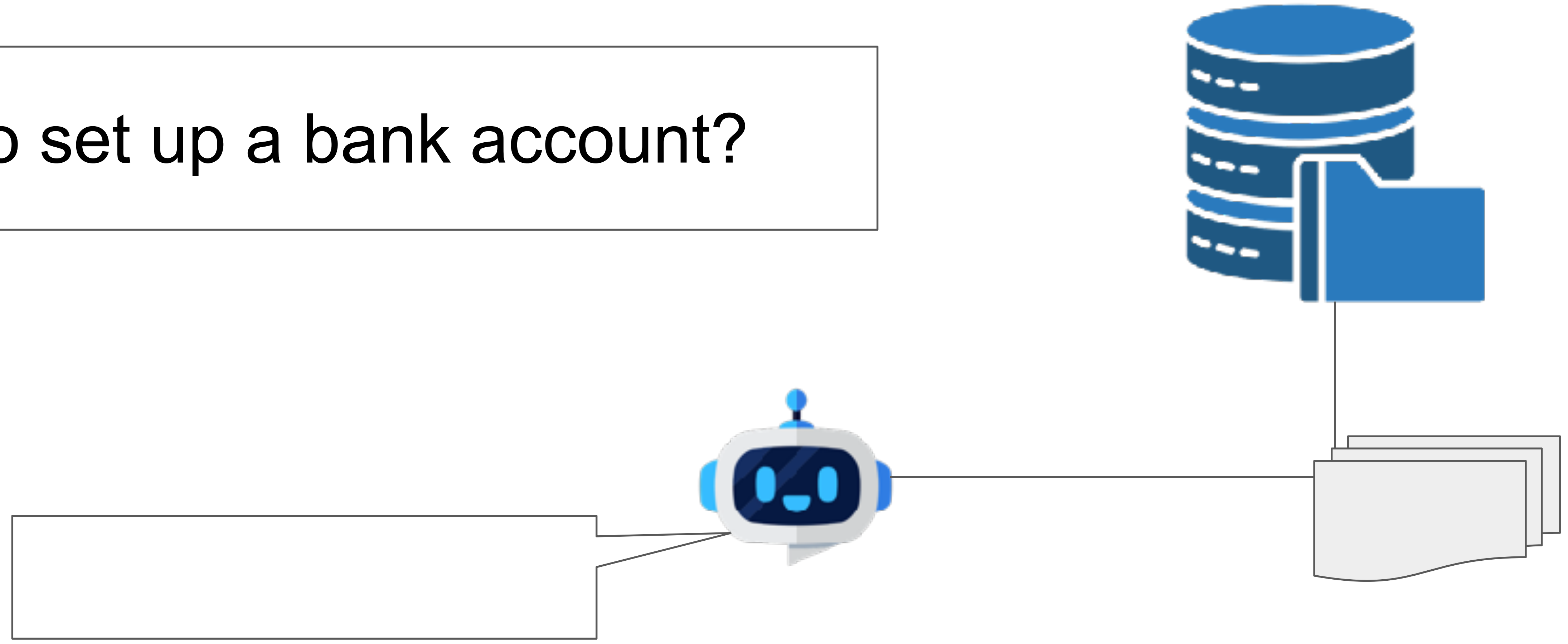
ORIGen Workshop
COLM 2025
10th October 2025

