Exploring Trustworthy Foundation Models: Benchmarking, Finetuning, and Reasoning

Prof. Bo Han
HKBU TMLR Group / RIKEN AIP Team
Associate Professor / BAIHO Visiting Scientist
https://bhanml.github.io

Trustworthy Foundation Models

Benchmarking

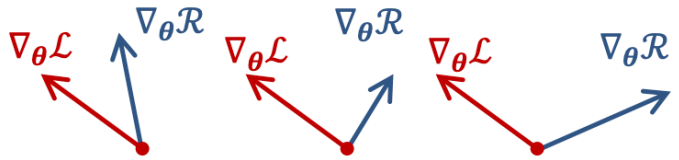
Existing datasets are NOT proper to assess if **VLMs** are robust.

CounterAnimal, a reliable benchmark for assessing VLMs.

- **Scaling backbone models** and **improving data quality** improve the robustness of VLMs.
- **Scaling raw training data** does not necessary enhance reliability.

Finetuning

Analyzing the dynamics of **LLMs** unlearning is critical yet hard.



- **Analyzing gradients** provides insights into unlearning dynamics.
- **Wrong token reweighting** within gradients leads to failures in previous methods.

Reasoning

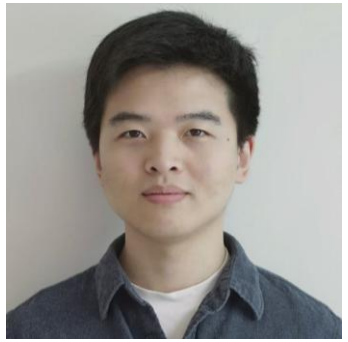
Existing **LLMs** are passive chatbots rather than active reasoners.

AR-Bench, a benchmark to evaluate LLMs' ability to ask the right questions.

- Existing LLMs and methods exhibit **a significant gap** compared to humans in active reasoning.
- LLMs struggle to consistently ask **high-quality questions**.

Part I: Benchmarking

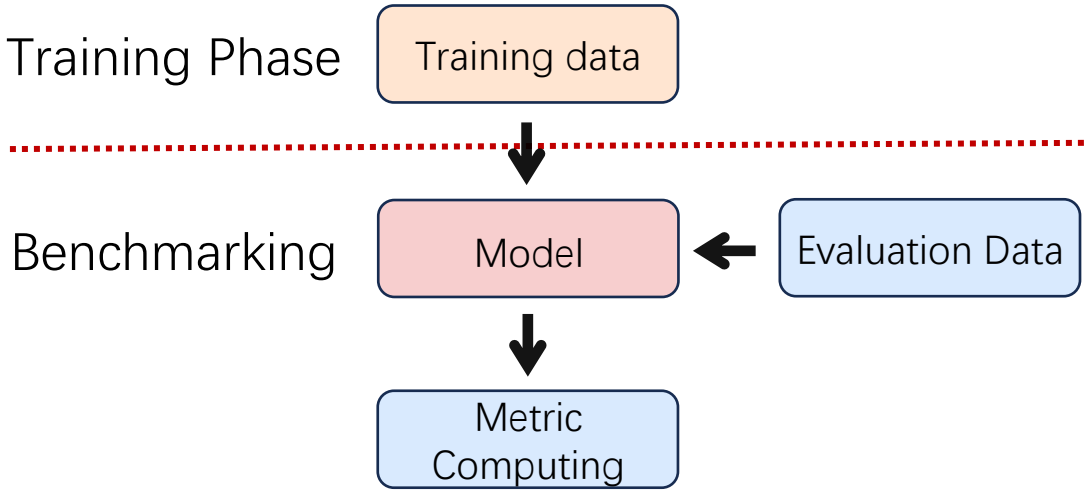
Qizhou Wang



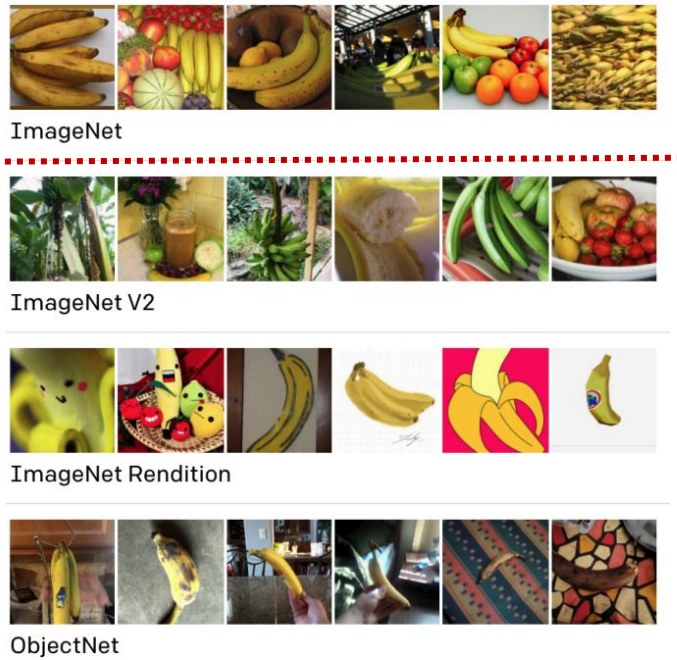
Yongqiang Chen

Benchmarking is critical to evaluate and compare model quality.

- Gathering **reliable evaluation data.**
- Conducting **proper metric evaluations.**



Training and evaluation data have **distribution shifts** to reflect **OOD Generalization**.

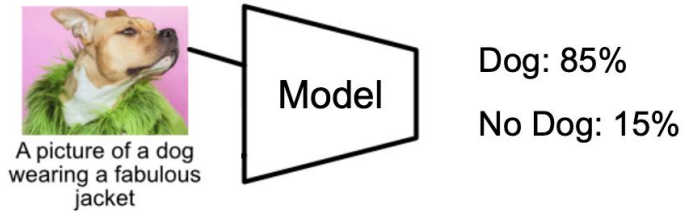


Qizhou Wang, Yong Lin, **Yongqiang Chen**, Ludwig Schmidt, **Bo Han**, and Tong Zhang. A Sober Look at the Robustness of CLIPs to Spurious Features. In *NeurIPS*, 2024.

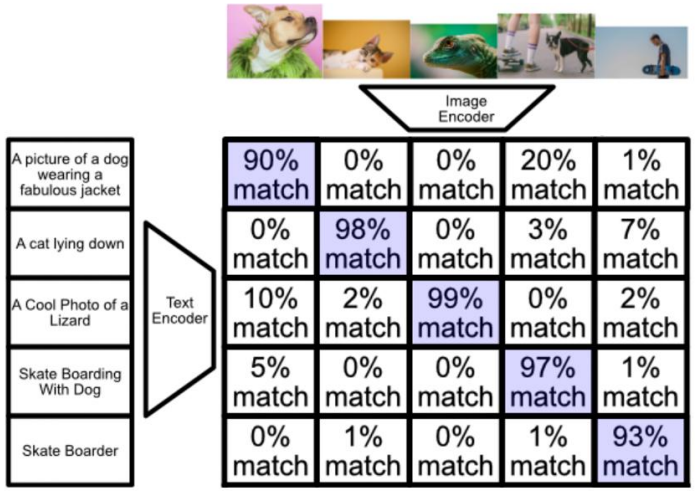
https://bhanml.github.io & https://github.com/tmlr-group

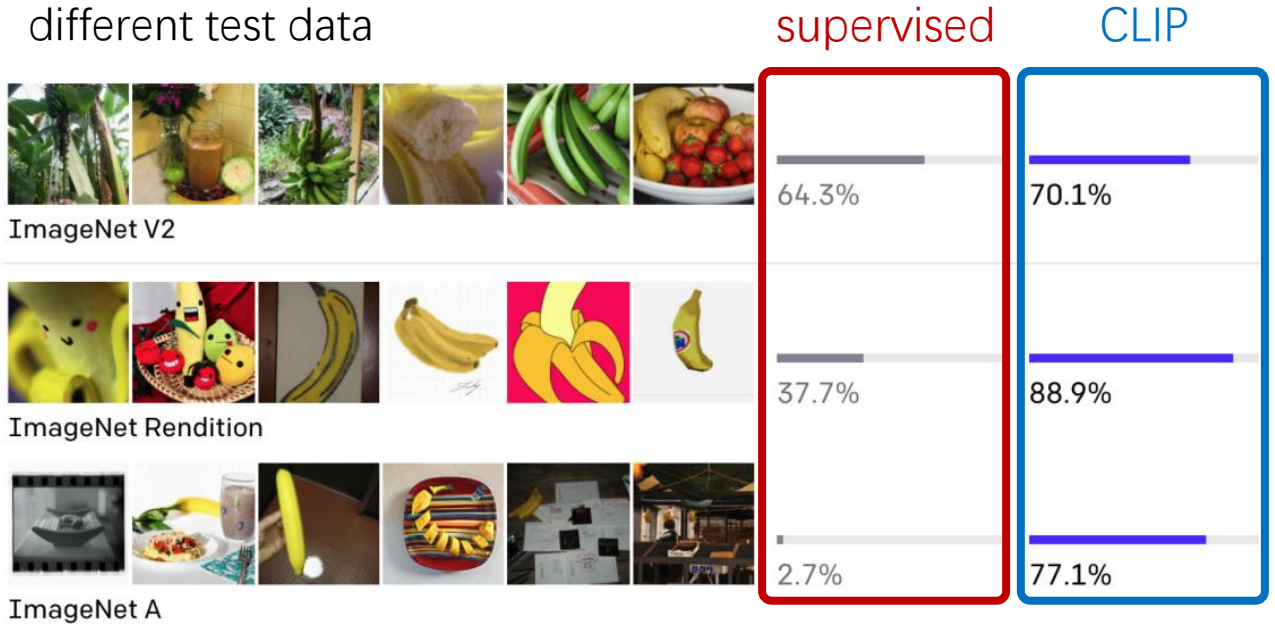
Supervised vs CLIP Training

Supervised Training *label supervision*



CLIP Training *cross-modal supervision*





*Comparison of the **OOD evaluation accuracy** between supervised and CLIP training shows that **CLIP performs better!***

Previous Belief: CLIP is **more robust to distribution shifts** than conventional supervised training.

(Radford et al., 2021)

Is the Conclusion Correct?

These OOD datasets are crafted for the distribution shifts **within ImageNet setups**, which are **NOT valid for CLIP models.**

- **Data Contamination:** Datasets considered **OOD for ImageNet-trained models** may be **ID for CLIP models.**

ImageNet V2

CLIP models may have seen ImageNet V2 during training, which is in fact ID for CLIP setups.

- **Biased Spuriousness:** Features that **mislead ImageNet-trained models** may **not mislead CLIP models** necessarily.

ImageNet A

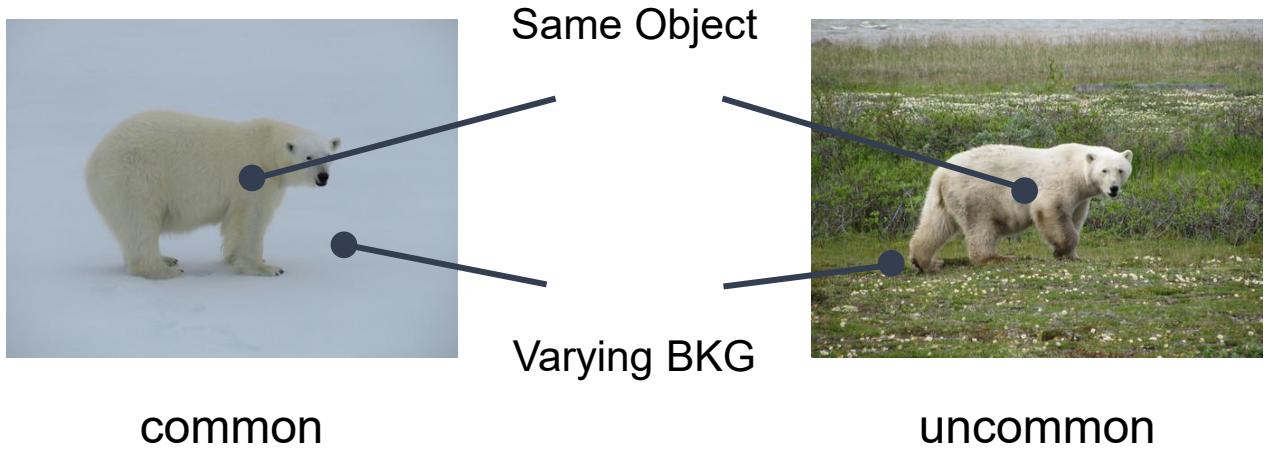
ImageNet A contains data that mislead ImageNet models, which may not make CLIP models fail.

ImageNet OOD datasets CANNOT reflect the OOD Generalization for CLIP setups!

CounterAnimal: A New Benchmark

Is there a benchmark capturing **true OOD performance** of CLIP?

- **Spuriousness:** Considering **background changes** as potential spurious features.
- **Generality:** The captured spurious features should impact **diverse CLIP configurations**.



The changes of backgrounds represent the impacts of spurious features, which is a typical distribution shift.

Basic Assumption: Since "ice bears" are more commonly appear with "ice" rather than "grass" backgrounds, CLIP may rely on ice-related spurious features.

CounterAnimal Construction

Step 1. Data Collection

Raw data from iNaturalist (https://www.inaturalist.org)

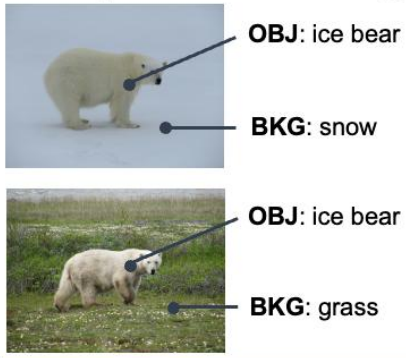
query and crawl **animal photos** given the names

Step 2. Data Curation

Raw data are susceptible to **noise** and **ambiguities**, which should be **cleansed manually**.

clean | noise | occlusion | obscurity

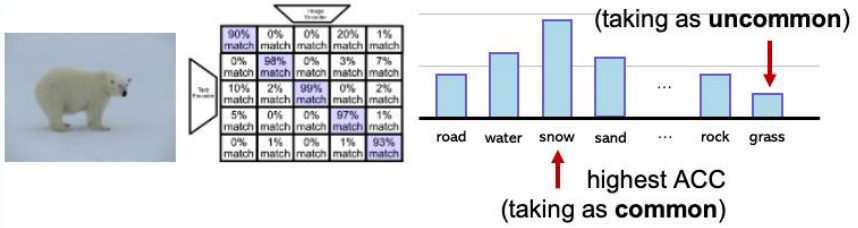
Step 3. Data Labelling



OBJ labels: *ostrich, African crocodile, water snake, ice bear*, and other totally **45 animal names**.

BKG labels: *ground, water, earth*, and other totally **16 background labels**.

Step 4. Spurious Discovery



The pair of backgrounds where the CLIP shows high-performance drops are preserved.

CounterAnimal Construction

Step 1. Data Collection

Raw data from iNaturalist (https://www.inaturalist.org)

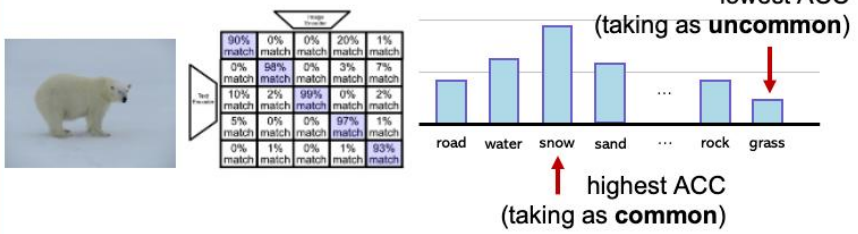
query and crawl **animal photos** given the names

Step 2. Data Curation

Raw data are susceptible to **noise** and **ambiguities**, which should be **cleansed manually**.

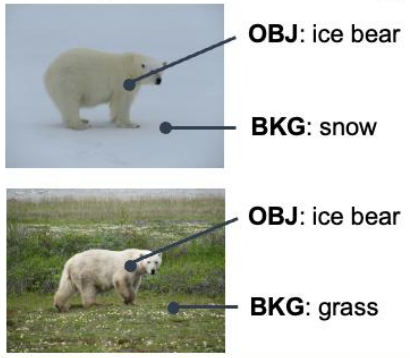
clean noise occlusion obscurity

Step 4. Spurious Discovery



The pair of backgrounds where the CLIP shows high-performance drops are preserved.

Step 3. Data Labelling



OBJ labels: *ostrich, African crocodile, water snake, ice bear,* and other totally **45 animal names**.

BKG labels: *ground, water, earth*, and other totally **16 background labels**.

CounterAnimal Construction

Step 1. Data Collection

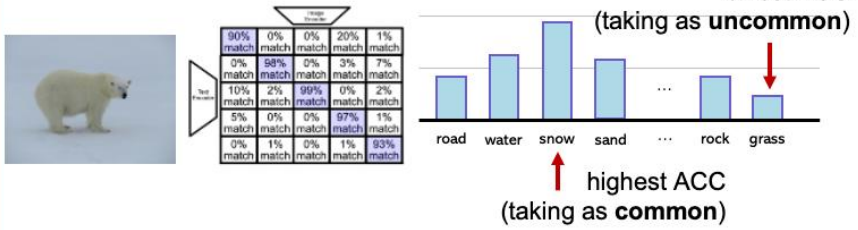
Raw data from iNaturalist (https://www.inaturalist.org)

Step 2. Data Curation

Raw data are susceptible to **noise** and **ambiguities**, which should be **cleansed manually**.

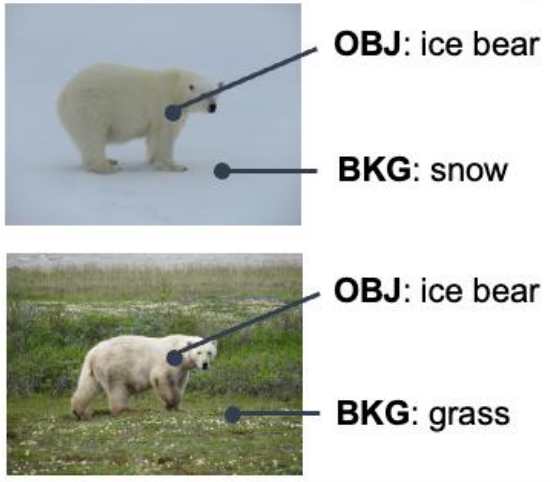
clean | noise | occlusion | obscurity

Step 4. Spurious Discovery



The pair of backgrounds where the CLIP shows high-performance drops are preserved.

Step 3. Data Labelling



OBJ labels: *ostrich, African crocodile, water snake, ice bear*, and other totally **45 animal names**.

BKG labels: *ground, water, earth*, and other totally **16 background labels**.

CounterAnimal Construction

Step 1. Data Collection

Raw data from iNaturalist (https://www.inaturalist.org)

Step 2. Data Curation

Raw data are susceptible to **noise** and **ambiguities**, which should be **cleansed manually**.