THE UNIVERSITY of EDINBURGH

Edinburgh NLP
University of Edinburgh
Natural Language Processing

Retrieval Augmented Question Answering

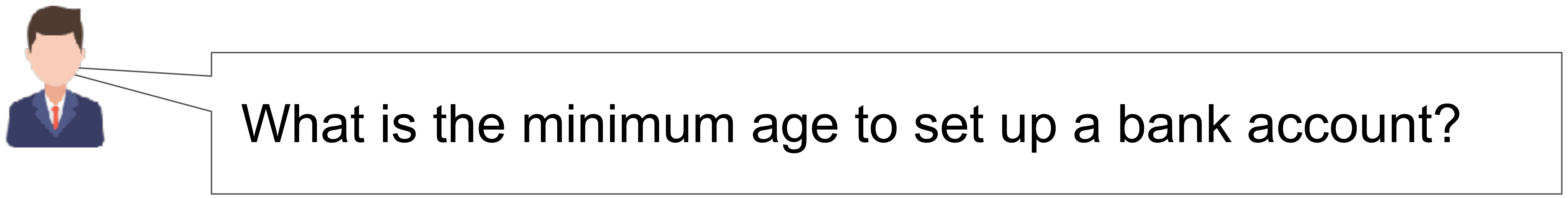


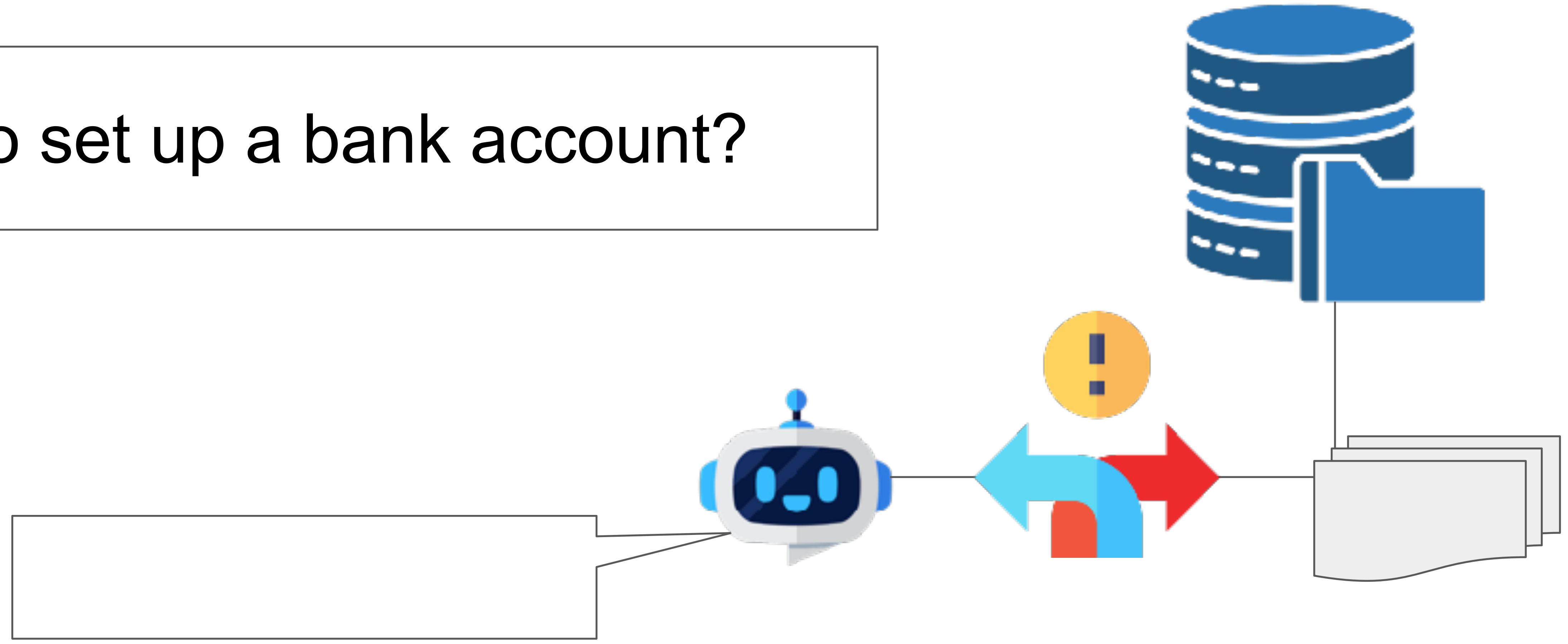


Benefits

Mallen et al., 2023; Shuster et al., 2021; Jiang et al., 2021; and many more

Retrieval Augmented Question Answering





Benefits

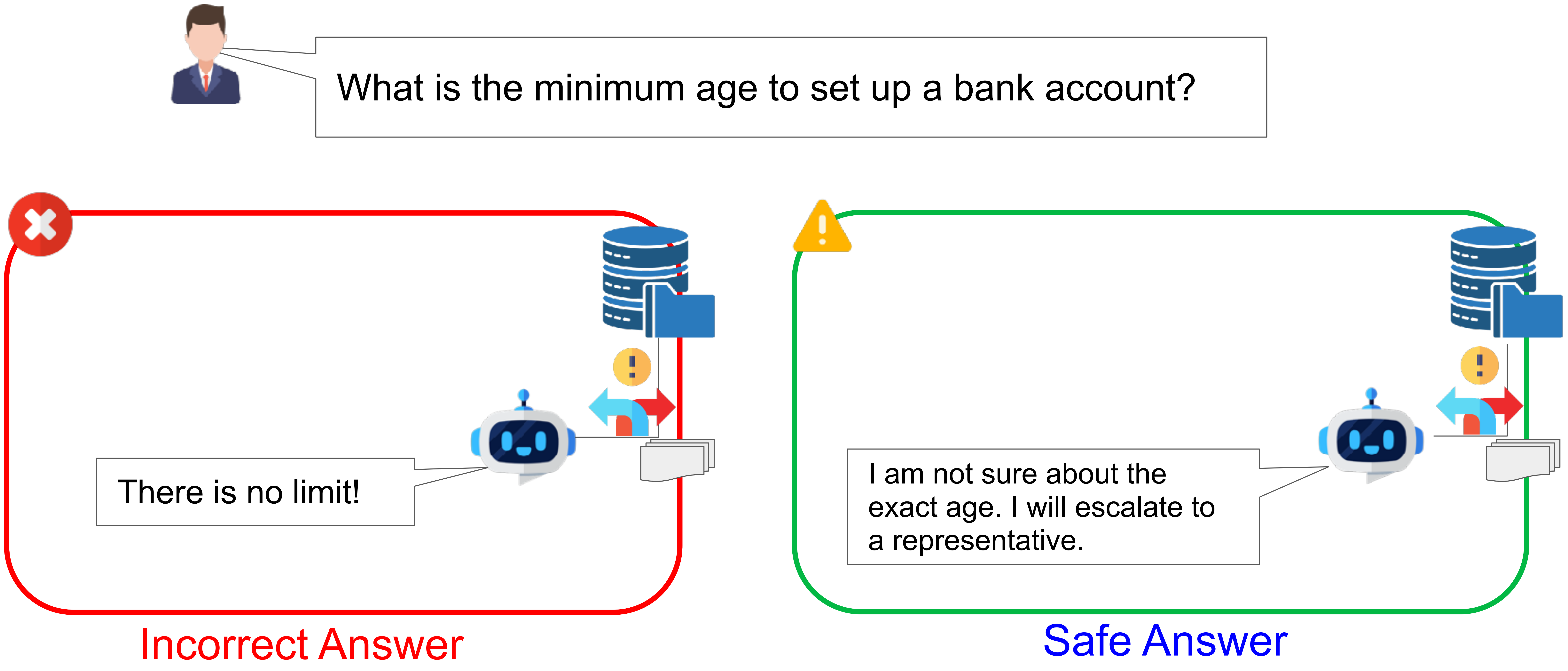