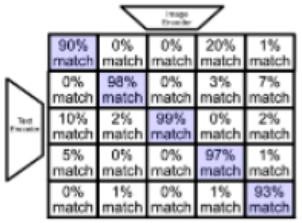
clean noise occlusion obscurity

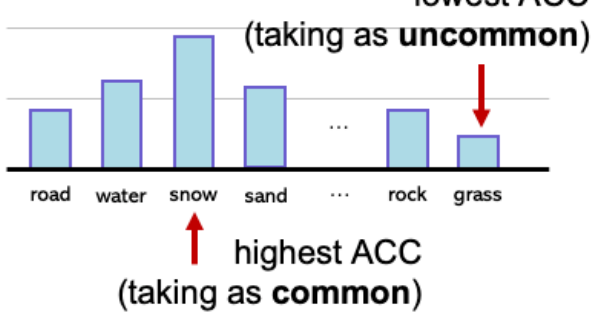
Step 3. Data Labelling

OBJ labels: *ostrich, African crocodile, water snake, ice bear,* and other totally **45 animal names**.

BKG labels: *ground, water, earth*, and other totally **16 background labels**.

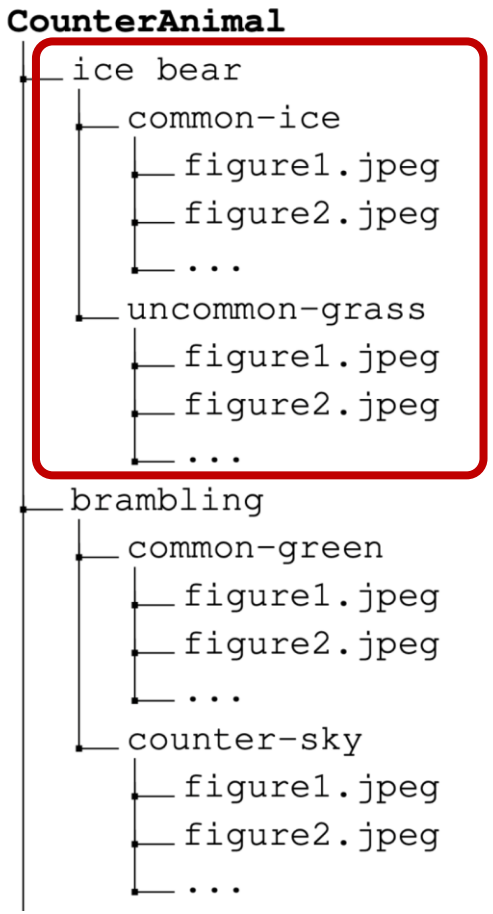
Step 4. Spurious Discovery





The pair of backgrounds where the CLIP shows high-performance drops are preserved.

CounterAnimal Characteristics



Data Structure. *Images are organized per class and each further divided into two groups: common and uncommon.*

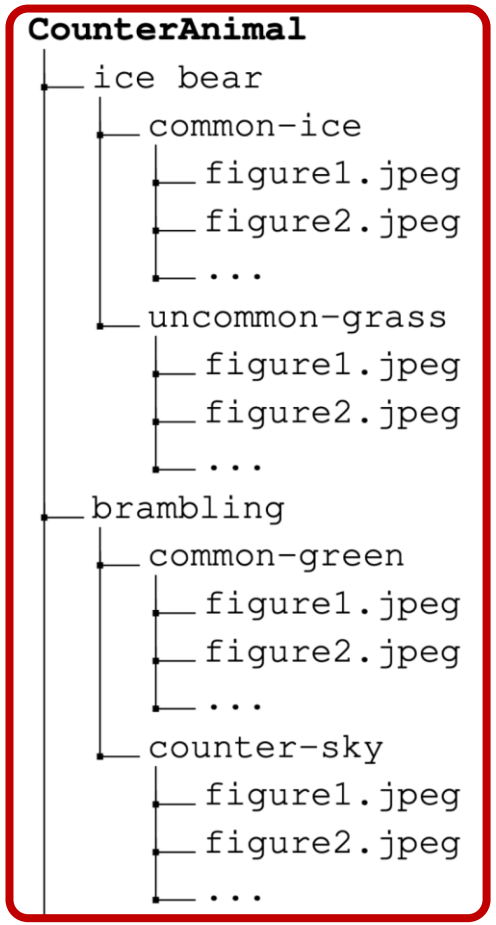
*Photos of **ice bear** in **snow** background*

*Photos of **ice bear** in **grass** background*

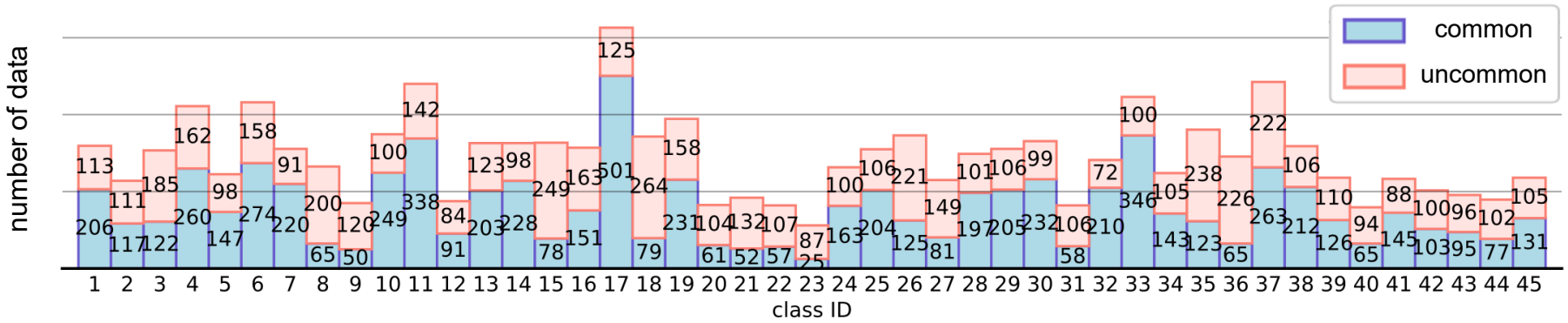
Common vs. Uncommon: Photos are grouped according to their backgrounds. For each class, we identify **group pairs** that cause **high performance drop** when evaluating with CLIP.

Assessing Robustness: The **performance drop** between common and uncommon groups indicates the robustness of evaluated models.

CounterAnimal Characteristics



Data Structure. *Images are organized per class and each further divided into two groups: common and uncommon.*



*The **data distributions** illustrate variations across different animal classes, categorized into **common** and **uncommon** groups. The horizontal axis denotes the **class IDs**, e.g., ID 1 to "ostrich", ID 2: to "brambling", ···, ID 8 to "box turtle", ID 9 to "common iguana",···, ID 18 to "scorpion", ID 19 to "tarantula", ···, ID 32 to "African hunting dog", ID 33 to "hyena", ···.*

We collect **45 classes** of animals with **7,000 common** and **6,000 uncommon** examples.

Experimental Results

common acc – uncommon acc

CLIP Training | CounterAnimal | (ImageNet) Supervised Training | Other LVLMs (large VLMs)

backbone	pre-train dataset	common	uncommon	drop
RN-101	OpenAI	64.27	45.15	19.12
RN-50×4	OpenAI	70.02	49.07	20.95
ViT-B/16	LAION400M	73.11	52.17	20.94
ViT-B/16	OpenAI	73.08	56.56	16.52
ViT-B/16	DataComp1B*	80.36	64.24	16.12
ViT-B/16	LAION2B	73.18	53.18	20.00
ViT-B/16	DFN2B*	85.03	70.61	14.42
ViT-B/32	LAION400M	67.13	36.95	30.18
ViT-B/32	OpenAI	69.13	45.62	23.51
ViT-B/32	DataComp1B*	75.96	53.74	22.22
ViT-B/32	LAION2B	72.94	48.74	24.20
ViT-L/14	LAION400M	80.90	63.31	17.59
ViT-L/14	OpenAI	85.38	70.28	15.10
ViT-L/14	DataComp1B*	89.29	79.90	9.39
ViT-L/14	LAION2B	82.23	66.27	15.96
ViT-L/14	DFN2B*	90.77	80.55	10.22
ViT-L/14-336	OpenAI	86.36	73.14	13.21
ViT-H/14	LAION2B	85.74	73.13	12.61
ViT-H/14	DFN5B*	88.55	79.13	9.42
ViT-G/14	LAION2B	86.81	73.32	13.49
ViT-bigG/14	LAION2B	87.57	76.96	10.61

increasing model scale

diverse data source

backbone	common	uncommon	drop
AlexNet	59.56	39.24	20.31
VGG-11	73.37	56.12	17.25
VGG-13	75.33	58.43	16.90
VGG-19	77.84	61.74	16.10
RN-18	74.36	56.07	18.29
RN-34	78.31	61.01	17.30
RN-50	81.44	66.07	15.37
RN-101	81.76	68.18	13.57
ViT-B/16	84.97	74.98	9.99
ViT-B/32	79.84	64.36	15.48
ViT-L/16	83.74	72.69	11.05
ViT-L/32	81.23	67.54	13.69
ConvNext-S	88.27	79.97	8.30
ConvNext-B	88.60	80.53	8.07
ConvNext-L	89.12	81.47	7.65

increasing model scale

LVLMs	common	uncommon	drop
MiniGPT4-Viccuna7B	47.99	39.73	8.26
LLaVA1.5-7B	40.06	30.09	9.97
CLIP-LAION400M-ViT-L/14	80.90	63.31	17.59
CLIP-OpenAI-ViT-L/14	85.38	70.28	15.10
CLIP-DataComp1B-ViT-L/14	89.29	79.90	9.39
CLIP-LAION2B-ViT-L/14	82.23	66.27	15.96
CLIP-DFN2B-ViT-L/14	90.77	80.55	10.22

different LVLM paradigms

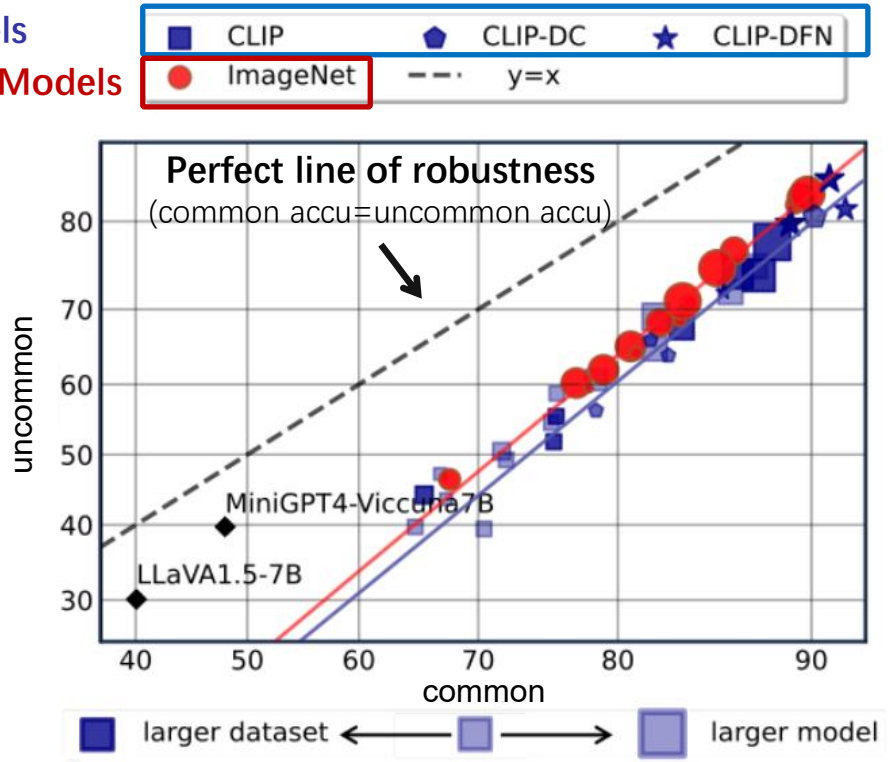
What observations can we draw from these results?

Observations

DataComp (DC) and Data Filtering Networks (DFN) are two **high-quality CLIP data sources.**

CLIP Models

ImageNet Models



*The **marker size** indicates the **backbone scale**, and the **color shade** indicates **pre-train data scale**.*

Observation 1 (ImageNet Models vs. CLIPs).

ImageNet models perform better than CLIPs against spuriousness within CounterAnimal.

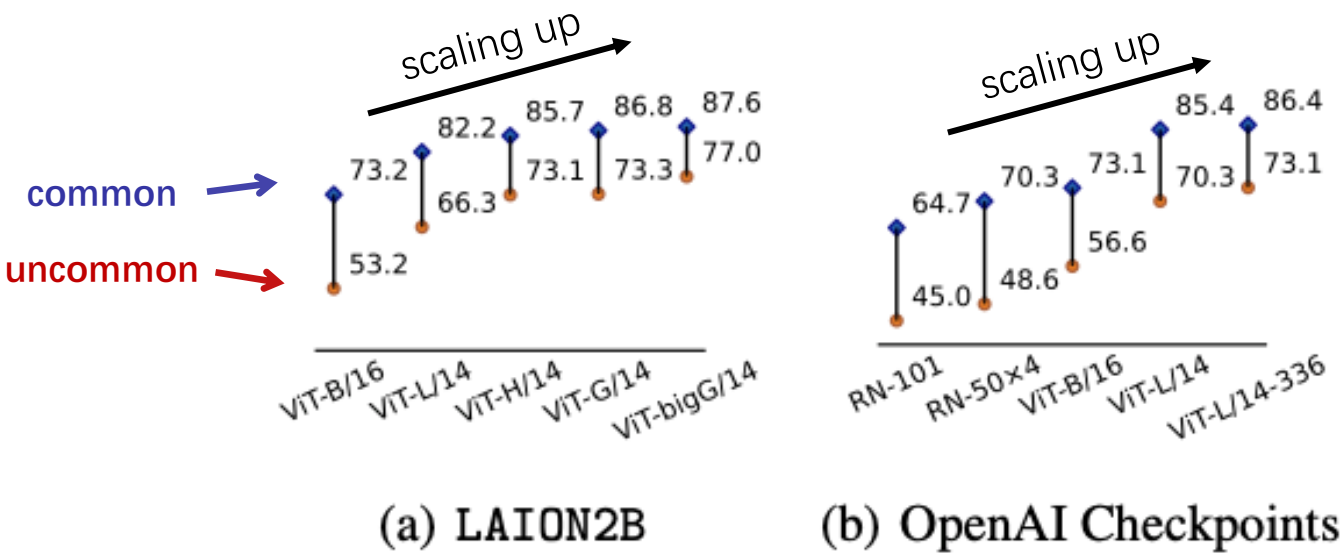
Note. CounterAnimal characterizes the spuriousness within CLIPs, thus proper for assessing CLIPs.

Observation 2 (CLIPs vs. More Advanced LVLMs).

LLaVA and MinGPT4 show **stronger robustness** (closer to $y = x$) yet with **lower performance** than CLIPs.

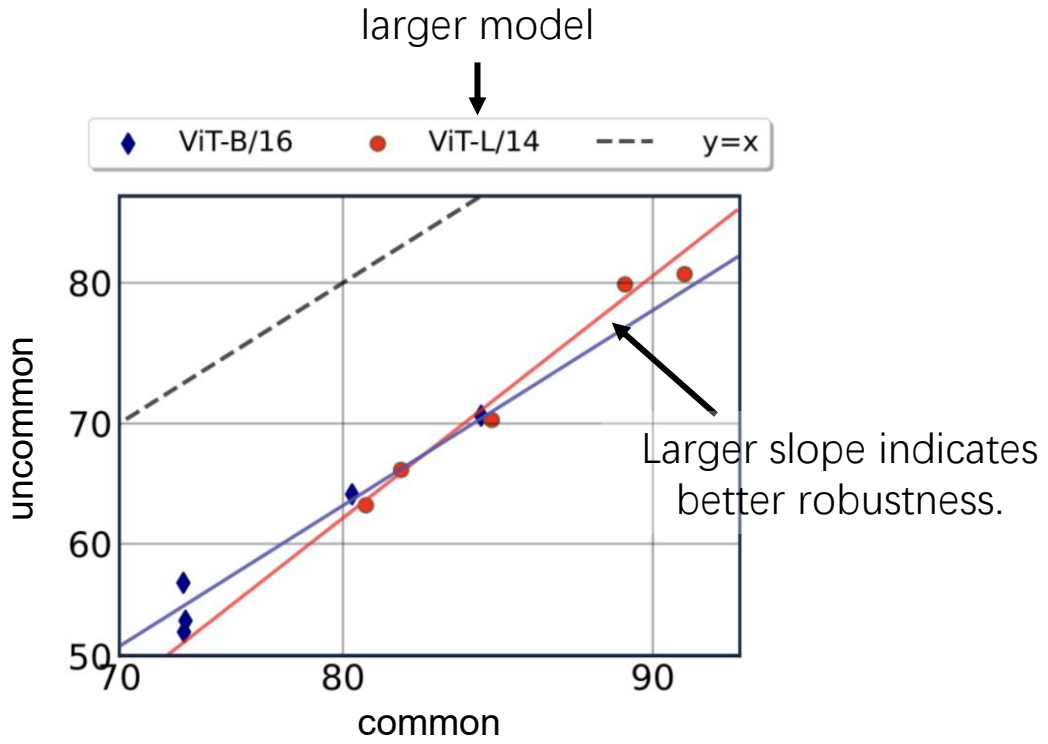
Note. More advanced VLMs built upon CLIPs are still affected by spuriousness within CounterAnimal.

Observations



(a) LAION2B (b) OpenAI Checkpoints

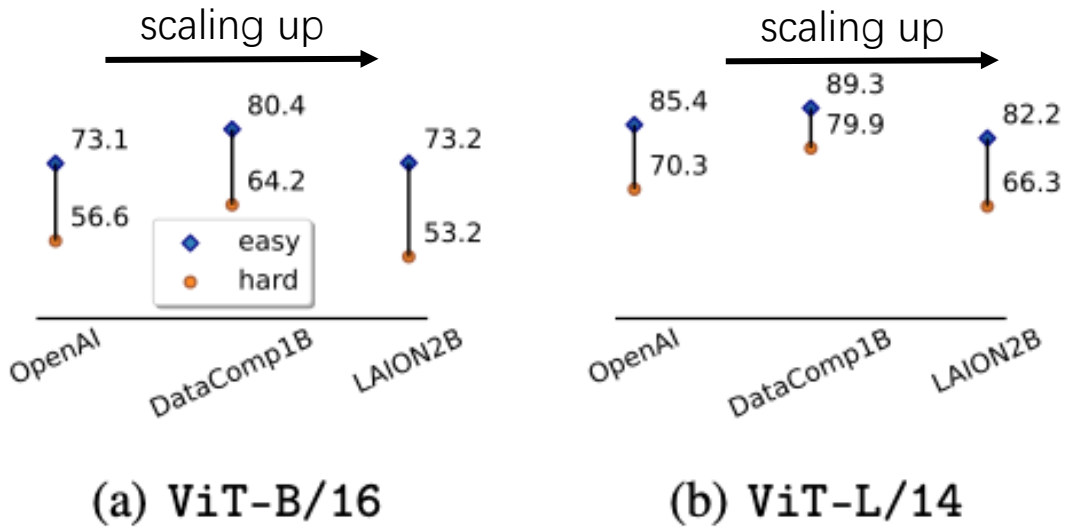
Accuracy and Performance Drop.



Effective Robustness.

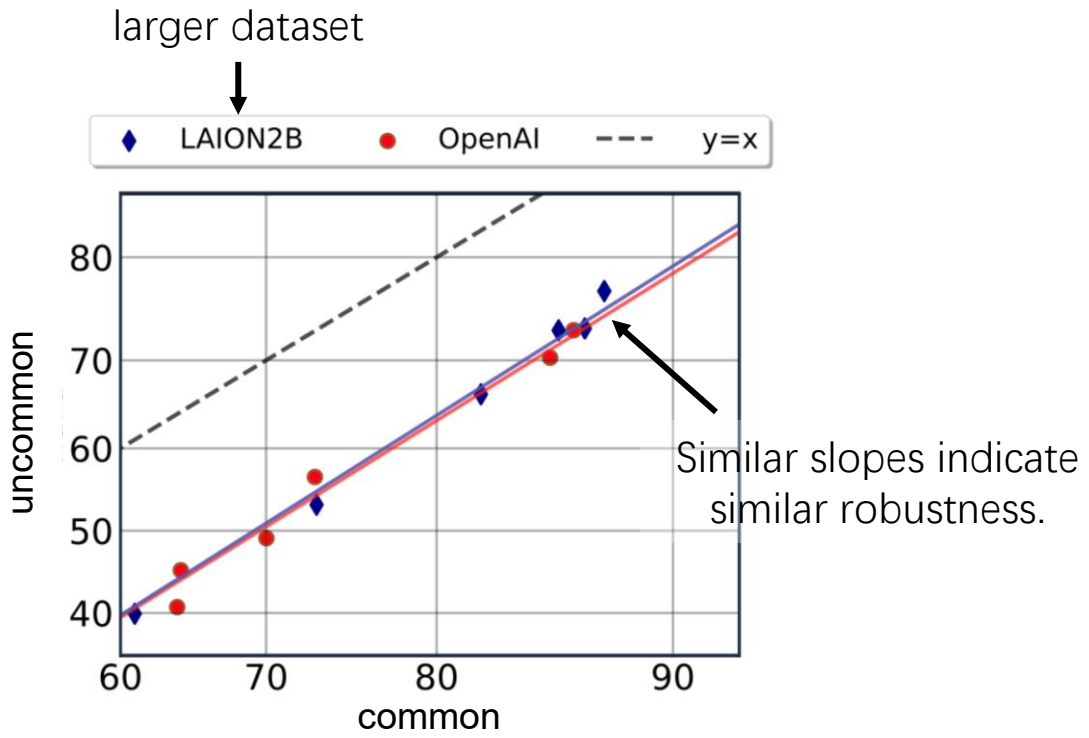
Observation 3 (Model Size). Scaling up model size CAN enhance CLIP robustness.

Observations



(a) ViT-B/16 (b) ViT-L/14

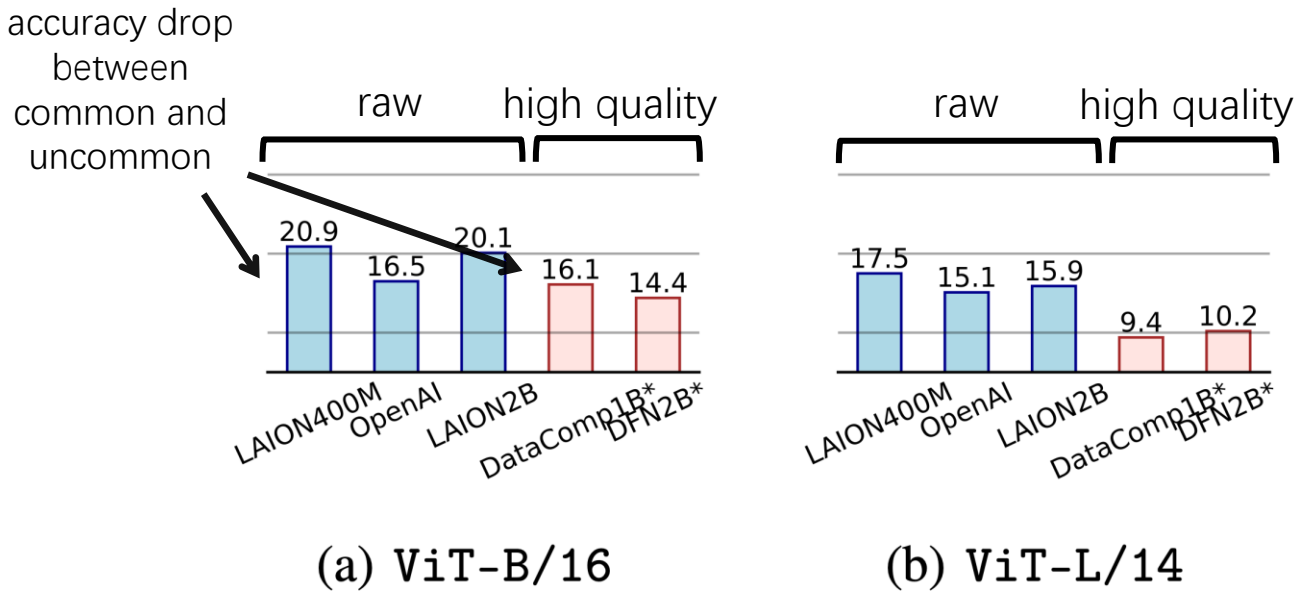
Accuracy and Performance Drop.



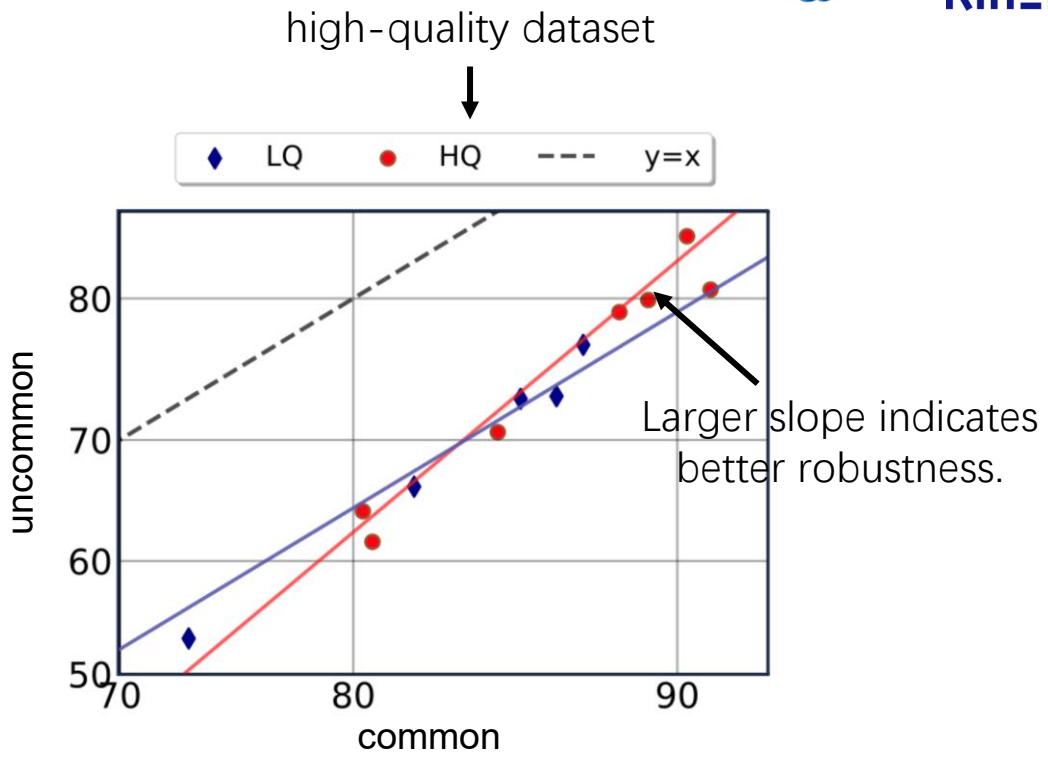
Effective Robustness.

Observation 4 (Data Size). Scaling up data size CANNOT enhance CLIP robustness.

Observations



Performance Drop.



Effective Robustness.

Observation 5 (Data Quality). Improving data quality CAN enhance CLIP robustness.

Theoretical Understanding

Assumption (Multi-modal Dataset). Considering n image-text pairs $\{(\boldsymbol{x}_{\mathrm{I}}^i, \boldsymbol{x}_{\mathrm{T}}^i)\}_{i=1}^n$, both $\boldsymbol{x}_{\mathrm{I}}^i$ and $\boldsymbol{x}_{\mathrm{T}}^i$ are generated from the latent factor $\boldsymbol{z}_i$, where $\boldsymbol{z} = [z_{inv}, z_{spu}] \in \mathbb{R}^2$ is composed of

- **invariant feature** $z_{inv} \sim \mathcal{N}(\mu_{inv} y, \sigma_{inv}^2)$
- **spurious feature** $z_{spu} \sim \mathcal{N}(\mu_{spu} a, \sigma_{spu}^2)$

with $\Pr(a = y) = p_{spr}$ otherwise $a = -y$. y is the label uniformly drawn from $\{-1,1\}$. The training data $\mathcal{D}^{tr}$ is drawn with $\frac{1}{2} \leq p_{spr} \leq 1$ and test data $\mathcal{D}^*$ is drawn with $p_{spu} = \frac{1}{2}$.

Note. The dataset is **biased** to spurious feature z_{spu} due to **different** p_{spr} between training and test.

Theorem 1. Given the multi-modal dataset with a large spurious correlation $p_{spu} = 1 - o(1)$. Then, under reasonable assumptions, w.p. at least $1 - O(1)$, the CLIP model achieves

- a **small zero-shot error** on test data where $a = y$: $\mathrm{Acc}(g_{\mathrm{I}}, g_{\mathrm{T}}) \geq 1 - \Phi(\kappa_2) - o(1)$,
- a **large zero-shot error** on test data where $a \neq y$: $\mathrm{Err}(g_{\mathrm{I}}, g_{\mathrm{T}}) \geq 1 - \Phi(\kappa_1) - o(1)$.

Therein, κ_1, κ_2 are constants that depend on $\mu_{inv}, \sigma_{inv}, \mu_{inv}$, and σ_{inv}.

Note. The model relies on whether $a = y$ (whether biased) to make right predictions.

Take Home Messages

We should be cautious about **test setups** when assessing new **training setups**.

CounterAnimal (https://counteranimal.github.io/) is a proper benchmark for assessing the robustness of CLIPs to spurious features.

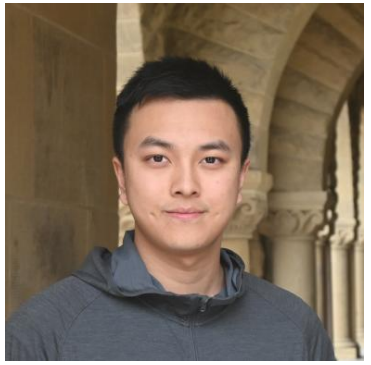
Distribution shifts remain an open question for CLIP and other VLMs.

Scaling up model size can enhance robustness, while **scaling up pre-train data** is not that effective.

Improving data quality is effective to enhance robustness.

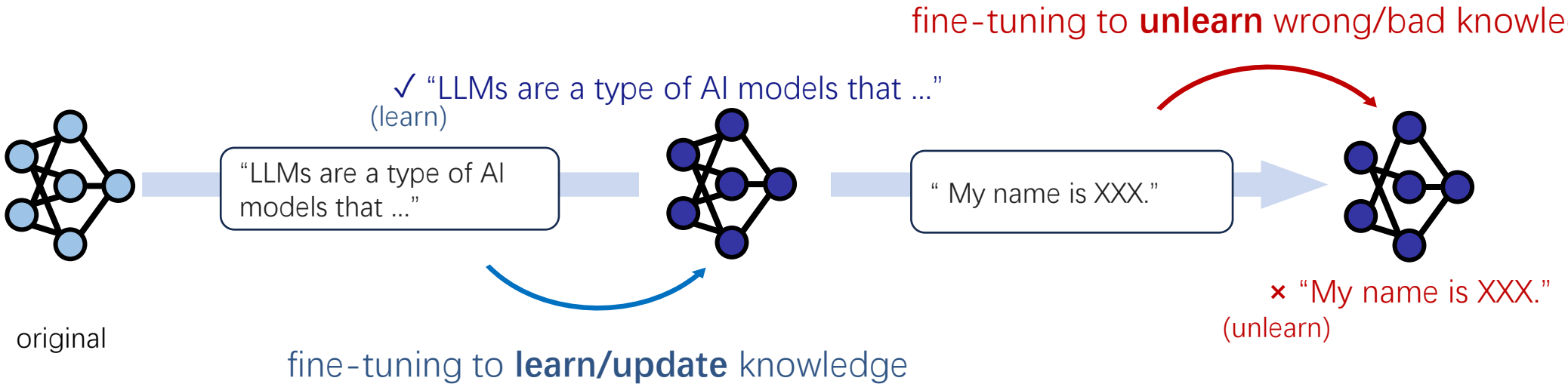
Part II: Finetuning

Qizhou Wang



Zhanke Zhou

Finetuning aims to adapts the model parameters to fit tasks or knowledge, of which the specific goals can be attributed to **learning** and **unlearning**.



Qizhou Wang, Jin Peng Zhou, **Zhanke Zhou**, Saebyeol Shin, **Bo Han**, Kilian Q. Weinberger. Rethinking LLM Unlearning Objectives: A Gradient Perspective and Go Beyond. In *ICLR*, 2025.

Right to be Forgotten

"The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data …"

"A consumer shall have the right to request that a business delete any personal information about the consumer which the business has collected from the consumer …"

LLM Unlearning

to be unlearned

Bi-objective Goal

- **Unlearn:** removing model capability to generate targeted data $\mathcal{D}_{\mathrm{u}} = \{s_{\mathrm{u}}\}_{n_{\mathrm{u}}}$
- **Retain:** maintain performance on other non-targeted data $\mathcal{D}_{\mathrm{r}} = \{s_{\mathrm{r}}\}_{n_{\mathrm{r}}}$

not to be unlearned

Gradient Ascent (GA)-based Method

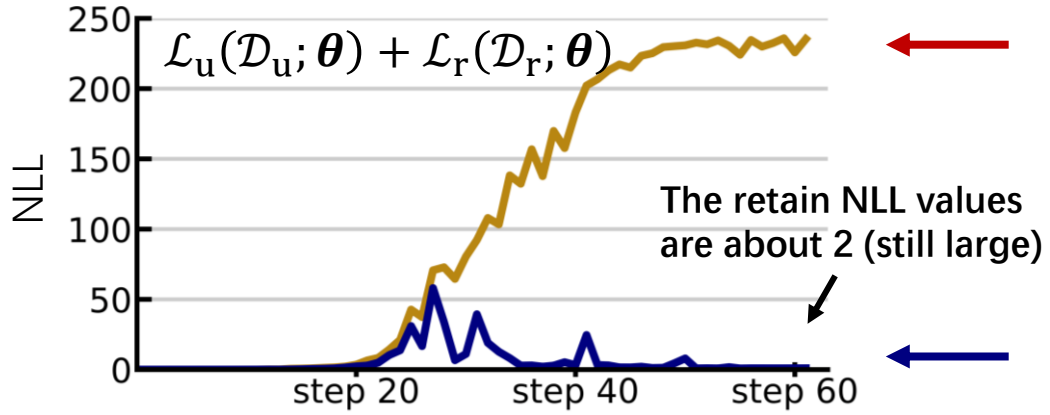
$$\min_{\boldsymbol{\theta}} \underbrace{\mathbb{E}_{\mathcal{D}_{\mathrm{u}}} \log P(s_{\mathrm{u}}; \boldsymbol{\theta})}_{\mathcal{L}_{\mathrm{u}}(\mathcal{D}_{\mathrm{u}}; \boldsymbol{\theta})} + \underbrace{\mathbb{E}_{\mathcal{D}_{\mathrm{r}}} -\log P(s_{\mathrm{r}}; \boldsymbol{\theta})}_{\mathcal{L}_{\mathrm{r}}(\mathcal{D}_{\mathrm{r}}; \boldsymbol{\theta})}$$

Unlearn Objective **Retain Objective**

Basic Assumption: If the negative log-likelihood is a proper objective for learning, then the log-likelihood should be appropriate for unlearning.

Impacts of GA

Negative log-likelihood (NLL) as the **metric** $\mathcal{R}$ to assess performance.



Performance regarding unlearning and retention.

$\mathcal{R}(\mathcal{D}_{\mathrm{u}};\boldsymbol{\theta})$: **large unlearn NLL** indicates **strong** unlearning.

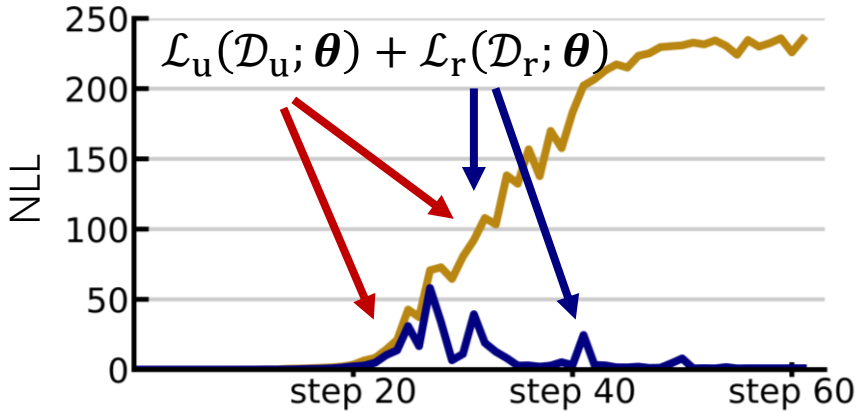
$\mathcal{R}(\mathcal{D}_{\mathrm{r}};\boldsymbol{\theta})$: **large retain NLL** indicates **damage** in retention.

Observation 1. GA-based methods CAN achieve strong unlearning but CANNOT ensure reliable retention, thus **NOT meeting the dual-objective goal.**

Delve Deeper?

Performance metrics offer **limited** insights towards deeper understandings.

Limitation 1. We CANNOT **disentangle** the impacts of $\mathcal{L}_{\mathrm{u}}(\mathcal{D}_{\mathrm{u}};\boldsymbol{\theta})$ and $\mathcal{L}_{\mathrm{r}}(\mathcal{D}_{\mathrm{r}};\boldsymbol{\theta})$ on model performance.



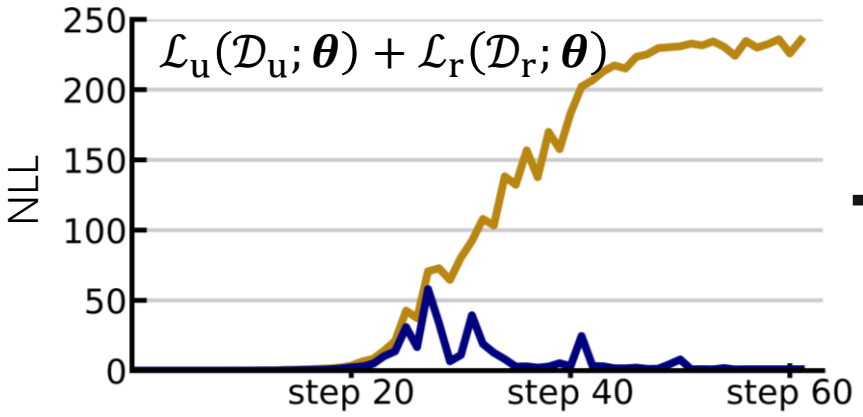
Using NLL to assess performance changes regarding unlearning and retention.