Mallen et al., 2023; Shuster et al., 2021; Jiang et al., 2021; and many more

Insufficient knowledge

- Irrelevant, incomplete, or misleading evidence
- Fail to reason over evidence and question
- Ingnore retrieved evidence
- Unanswerable questions

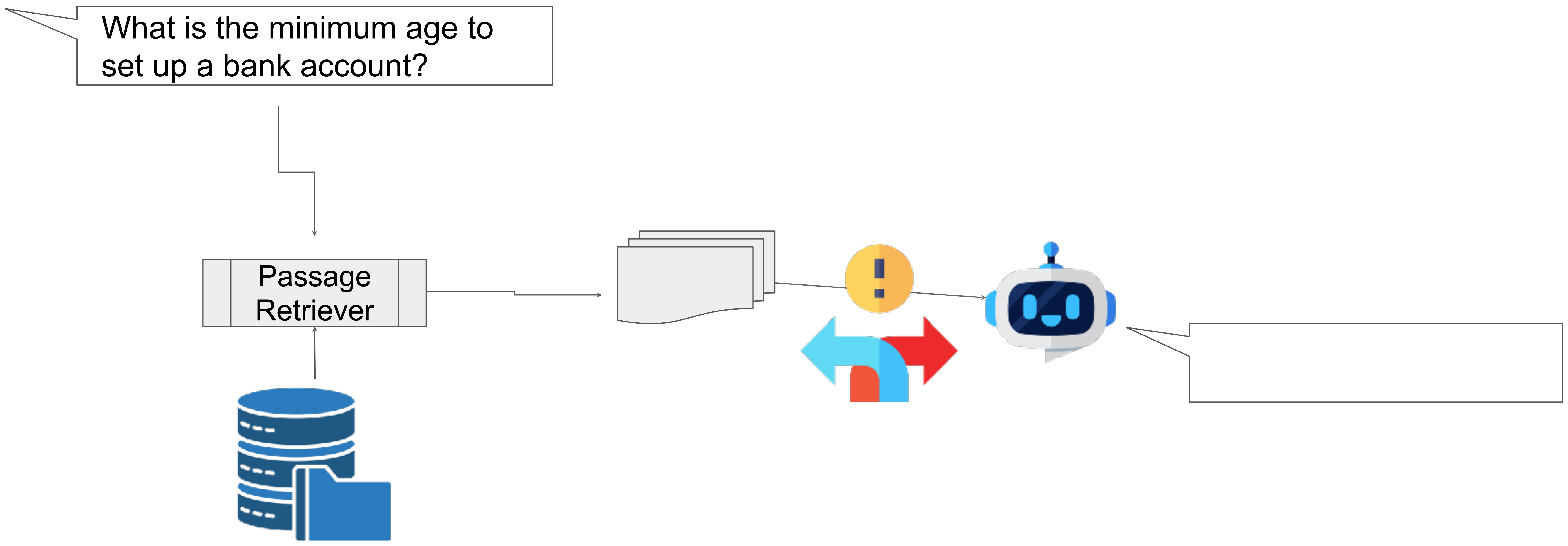
Sciavolino et al., 2021; Yoran et al., 2024; Kasai et al., 2024; Izacard et al., 2024; Liu et al., 2024b; Sun et al., 2025; Xie et al., 2024; Joren et al., 2025