Both $\mathcal{L}_{\mathrm{u}}(\mathcal{D}_{\mathrm{u}};\boldsymbol{\theta})$ and $\mathcal{L}_{\mathrm{r}}(\mathcal{D}_{\mathrm{r}};\boldsymbol{\theta})$ have impacts on $\mathcal{R}(\mathcal{D}_{\mathrm{u}};\boldsymbol{\theta})$ and $\mathcal{R}(\mathcal{D}_{\mathrm{r}};\boldsymbol{\theta})$ in an **intertwined** manner.

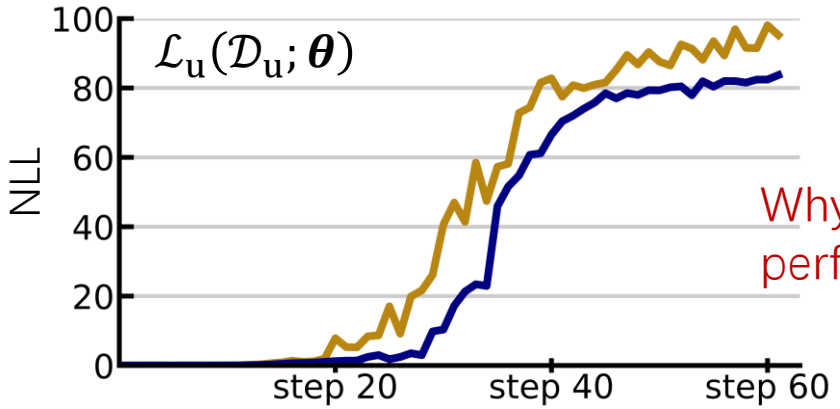
Delve Deeper?

Performance metrics offer **limited** insights towards deeper understandings.

Limitation 2. Even disentangled, we CANNOT fully **understand the factors** that lead to the observed behaviors.



Unlearning with $\mathcal{L}_u(\mathcal{D}_u;\boldsymbol{\theta}) + \mathcal{L}_r(\mathcal{D}_r;\boldsymbol{\theta})$.



For illustration, we approximate the disentanglement by unlearning only with $\mathcal{L}_u(\mathcal{D}_u;\boldsymbol{\theta})$.

Why does the retention performance drop so quick? 🤔

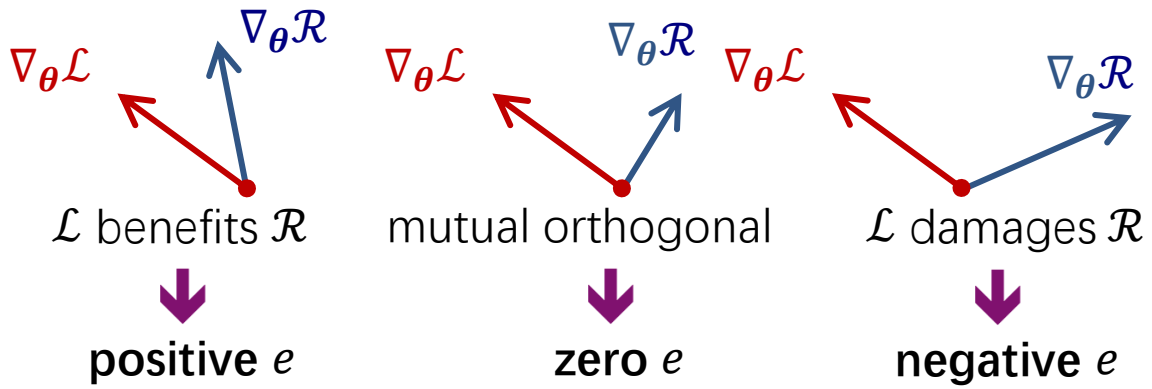
G-effect: A Gradient View

Studying the impacts of **unlearning methods** (e.g., GA) on **performance metrics** (e.g., NLL) from a gradient view.

gradients of **objective** (unlearning method)

$$e = \nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathcal{D};\boldsymbol{\theta})^{\top}\nabla_{\boldsymbol{\theta}}\mathcal{R}(\mathcal{D};\boldsymbol{\theta})$$

gradients of **metric**

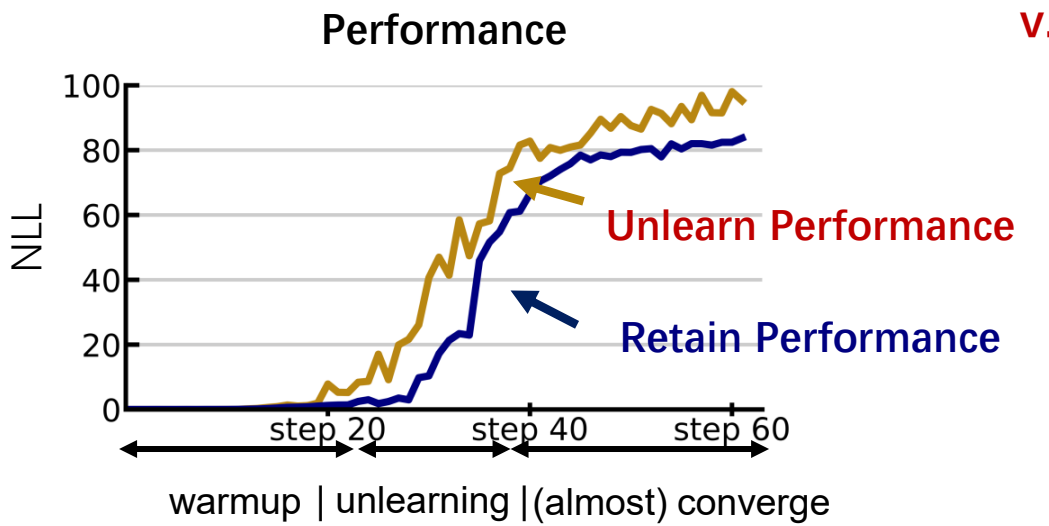


- **Fulfill Goal 1** as the G-effect can be computed for $\mathcal{L}_{\mathrm{u}}(\mathcal{D}_{\mathrm{u}};\boldsymbol{\theta})$ and $\mathcal{L}_{\mathrm{r}}(\mathcal{D}_{\mathrm{r}};\boldsymbol{\theta})$ separately.
- **Fulfill Goal 2** as gradients provide more messages than merely CE performance.

G-effect: An Example

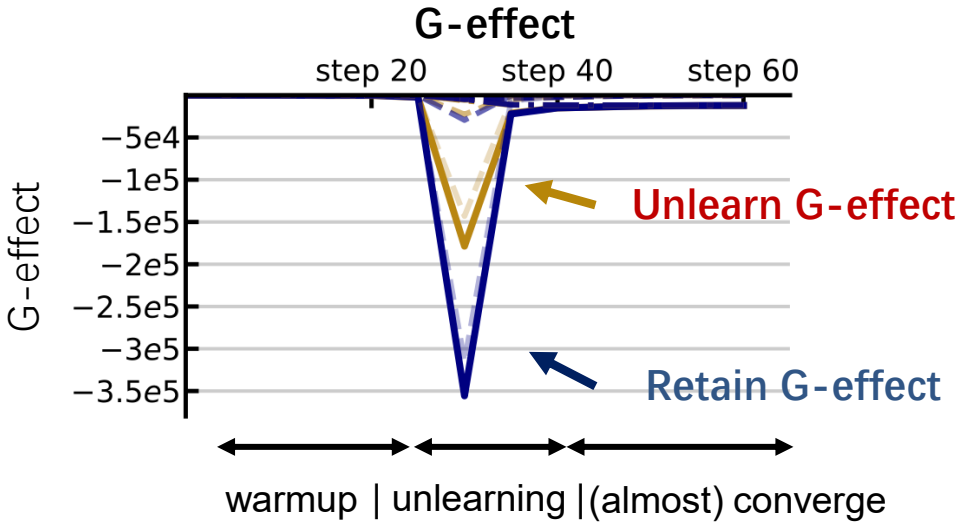
Retain G-effect: $e_{\mathrm{r}} = \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D}_{\mathrm{u}}; \boldsymbol{\theta})^{\top} \nabla_{\boldsymbol{\theta}} \mathcal{R}(\mathcal{D}_{\mathrm{r}}; \boldsymbol{\theta})$. A **positive** e_{r} is preferred to enhance retention.

Unlearn G-effect: $e_{\mathrm{u}} = \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D}_{\mathrm{u}}; \boldsymbol{\theta})^{\top} \nabla_{\boldsymbol{\theta}} \mathcal{R}(\mathcal{D}_{\mathrm{u}}; \boldsymbol{\theta})$. A **negative** e_{u} is preferred for strong unlearning.



Using NLL to assess performance.

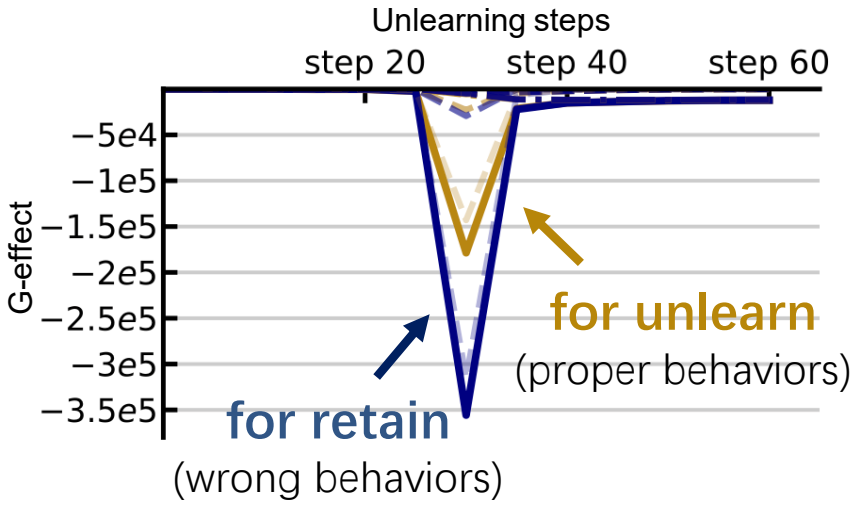
V.S.



Using G-effect to assess performance change.

Note. The G-effect quantifies the **rate of change** (increase/decrease) in performance, which can be calculated **separately** for retention and unlearning.

GA: Objective 1



The G-effects of GA.

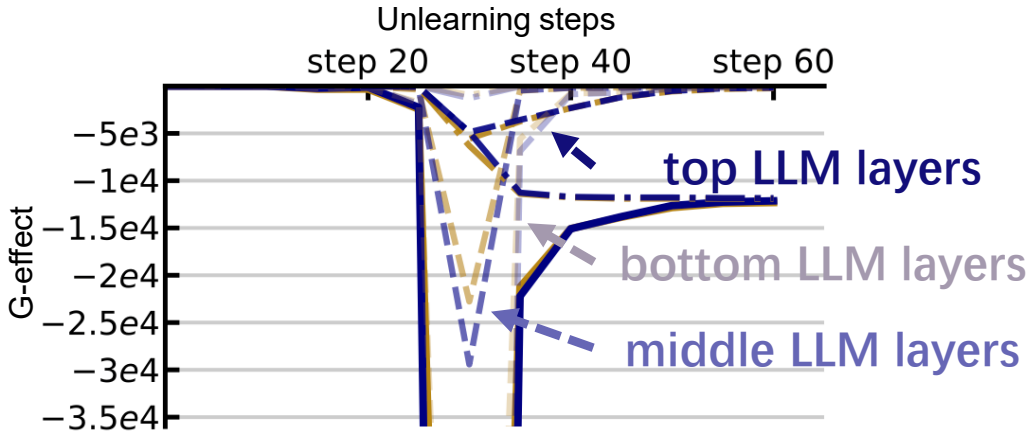
Objective: $\mathbb{E}_{\mathcal{D}_\mathrm{u}} \sum_i \log P\left(s_\mathrm{u}^i | s_\mathrm{u}^{<i}; \boldsymbol{\theta}\right)$

Gradient: $\mathbb{E}_{\mathcal{D}_\mathrm{u}} \sum_i \underbrace{\frac{1}{P\left(s_\mathrm{u}^i | s_\mathrm{u}^{<i}; \boldsymbol{\theta}\right)}}_{\text{inverse likelihood}} \nabla_{\boldsymbol{\theta}} P\left(s_\mathrm{u}^i | s_\mathrm{u}^{<i}; \boldsymbol{\theta}\right)$

Observation 2. **Excessive extent of removal** incurs negative costs to retention.

Reason. The inverse likelihood wrongly focuses more on sufficiently unlearned tokens, leading to **over-unlearning** that negatively impacts model utility.

GA: Objective 1



The G-effects of GA (closer look).

Objective: $\mathbb{E}_{\mathcal{D}_\mathrm{u}} \sum_i \log P\left(s_\mathrm{u}^i | s_\mathrm{u}^{<i}; \boldsymbol{\theta}\right)$

Gradient: $\mathbb{E}_{\mathcal{D}_\mathrm{u}} \sum_i \underbrace{\frac{1}{P\left(s_\mathrm{u}^i | s_\mathrm{u}^{<i}; \boldsymbol{\theta}\right)}}_{\text{inverse likelihood}} \nabla_{\boldsymbol{\theta}} P\left(s_\mathrm{u}^i | s_\mathrm{u}^{<i}; \boldsymbol{\theta}\right)$

Observation 3. Unlearning **affects on bottom layers** of LLMs more than others.

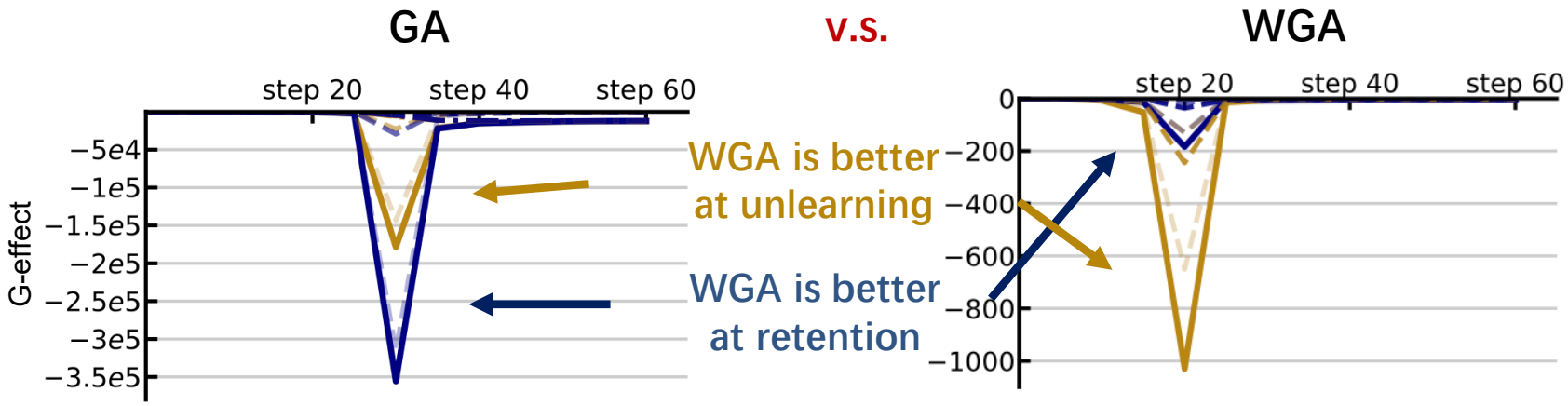
Reason. Large gradients will **accumulate** due to the chain rule, a general scenario holds for many other unlearning objectives.

WGA: Improvement 1

Motivation: Combating the inverse likelihood term via **loss reweighting.**

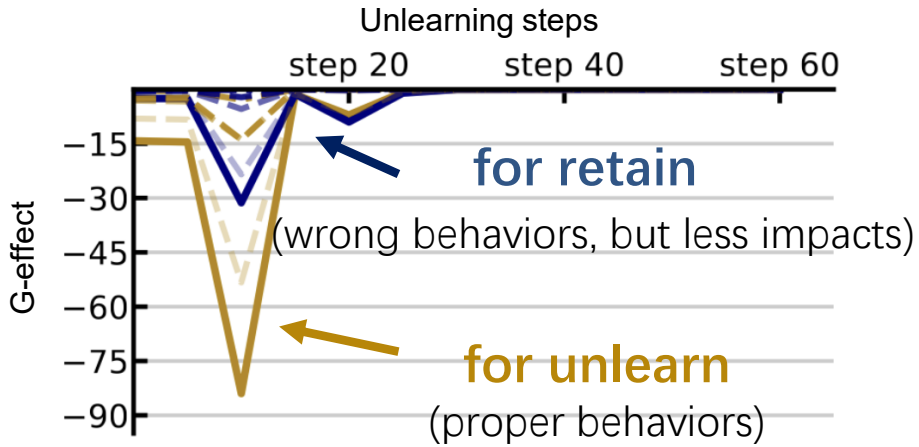
Original GA: $\mathbb{E}_{\mathcal{D}_\mathrm{u}} \sum_i \log P\left(s_\mathrm{u}^i | s_\mathrm{u}^{<i}; \boldsymbol{\theta}\right)$ ➔ **Weighted GA:** $\mathbb{E}_{\mathcal{D}_\mathrm{u}} \sum_i P\left(s_\mathrm{u}^i | s_\mathrm{u}^{<i}; \boldsymbol{\theta}\right)^{\alpha} \log P\left(s_\mathrm{u}^i | s_\mathrm{u}^{<i}; \boldsymbol{\theta}\right)$

Gradients: $\mathbb{E}_{s_\mathrm{u} \sim \mathcal{D}_\mathrm{u}} \sum_i \underbrace{P\left(s_\mathrm{u}^i | s_\mathrm{u}^{<i}; \boldsymbol{\theta}\right)^{\alpha-1}}_{\text{counteract the inverse likelihood}} \nabla_{\boldsymbol{\theta}} P\left(s_\mathrm{u}^i | s_\mathrm{u}^{<i}; \boldsymbol{\theta}\right)$



Comparison of the G-effects between GA and WGA.

NPO: Objective 2



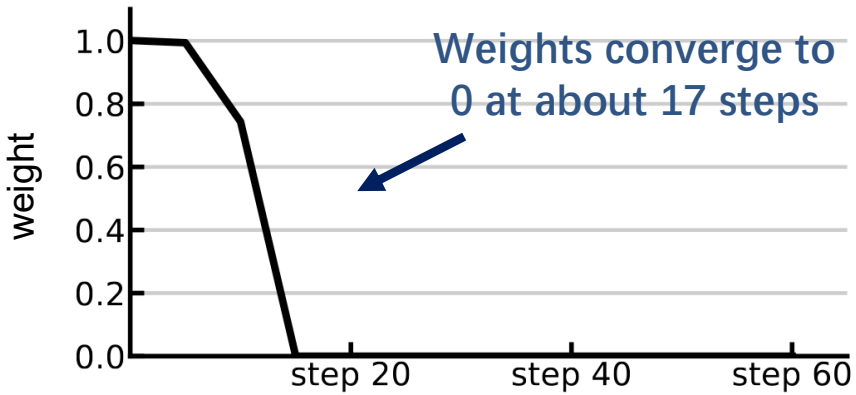
The G-effects of NPO.

Objective: $\mathbb{E}_{\mathcal{D}_\mathrm{u}} \frac{1}{\beta}\log\left(1+\left(\frac{p(s_\mathrm{u};\boldsymbol{\theta})}{p(s_\mathrm{u};\boldsymbol{\theta}_\mathrm{o})}\right)^{\beta}\right)$

Gradient: $\mathbb{E}_{\mathcal{D}_\mathrm{u}} \sum_i \underbrace{\frac{2P(s_\mathrm{u};\boldsymbol{\theta})^{\beta}}{P(s_\mathrm{u};\boldsymbol{\theta})^{\beta}+P(s_\mathrm{u};\boldsymbol{\theta}_\mathrm{o})^{\beta}}}_{w_\mathrm{npo}\text{ reweighting}} \nabla_{\boldsymbol{\theta}} \log P(s_\mathrm{u};\boldsymbol{\theta})$

Observation 4. NPO (Negative Preference Optimization) has **fewer negative impacts** on retention compared to GA.
Reason. The gradients of NPO are very similar to GA, yet further **reweighting** by w_npo, which mainly contributes to its improvements over GA.

NPO: Objective 2



The curve of w_{npo} during unlearning.

Objective: $\mathbb{E}_{\mathcal{D}_\mathrm{u}} \frac{1}{\beta} \log\left(1 + \left(\frac{p(s_\mathrm{u};\boldsymbol{\theta})}{p(s_\mathrm{u};\boldsymbol{\theta}_\mathrm{o})}\right)^{\beta}\right)$

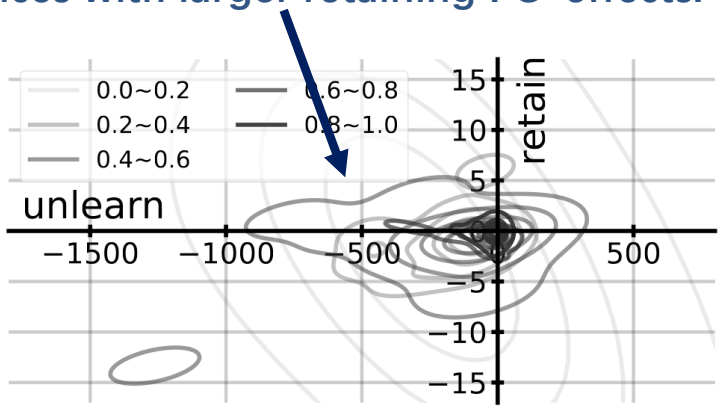
Gradient: $\mathbb{E}_{\mathcal{D}_\mathrm{u}} \sum_i \underbrace{\frac{2P(s_\mathrm{u};\boldsymbol{\theta})^{\beta}}{P(s_\mathrm{u};\boldsymbol{\theta})^{\beta}+P(s_\mathrm{u};\boldsymbol{\theta}_\mathrm{o})^{\beta}}}_{w_\mathrm{npo}\text{ reweighting}} \nabla_{\boldsymbol{\theta}} \log P(s_\mathrm{u};\boldsymbol{\theta})$

Observation 5. The NPO weight w_npo serves a role like **early stopping.**

Reason. w_npo approaches 0 when $P(s_\mathrm{u};\boldsymbol{\theta}) \to 0$.

NPO: Objective 2

Larger weights are assigned to those instances with larger retaining PG-effects.



The distributions of the point-wise G-effects across different range of w_{npo}.

Gradient: $\mathbb{E}_{\mathcal{D}_\mathrm{u}} \sum_i \frac{2P(s_\mathrm{u};\boldsymbol{\theta})^\beta}{P(s_\mathrm{u};\boldsymbol{\theta})^\beta + P(s_\mathrm{u};\boldsymbol{\theta}_\mathrm{o})^\beta} \nabla_{\boldsymbol{\theta}} \log P(s_\mathrm{u};\boldsymbol{\theta})$

G-effect: $\mathbb{E}_{\mathcal{D}_\mathrm{u}} \underbrace{w_\mathrm{npo}}_{\text{weights}} \underbrace{\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{s}_\mathrm{u};\boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}} \mathcal{R}(\mathcal{D};\boldsymbol{\theta})}_{\text{point-wise G-effect (PG-effect)}}$

(The impacts of a particular data point on model performance.)

Observation 6. The NPO reweighting mechanism w_npo **prioritizes instances** that less damages retention.

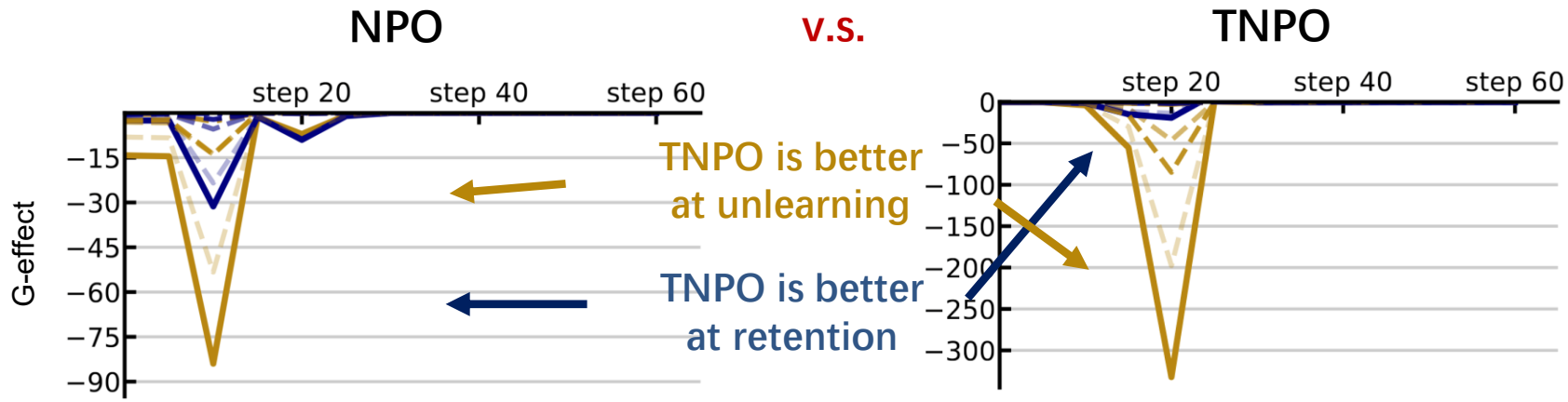
Reason. Data that have small impacts on **retention** also have small impacts on **unlearning.**

TNPO: Improvement 2

Motivation: Generalized the reweighting mechanism of NPO for tokens.

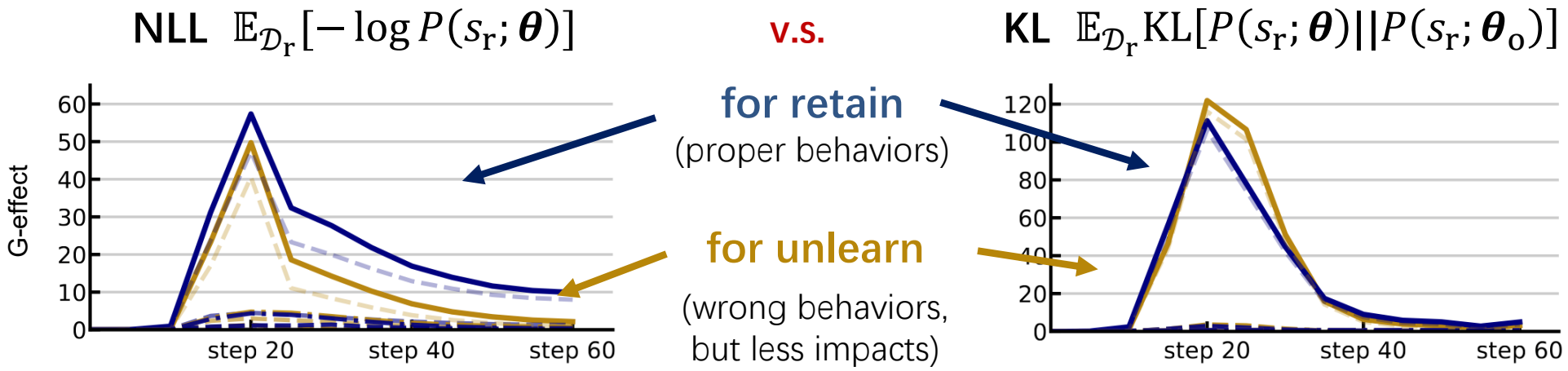
Token-wise NPO $\sum_i w^i_{\text{tnpo}} \log P\left(s^i_{\text{u}} | s^{<i}_{\text{u}}; \boldsymbol{\theta}\right)$ with $w^i_{\text{tnpo}} = \underbrace{\frac{2P\left(s^i_{\text{u}} | s^{<i}_{\text{u}}; \boldsymbol{\theta}\right)^{\alpha}}{P\left(s^i_{\text{u}} | s^{<i}_{\text{u}}; \boldsymbol{\theta}\right)^{\alpha} + P\left(s^i_{\text{u}} | s^{<i}_{\text{u}}; \boldsymbol{\theta}_{\text{o}}\right)^{\alpha}}}$

same reweighting scheme yet applied point-wise.



Comparison of the G-effects between NPO and TNPO.

Retain Objectives



Comparison between two representative retain objectives.

Observation 7. **NLL** and **KL** are both effective for retention, while KL can lead to overall larger retain G-effect, thus preferred.

Note. The unlearn G-effect for the unlearning objective is much larger than for the retain objectives. Thus, we do not need to worry about the side effect on unlearning.

Empirical Evaluations

LLM setup	method	Phi-1.5 ES-exact retain ↑	Phi-1.5 ES-exact unlearn ↓	Phi-1.5 ES-perturb retain ↑	Phi-1.5 ES-perturb unlearn ↓	Phi-1.5 MU ↑	Phi-1.5 FQ ↑	Llama-2-7B ES-exact retain ↑	Llama-2-7B ES-exact unlearn ↓	Llama-2-7B ES-perturb retain ↑	Llama-2-7B ES-perturb unlearn ↓	Llama-2-7B MU ↑	Llama-2-7B FQ ↑
before unlearning		0.44	0.59	0.21	0.16	0.52	-5.80	0.82	0.80	0.53	0.40	0.63	-7.59
1%	GA	0.11	0.05	0.08	0.08	0.37	-0.54	0.42	**0.05**	0.26	**0.04**	0.53	-0.54
	PO	**0.36**	0.84	**0.16**	0.36	**0.51**	-4.24	**0.75**	0.83	**0.47**	0.52	0.62	-5.80
	WGA	**0.36**	**0.03**	**0.18**	**0.02**	**0.51**	**-0.54**	**0.67**	0.08	0.38	0.06	**0.65**	**-0.08**
	NPO	0.27	0.09	0.11	0.07	0.48	-2.91	0.47	0.12	0.38	0.09	0.62	-1.32
	TNPO	0.33	**0.03**	0.12	**0.04**	0.49	**-0.08**	0.51	**0.03**	**0.43**	**0.03**	**0.64**	**-0.08**
	RMU	0.23	0.08	0.15	0.05	0.43	-0.54	0.23	0.08	0.15	0.05	0.52	-1.32
before unlearning		0.44	0.56	0.21	0.23	0.52	-29.65	0.82	0.77	0.53	0.41	0.63	-32.13
5%	GA	0.00	**0.00**	0.00	**0.00**	0.00	-11.40	0.03	**0.00**	0.02	**0.00**	0.00	**-12.42**
	PO	**0.26**	0.79	**0.16**	0.49	**0.51**	-26.50	**0.55**	0.84	**0.36**	0.49	**0.64**	-28.84
	WGA	**0.29**	0.01	**0.16**	0.01	**0.51**	**-1.30**	0.47	0.00	**0.39**	0.00	**0.64**	-16.32
	NPO	0.08	0.12	0.08	0.06	0.38	-7.75	0.17	0.07	0.12	0.08	0.52	**-9.95**
	TNPO	0.16	0.01	0.08	0.00	0.46	-2.18	**0.50**	0.01	0.34	0.00	0.63	-32.13
	RMU	0.21	**0.00**	0.12	**0.00**	0.27	**-1.95**	0.12	**0.00**	0.12	**0.00**	0.58	-21.44
before unlearning		0.44	0.47	0.21	0.18	0.52	-39.00	0.82	0.83	0.53	0.30	0.63	-44.45
10%	GA	0.00	**0.00**	0.00	**0.00**	0.00	-45.26	0.00	**0.00**	0.00	**0.00**	0.00	-20.86
	PO	**0.32**	0.73	0.14	0.26	0.50	-38.25	**0.55**	0.84	**0.37**	0.43	**0.62**	-39.76
	WGA	**0.34**	**0.00**	**0.16**	**0.00**	**0.51**	-9.06	**0.66**	0.02	**0.42**	**0.01**	0.62	-24.85
	NPO	0.08	0.09	0.07	0.07	0.38	-10.57	0.12	0.13	0.10	0.14	0.50	**-12.19**
	TNPO	0.20	0.01	0.09	0.01	**0.50**	**-7.66**	0.45	**0.01**	0.26	0.01	**0.63**	**-13.47**
	RMU	0.03	0.05	0.03	0.06	0.31	**-7.00**	0.25	0.01	0.20	0.01	0.59	-16.72

Comparison between unlearning objective on TOFU with KL regularization.

Observation 8. Larger unlearning datasets and smaller model sizes make it more challenging to unlearn.

Observation 9. GA-based works (GA & TNPO) are superior to other lines of works like PO or RMU.

Observation 10. Instance-wise reweighting is promising for unlearning efficacy.

Take Home Messages

General knowledge within **shallow layers undergoes substantial alterations** over deeper layers during unlearning.

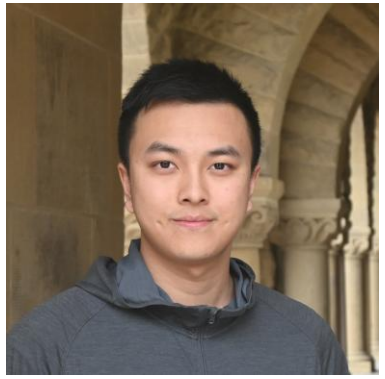
Although conceptually existing, **current objectives all fail** to retain the overall performance when conducting unlearning.

Prioritizing some tokens is effective for unlearning. However, there still exists a large space to further refine weighting mechanisms.

With **excessive unlearning**, the deterioration in common model responses can outweigh improvements in unlearning.

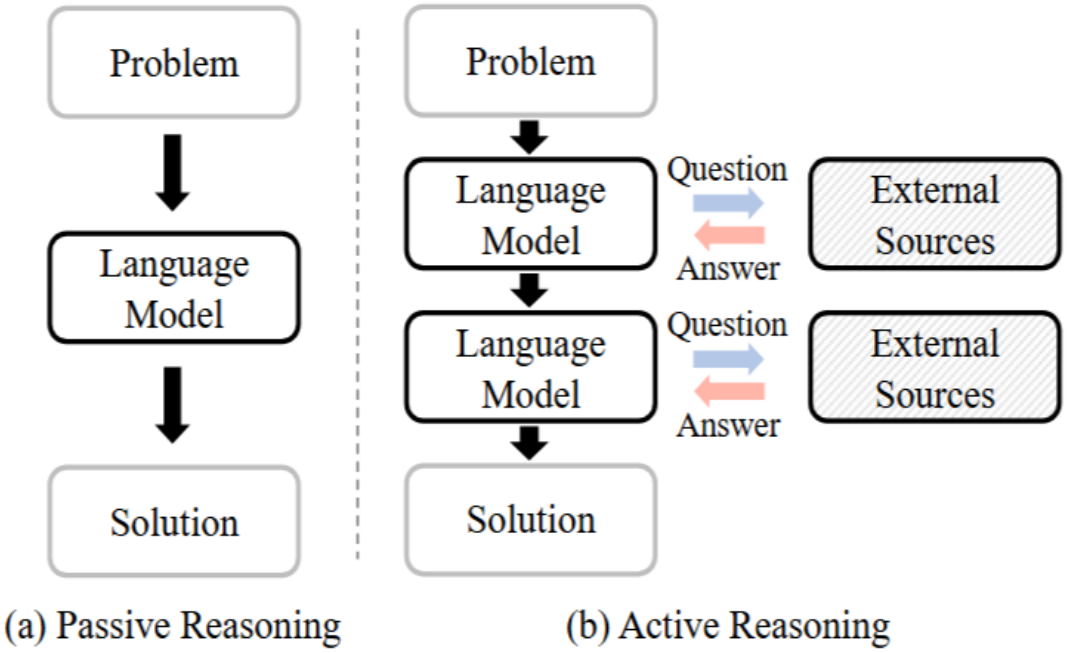
Part III: Reasoning

Can LLMs Ask the Right Questions under Incomplete Information?



Zhanke Zhou

Xiao Feng



(a) Passive Reasoning (b) Active Reasoning

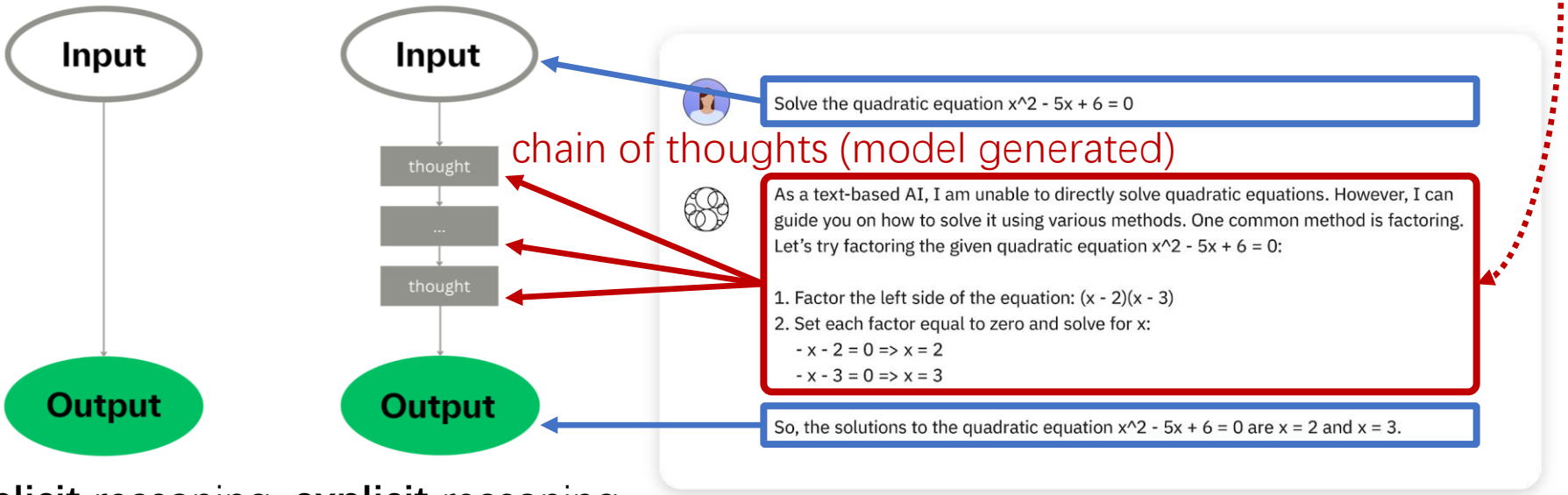
Zhanke Zhou, **Xiao Feng**, Zhaocheng Zhu, Jiangchao Yao, Sanmi Koyejo, **Bo Han**.
From Passive to Active Reasoning: Can Large Language Models Ask the Right Questions under Incomplete Information? In *ICML*, 2025

Background

Reasoning is the pathway to achieve powerful intelligence.