Answer Uncertainty Estimation for
QA Trustworthiness



Humans overrely on overconfident language models, across languages (Rathi et al., COLM 2025)

Answer Uncertainty Estimation?



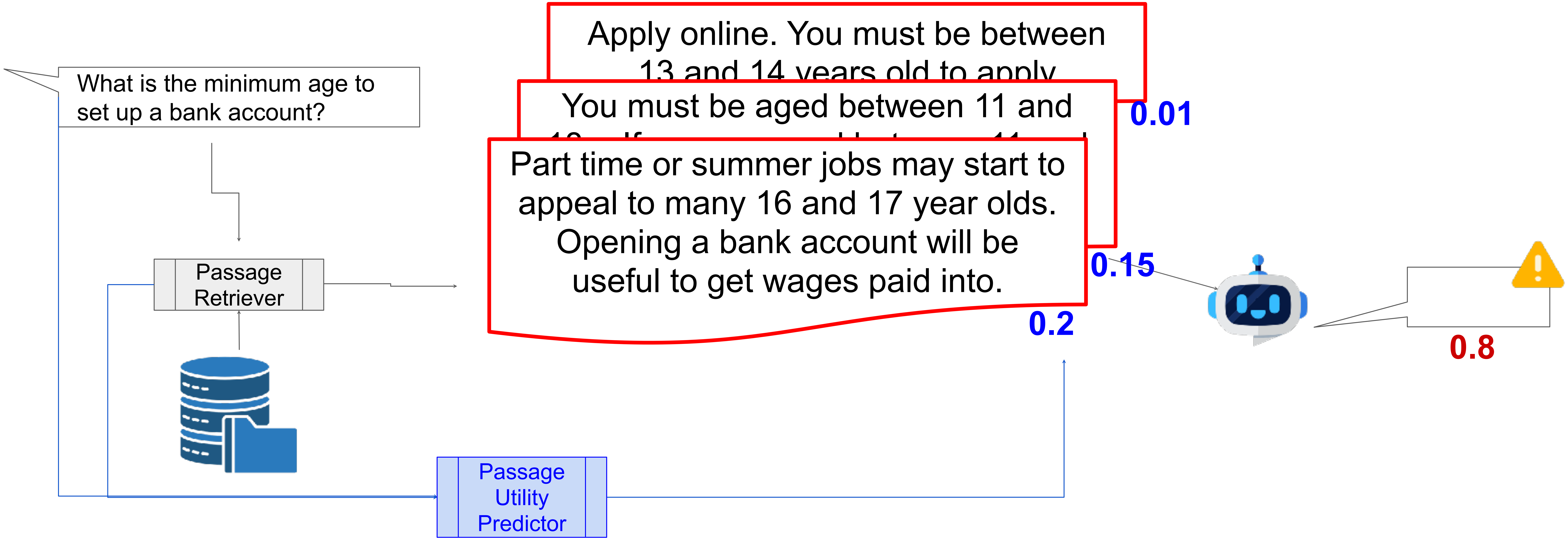
Existing Uncertainty Estimation Approaches

	Low Latency	Recognises Erroneous Evidence	Mitigates Overconfidence
PPL	✓		
Regular Entropy			
Semantic Entropy			
p(true)		✓	✓