- Decompose a complex problem into feasible steps.
- Combine knowledge pieces into new knowledge.

Generating **chain of thoughts (CoT)** is the key of several reasoning models.



implicit reasoning **explicit** reasoning

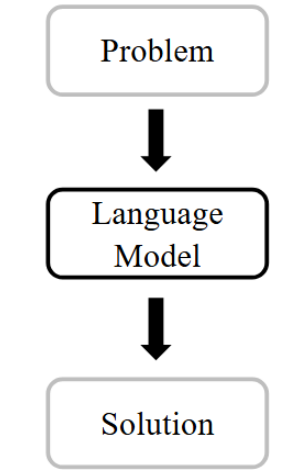
deepseek R1

OpenAI o3

Passive Reasoning

Existing works primarily focus on passive reasoning.

- A model is provided **necessary information** (e.g., conditions in a 24-Game problem).
- The model conducts **step-by-step reasoning** to derive the solution.



(a) Passive Reasoning

Passive Reasoning

Problem: Numbers: 3, 7, 8, 9
Goal: Use addition (+), subtraction (-), multiplication (×), and division (÷) to make 24. Each number must be used exactly once.

Solution:

1. 9 - 8 = 1
2. 7 + 1 = 8
3. 3 x 8 = 24

Done.

Beyond the Common Assumption

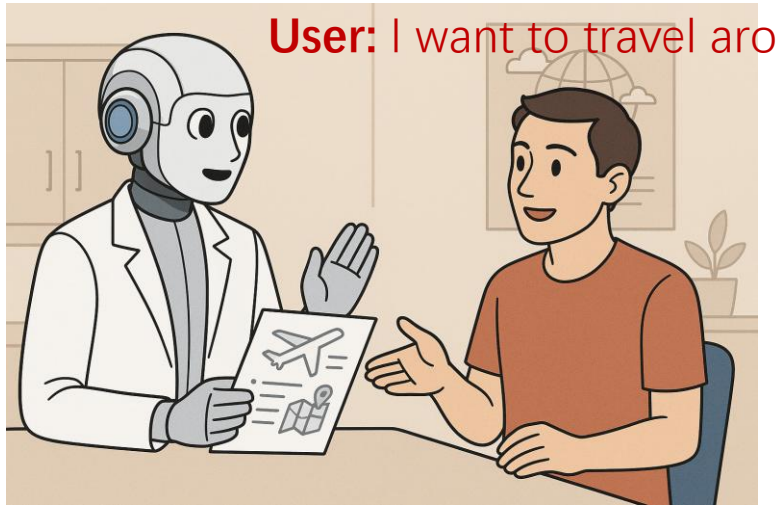
Prior works all assume that the given information is **complete.**

But what if the initially given information is **incomplete?** 🤔

Scenario 1: Travel Planner

User: I want to travel around HK.

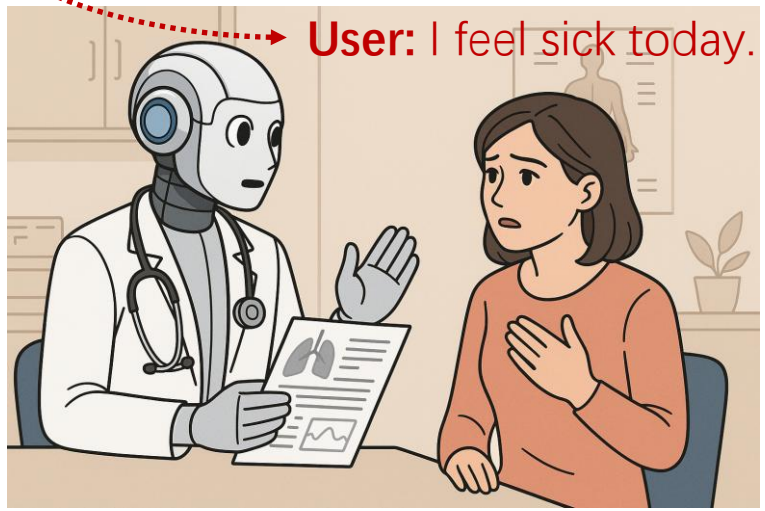
Unknown:
budget?
preferences?
time slots?
...



Scenario 2: Healthcare

User: I feel sick today.

Unknown:
symptoms?
history?
examination?
...

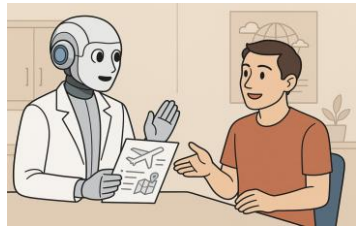


Pictures are generated by GPT-4o.

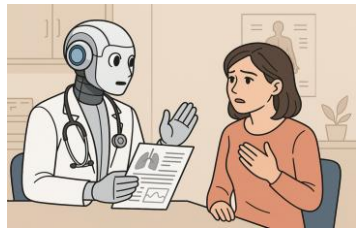
New Research Problem: Active Reasoning

Active Reasoning: The model has to actively interact with **external information sources** to **seek more essential information** and then draw the conclusion.

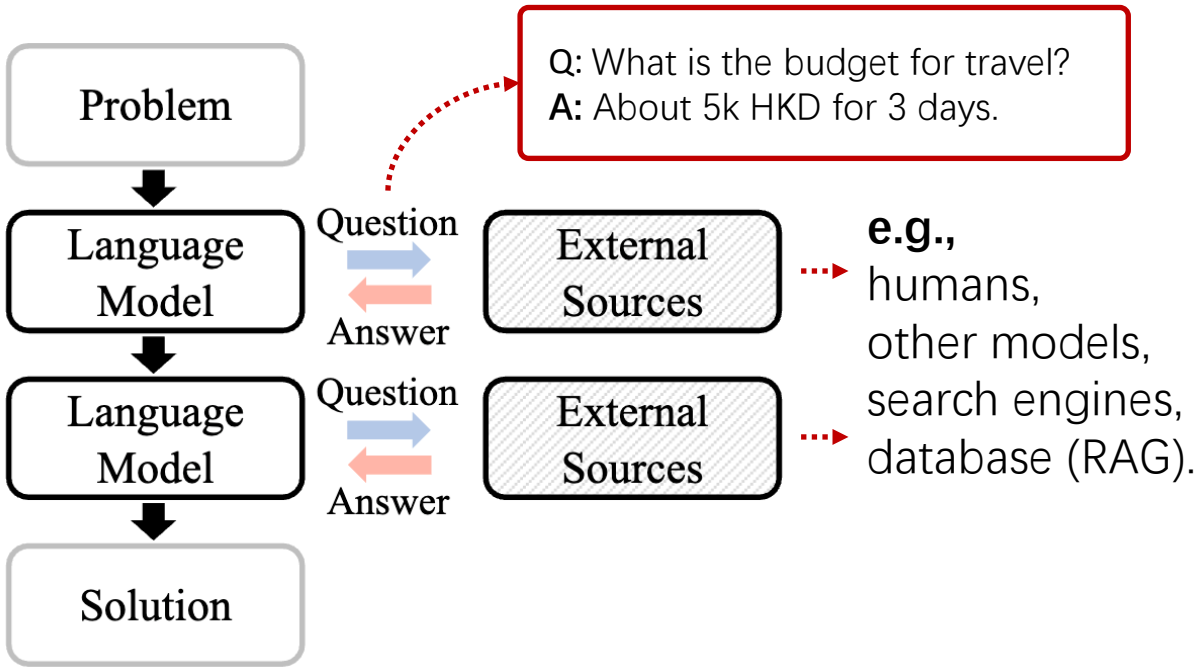
Travel Planner



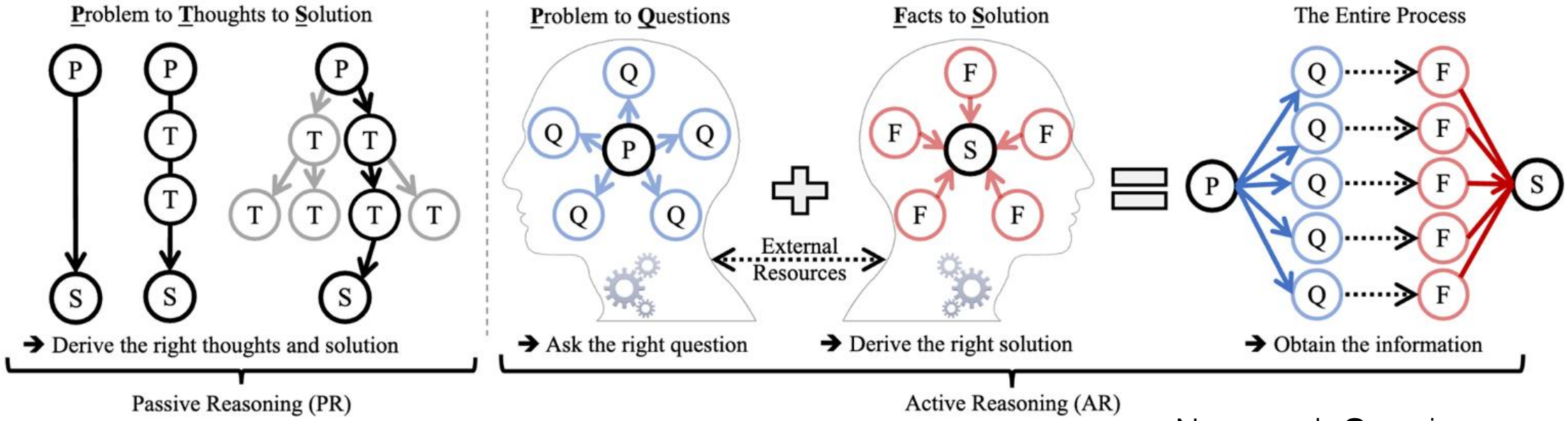
Healthcare

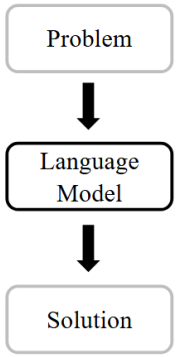


abstract to

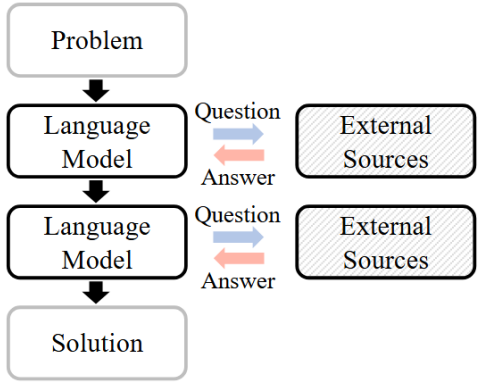


Passive Reasoning v.s. Active Reasoning





(a) Passive Reasoning



(b) Active Reasoning

Note: each **Q**uestion corresponds to an Answer (**F**act), all facts lead to **S**olution.

The Research Gap

So far, the AR capabilities of LLMs remain **largely under-explored**.

Existing AR datasets are relatively **simple** for the advanced models.

"LLMs are much more intelligent than most humans on most exams, but much less useful than most humans on most real-world tasks."

—The second half (by Shunyu Yao, Researcher at Tencent).

Therefore, it is necessary to conduct **a systematic evaluation with a new benchmarking dataset** that is tailored to active reasoning.

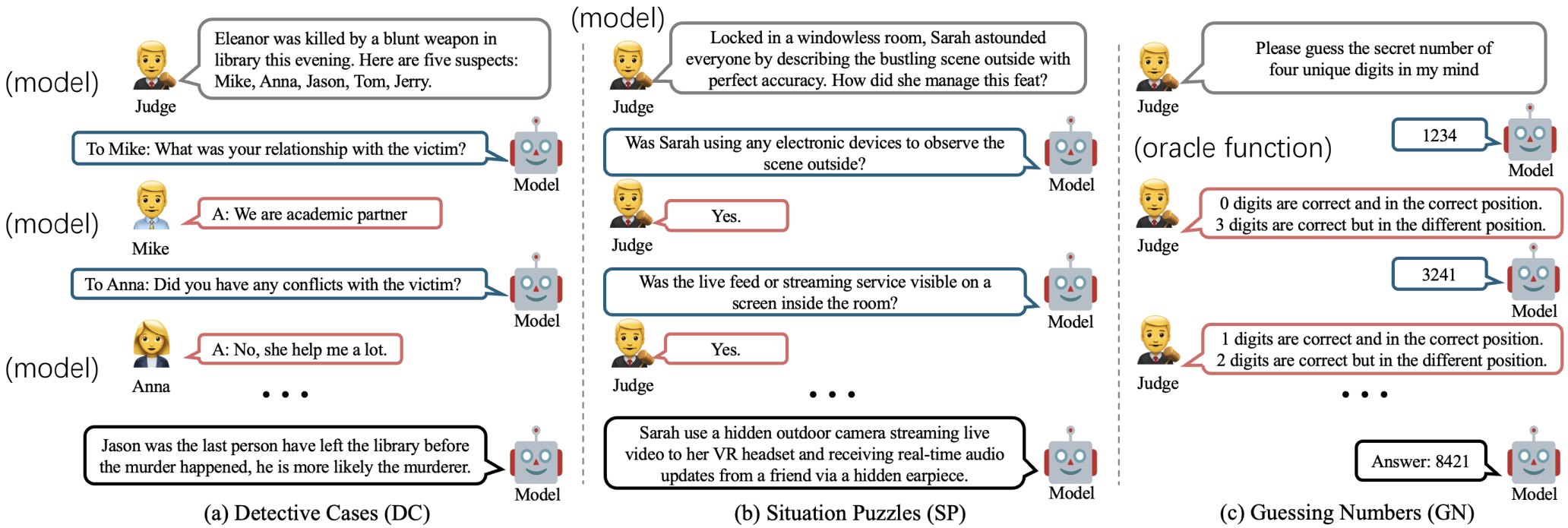
AR-Bench

Task	DC	SP	GN
Size (train/test)	400/100	400/100	4940/100
Avg. problem tokens	564.06	178.53	176.00
Interaction feedback	Narrative	Yes/No	Info. about correct digits
Answer space	5	-	5040
Metric	Accuracy	F1 score	Exact match

Table 2: Dataset statistics for the three tasks in AR-Bench.

AR-Bench (Active Reasoning Benchmark) contains **6040** problems, covering **3** AR tasks:

- **Detective Cases (DC):** The interrogation between a detective and 5 suspects.
- **Situation Puzzles (SP):** The game to reveal the truth from a puzzling mystery.
- **Guessing Numbers (GN):** The game to uncover a 4-unique-digits number.



The interaction rounds are set to 25.

Position

The uniqueness of AR-Bench

- Evaluate LLMs reasoning under **different types of feedback.**
- Require **active information-seeking** and **deliberate reasoning.**

Paradigm	Dataset	Incomplete problem	External interaction	Hypothesis verification	Language feedback	Symbolic feedback	Complex reasoning
Passive Reasoning	CommonsenseQA (Talmor et al., 2019)	✗	✗	✗	✗	✗	✗
	SocialIQA (Sap et al., 2019)	✗	✗	✗	✗	✗	✗
	GSM8K (Cobbe et al., 2021)	✗	✗	✗	✗	✗	✓
	MMLU (Hendrycks et al., 2021a)	✗	✗	✗	✗	✗	✓
	Game24 (Yao et al., 2023)	✗	✗	✗	✗	✗	✓
	Crosswords (Yao et al., 2023)	✗	✗	✗	✗	✗	✓
	Blocksworld (Valmeekam et al., 2023)	✗	✗	✗	✗	✗	✓
Active Reasoning	Qulac (Aliannejadi et al., 2019)	✓	✓	✗	✓	✗	✗
	Abg-CoQA (Guo et al., 2021)	✓	✓	✗	✓	✗	✗
	20 Questions (Abdulhai et al., 2023; Hu et al., 2024)	✓	✓	✓	✓	✗	✗
	Guess My City (Abdulhai et al., 2023)	✓	✓	✓	✓	✗	✗
	Trouble Shooting (Hu et al., 2024)	✓	✓	✓	✓	✗	✗
	MediQ (Li et al., 2024c)	✓	✓	✓	✓	✗	✗
	AR-Bench (ours)	✓	✓	✓	✓	✓	✓

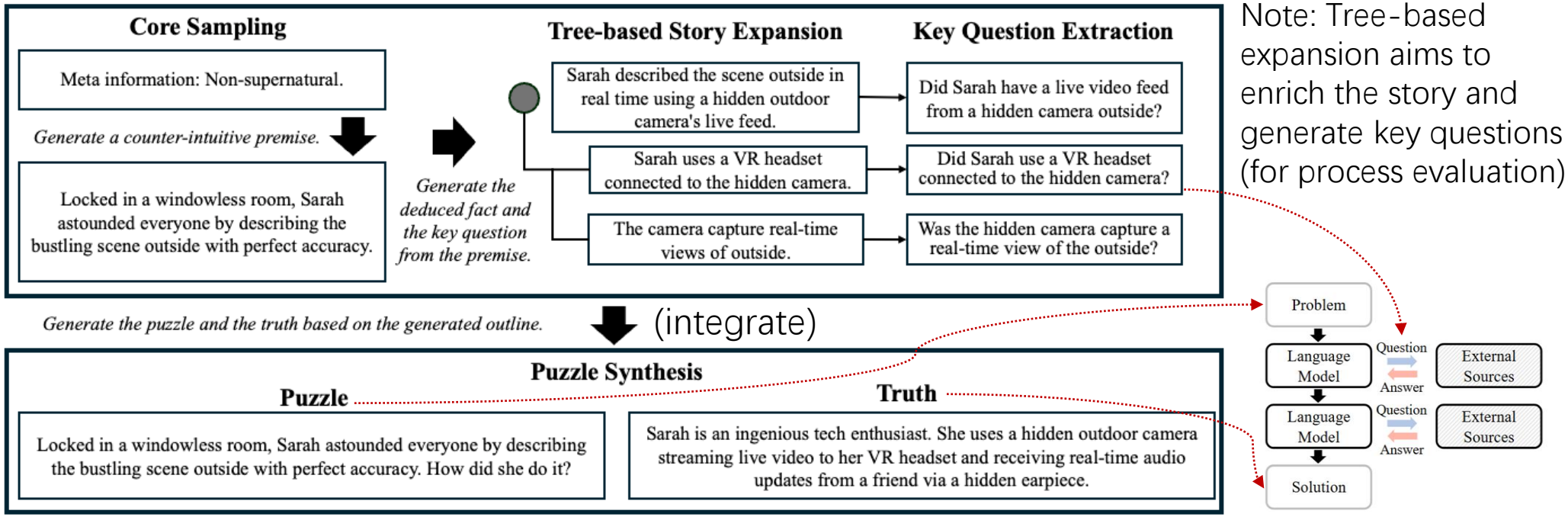
- **Hypothesis verification:** Generates initial hypothesis and verifies via asking questions.
- **Language feedback:** Receives natural-language feedback from external sources.
- **Symbolic feedback:** Receives symbolic feedback from external sources.
- **Complex reasoning:** Derives the solution through multi-step reasoning.

Dataset Construction

4-step LLM-driven construction (with a follow-up manual check)

1. **Core Sampling** to collect outlines/topics.
2. **Tree-based Story Expansion** to enrich the details of outlines/topics.
3. **Key Question Extraction** for fine-grained process analysis.
4. **Puzzle Synthesis** to construct the narrative of puzzles and truths.

(LLM-driven data generation)



Note: Tree-based expansion aims to enrich the story and generate key questions (for process evaluation).

Evaluation Metrics

Outcome score: The quality of the conclusion after the conversation.

- Detective Cases (DC): Accuracy.
- Situation Puzzles (SP): Character-level F1 score.
- Guessing Numbers (GN): Exact-match rate.

Process score: The quality of the question-answering record.

- DC & SP: How many key questions can be answered with the record.
- GN: How many digits of the guessing number match the ground truth.

Outcome Score. Process Score.

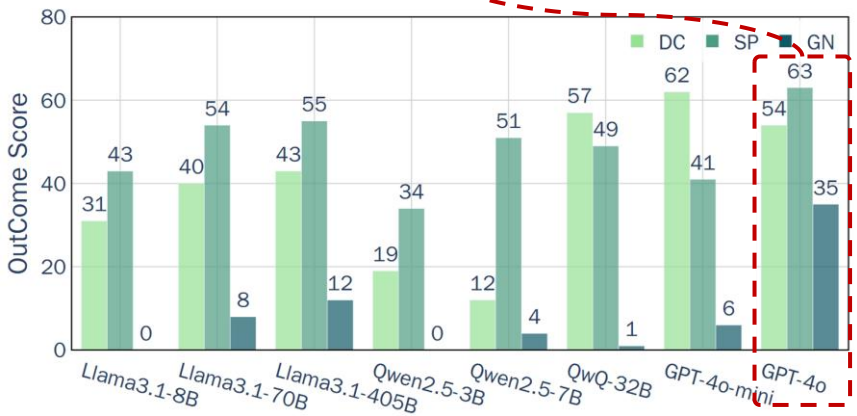
Process score of GN: (**Number of correct digits in correct positions**)/4 + 0.5 × (**Number of correct digits in wrong positions**)/4.
An example of outcome score and process score at GN (Ground truth: 2048).
Proposed number: **2048**, outcome score: 1, process score: 1.
Proposed number: **2084**, outcome score: 0, process score: 0.75.

Empirical Findings (Outcome Score)

Observations 1 & 2: AR-Bench demonstrates **challenges** across existing models and methods.

(compare different models)

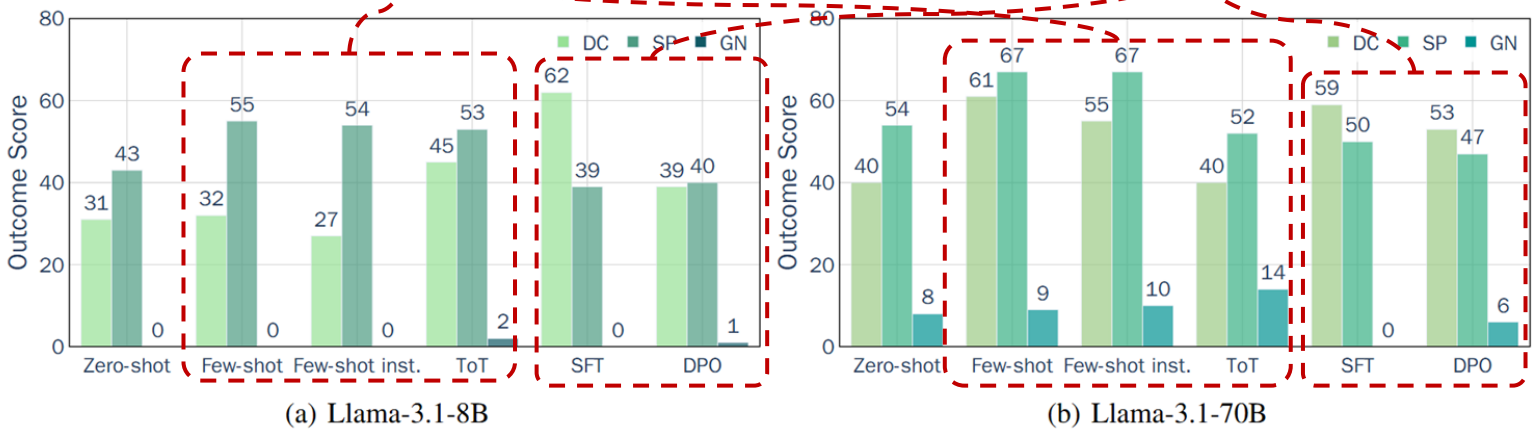
GPT-4o only has **35%** match rate in GN; other models even worse.



(compare different methods; zero-shot as the baseline)

Prompting-based methods exhibit **marginal** improvement on average across 3 tasks.

Post-training methods can even **degenerate** AR performance (in SP and GN).

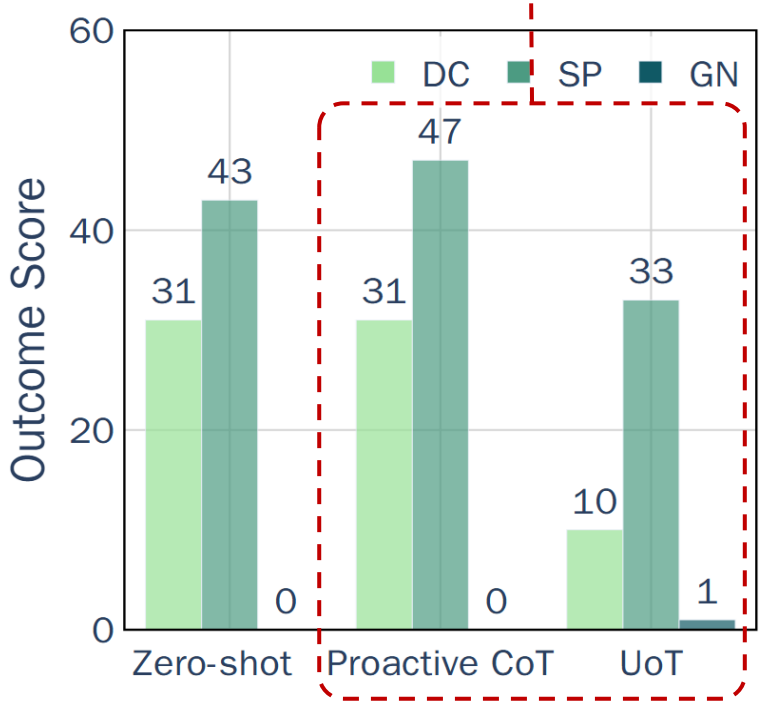


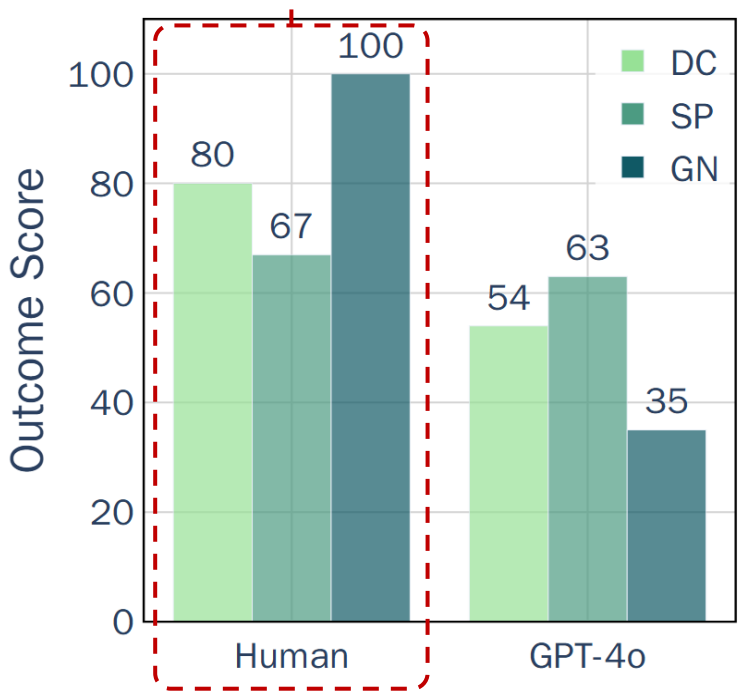
(a) Llama-3.1-8B (b) Llama-3.1-70B

Empirical Findings (Outcome Score)

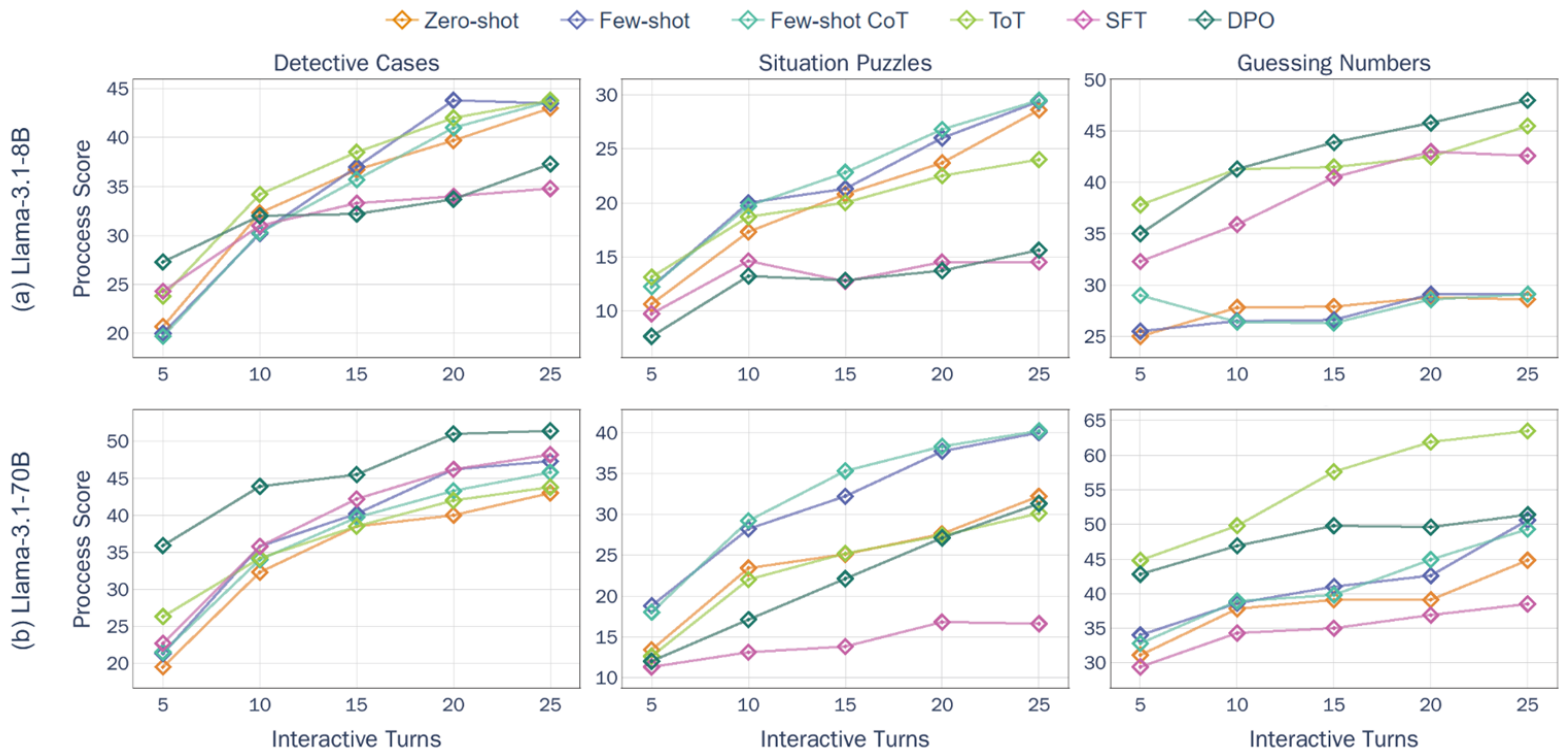
Observation 3:
Existing active reasoning methods **fail** in AR-Bench.

Observation 4:
Human baselines significantly surpass cutting-edge language models.





Empirical Findings (Process Score)



Observations 5, 6, 7:
LLMs struggle to consistently propose **high-quality questions.**
Unreliable verifier (using LLM-as-a-judge in ToT) limits the performance of search methods on tasks with language feedback (DC and SP).

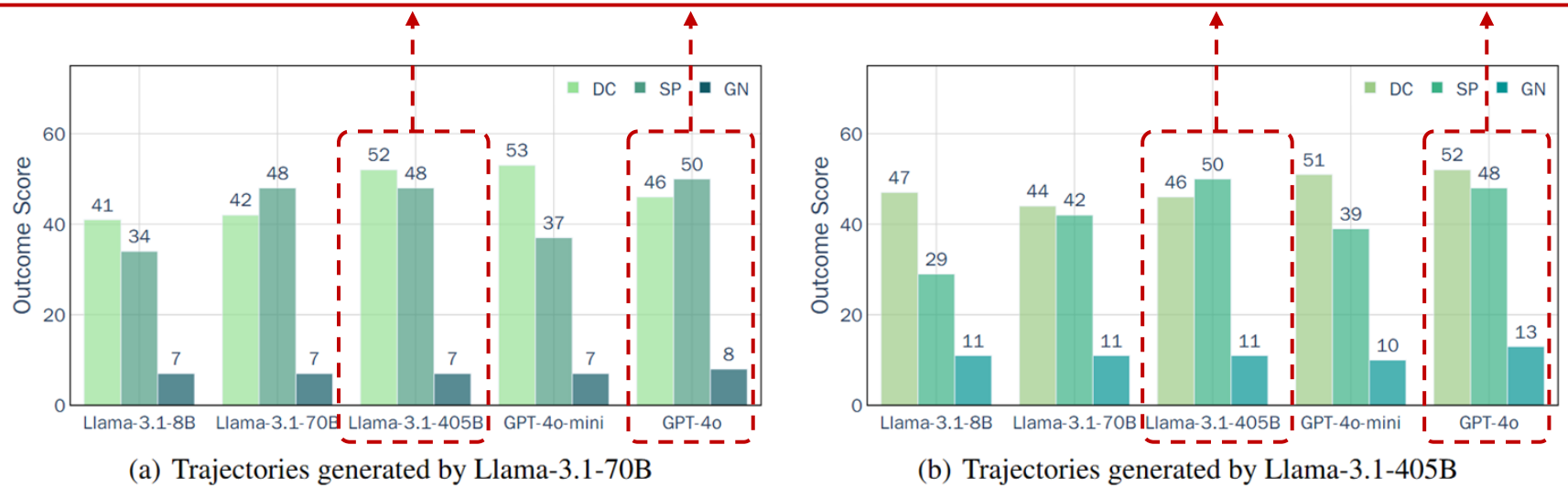
Empirical Findings (Process Score)

Observation 8: Underperforming LLMs ask **low-quality** questions.

Observation 9: **Larger models** can retrieve **more useful information** by proposing questions in interactions.

Empirical Findings (Process Score)

Observation 10: Given process conversations (trajectories) and directly ask for the solution, **larger models** demonstrate **higher robustness** to **insufficient information** (on average).



(a) Trajectories generated by Llama-3.1-70B

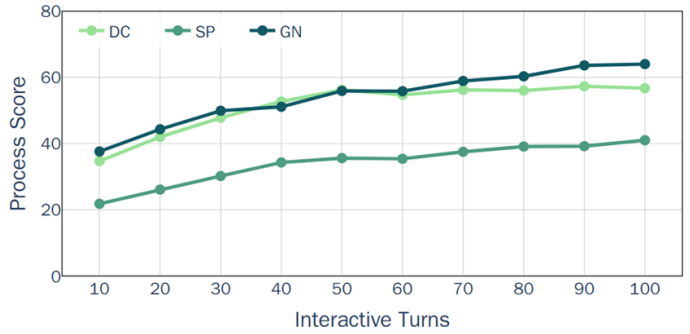
(b) Trajectories generated by Llama-3.1-405B

Note: We collect the QA record of 70B/405B model,
and feed the QA record to different models to generate solutions.

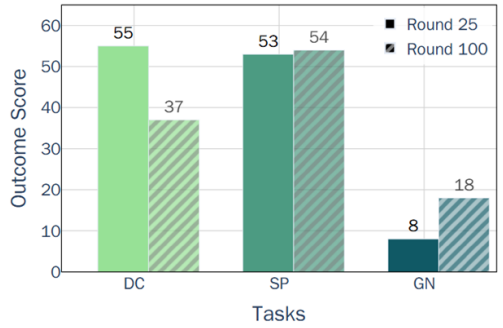
Empirical Findings (Ablation Study)

The question-asking scaling effect in AR

Observation 11: Question-asking **scaling cannot fully solve** the active reasoning tasks.



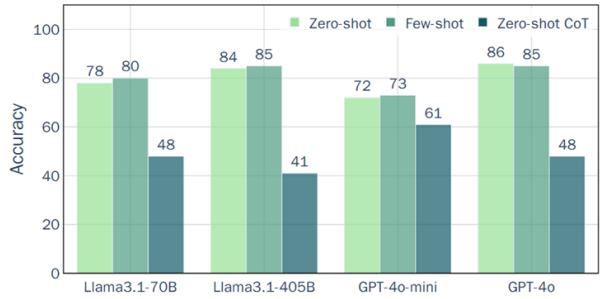
(a) Process Score



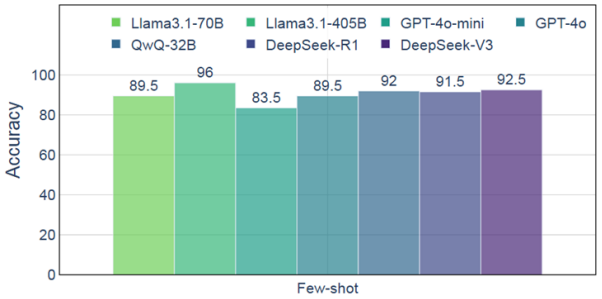
(b) Outcome Score

Verifying the reliability of the LLM judge

Observation 12: Larger models, as the judge, tend to provide more **reliable responses.**



(a) TurtleBenchmark



(b) AR-Bench

Empirical Findings (Case Study)

Incorrect case in DC

A Distinguished historian Dr. Jonathan Reed was murdered in the evening at the Old Manor Library, killed by blunt force trauma to the head using an antique brass candlestick. ...