+ Previously evaluated on closed-book QA

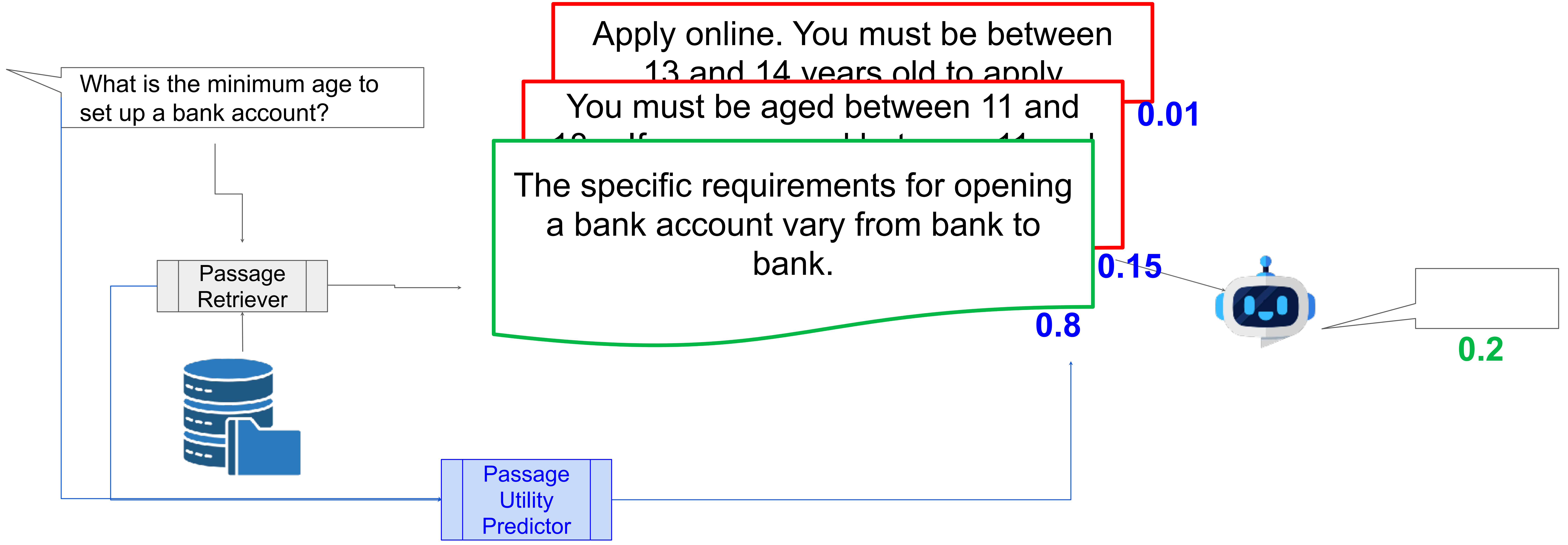
Farquhar et al., 2024; Simhi et al., 2025; Soudani et al., 2025; Sung et al., 2025

Answer Uncertainty via Passage Utility



- Predict Passage Utility, Estimate Answer Uncertainty

Answer Uncertainty via Passage Utility

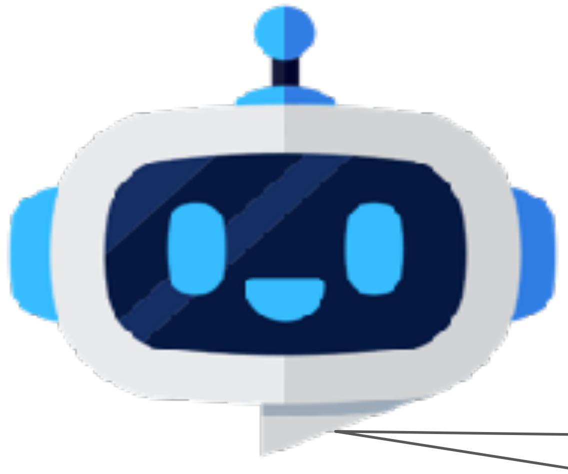


+ Predict Passage Utility, Estimate Answer Uncertainty

Existing Uncertainty Estimation Approaches

	Low Latency	Recognises Erroneous Evidence	Mitigates Overconfidence
PPL	✓		
Regular Entropy			
Semantic Entropy			
p(true)		✓	✓
Passage Utility	✓	✓	✓

Farquhar et al., 2024; Simhi et al., 2025; Soudani et al., 2025; Sung et al., 2025



See you at our poster!

We thank Roundicons, Luch Phou, littleicon, Freepik, Aranagraphics, HAJICON,
Good Ware from the Flaticon library.