Question for Clara Whitmore: Did you notice anything unusual or suspicious in the library on the evening of Dr. Reed's murder?
Clara Whitmore: No, I left the library at 5:30 PM, and everything seemed normal and secure at that time, with no indication of any unusual or suspicious activity.

Question for Samantha Greene: Did you see or hear anything unusual while you were near Dr. Reed's office in the library?
Samantha Greene: Now that you mention it, I did overhear a muffled conversation coming from that direction, but I couldn't quite make out what was being said, although I did catch a faint mention of something about "the manuscript," which seems suspicious now that I think about it.

Observation 13:
Models ask broad and non-specific questions.

Incorrect case in SP

Puzzle: Sarah found herself locked out of her own house without ever leaving the room she was in. How could this happen?

Q: Was Sarah in a bathroom when she found herself locked out?
A: No

Q: Was Sarah in a basement or attic when she found herself locked out?
A: Unknown

Observation 14:
Models ask repeated or unhelpful questions.

Incorrect case in GN

Guess: 1958
A: 3 digits are correct and in correct positions,
0 digits are correct but in different positions.

Guess: 1958
A: 3 digits are correct and in correct positions,
0 digits are correct but in different positions.

Task	Error Pattern	Model	
		Llama-3.1-8B	GPT-4o
DC	Timeline Misinterpretation	10%	31%
	Evidence Overlooked	61%	15%
SP	Evidence Overlooked	36%	44%
	Unsupported Assumptions	90%	72%
GN	Feedback Misunderstanding	78%	61%
	Incomplete Testing	81%	55%

Table 3: Error pattern analysis for AR-Bench. Proportions indicate the frequency of specific error types in error cases.

Observation 15 (Error Patterns):
Timeline Misinterpretation: models request information about irrelevant periods.
Evidence Overlooked: models fail to identify critical evidence needed to uncover the truth.
Unsupported Assumptions: models introduce fabricated details in their conclusions.
Feedback Misunderstanding: models fail to track the correct and eliminate the incorrect with feedback.
Incomplete Testing: models fail to identify the position of a correct digit.

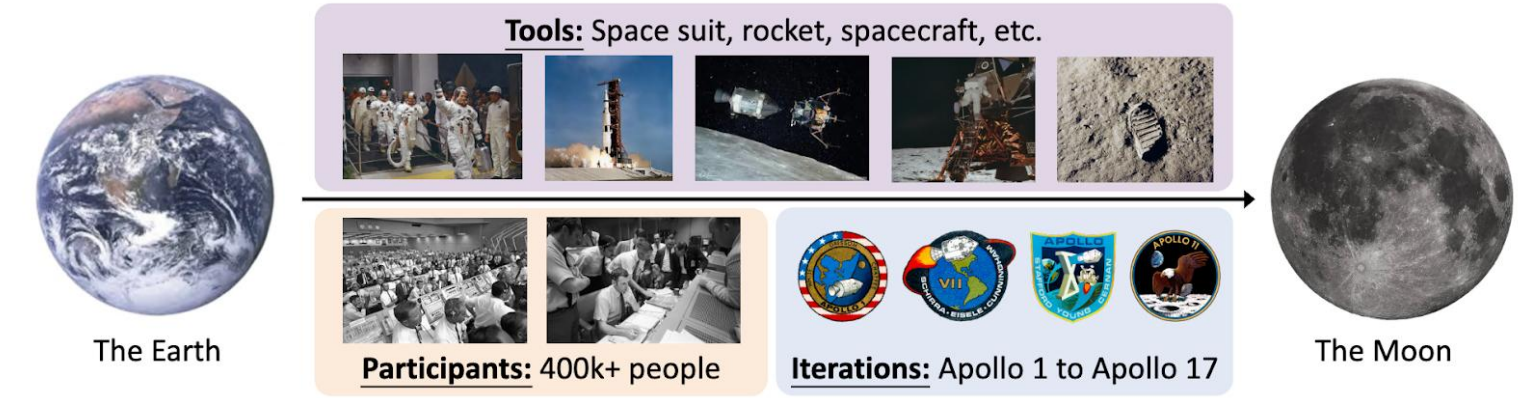
Take Home Messages

We construct **AR-Bench** to systematically evaluate the active reasoning capability of prevailing LLMs, and reveal **a general vulnerability** of LLMs in solving AR tasks.

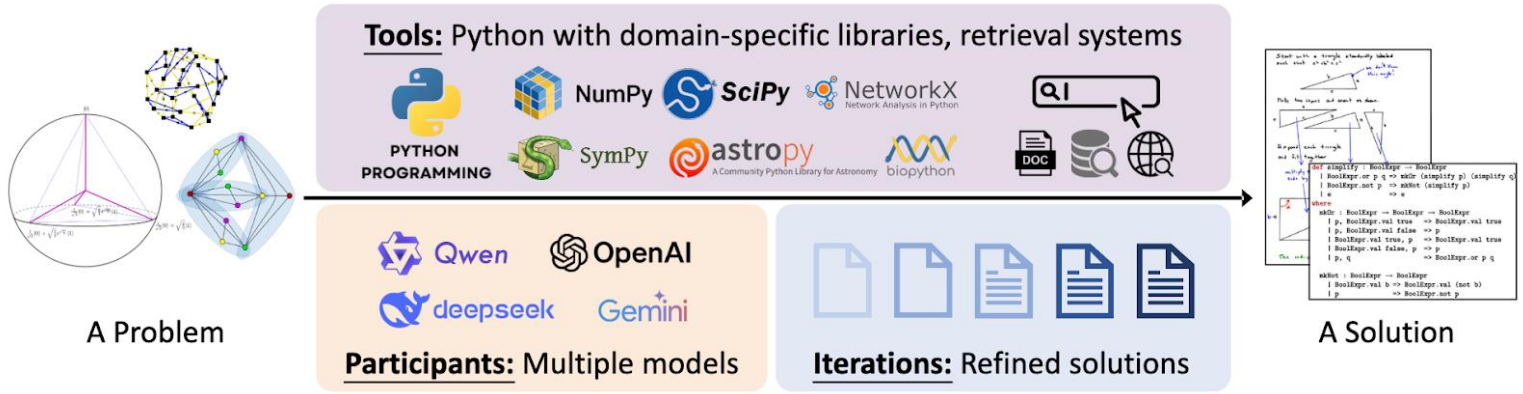
- Active reasoning demonstrates challenges across **all models and methods.**
- Human reasoners significantly **surpass** cutting-edge LLMs.
- LLMs struggle to consistently ask **high-quality questions** to retrieve critical information across long-horizon conversations.

AlphaApollo: A System for Deep Agentic Reasoning

- Orchestrating Foundation Models and Professional Tools into a Self-Evolving System.



(a) The Apollo Program (in 1960s) for moon landing with humans

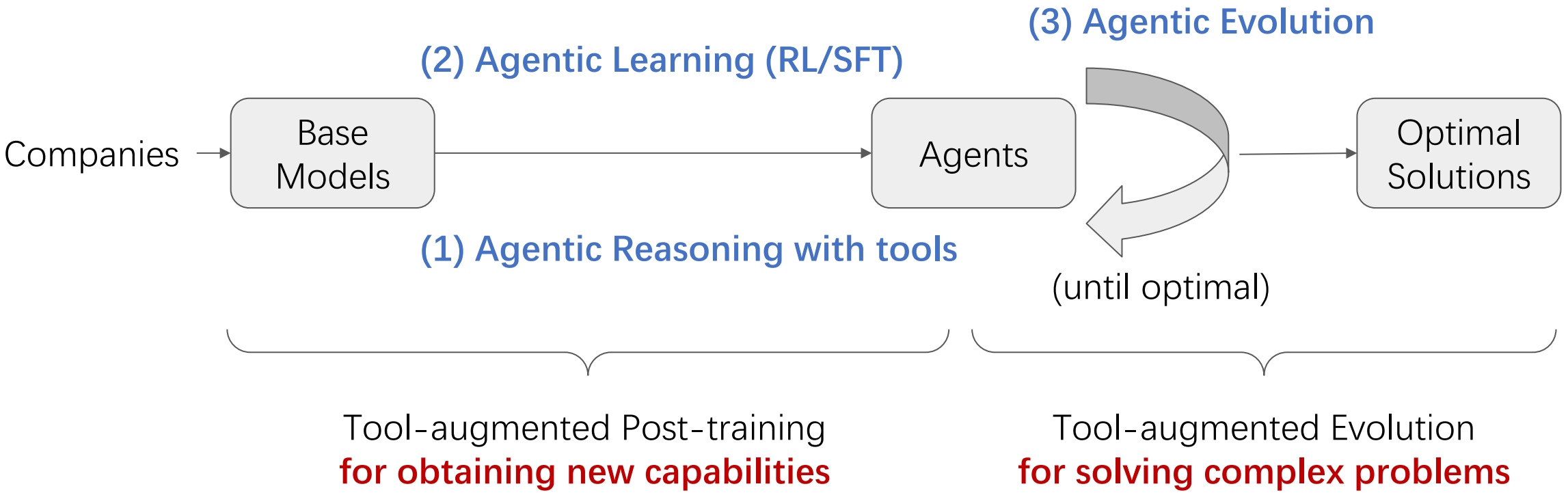


(b) The AlphaApollo System (ours) for problem solving with foundation models

Project homepage: https://alphaapollo.org/.

An Overview of AlphaApollo

- AlphaApollo provides a **unified platform** of agentic **reasoning**, **learning**, and **evolution**.



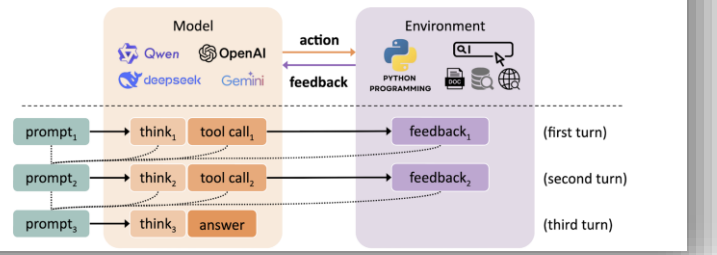
Project homepage: https://alphaapollo.org/.

An Overview of AlphaApollo

Feature 1: Agentic Reasoning

Agentic Reasoning (multi-turn interaction between model and environment).

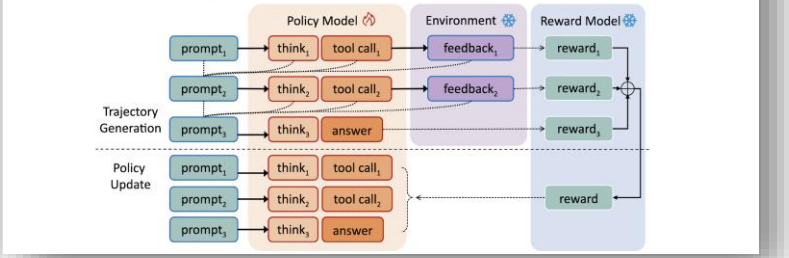
- Given the prompt, the model generate output (think/tool call/answer tokens).
- The environment parses output, executes tool, and give feedback to the model.



Feature 2: Agentic Learning

Agentic Learning (multi-turn optimization on the output of model).

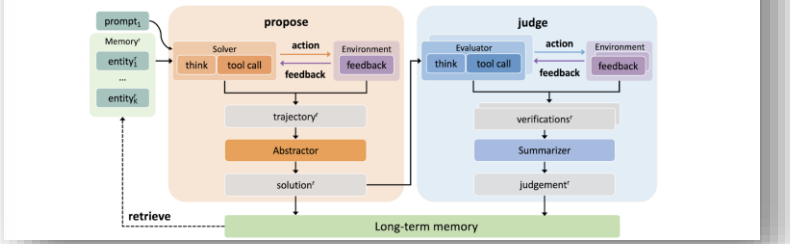
- Incorporates VeRL into a stable, turn-level agentic learning.
- Supports multiple algorithms (e.g., GRPO/SFT) and models (e.g., Qwen).



Feature 3: Agentic Evolution

Agentic Evolution (a test-time mechanism to evolve solutions).

- Operates through a propose-judge-update loop of multi-round evolution.
- Long-term memory to enable long-horizon evolution.
- Parallel (distributed) evolution to support scalable evolution with multiple mode



Quick Start Guidance

Installation

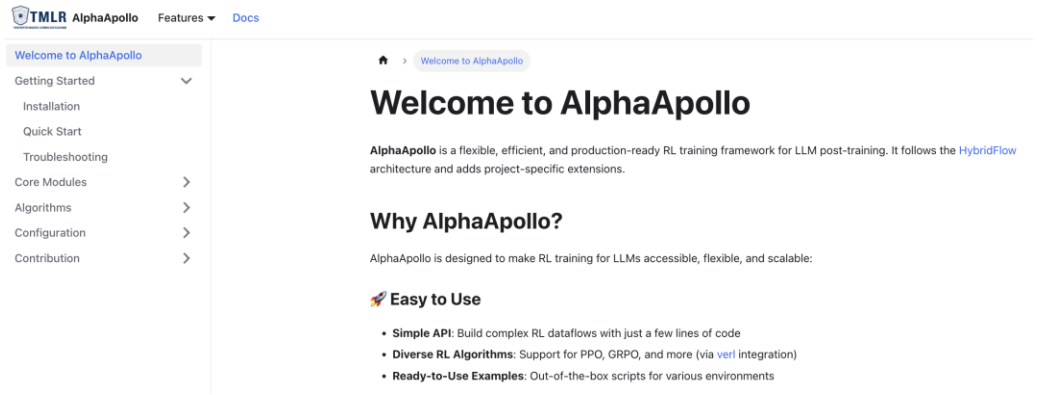
BASH Copy

```
conda create -n alphaapollo python==3.12 -y
conda activate alphaapollo

git clone https://github.com/tmlr-group/AlphaApollo.git
cd AlphaApollo

bash installation.sh
```

Document & Tutorial



Project homepage: https://alphaapollo.org/.

Demonstrations of AlphaApollo

Code: https://github.com/tmlr-group/AlphaApollo

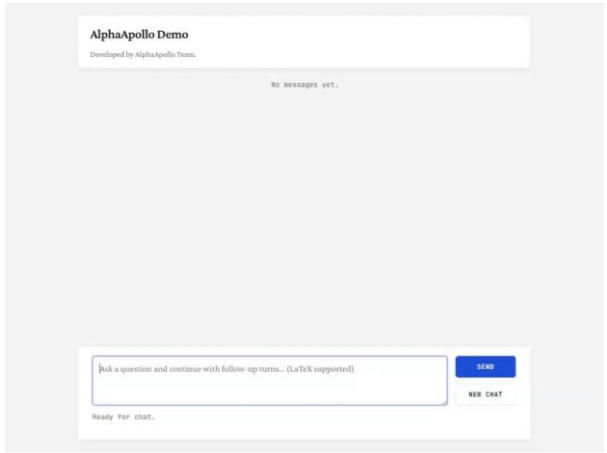
Host a model:

```
CUDA_VISIBLE_DEVICES=0,1,2,3
python -m
vllm.entrypoints.openai.api_server
\
--model
/data1/models/hub/Qwen2.5-
72B-Instruct \
--tensor-parallel-size 4 \
--port 8000
```

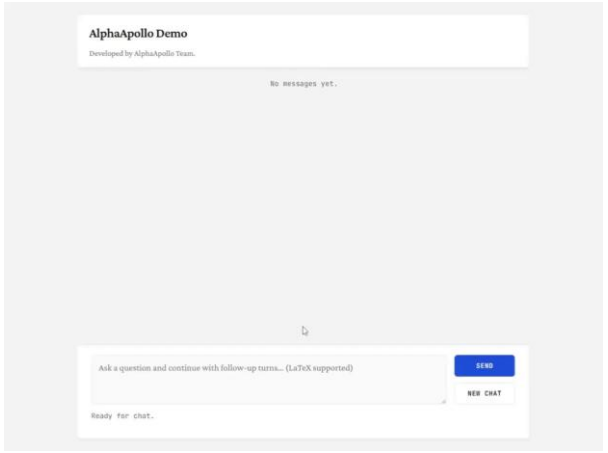
Run local web UI:

```
MODE=web
VLLM_MODEL="/data1/models/h
ub/Qwen2.5-72B-Instruct" \
bash
examples/demo/run_terminal_de
mo_vllm.sh
```

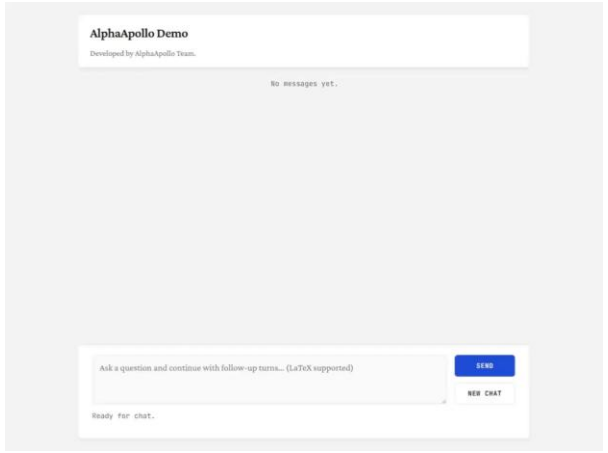
QUESTION 1



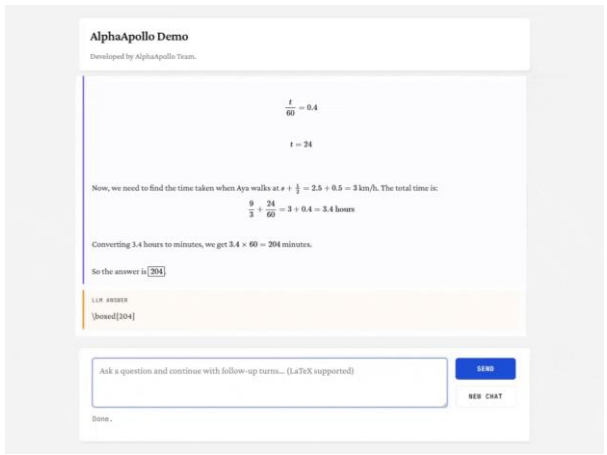
QUESTION 2



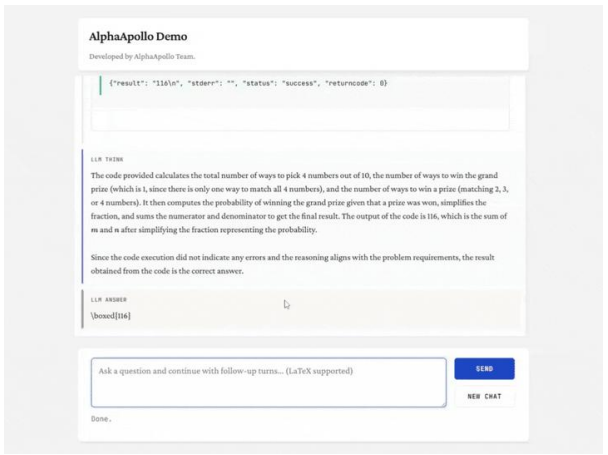
QUESTION 3



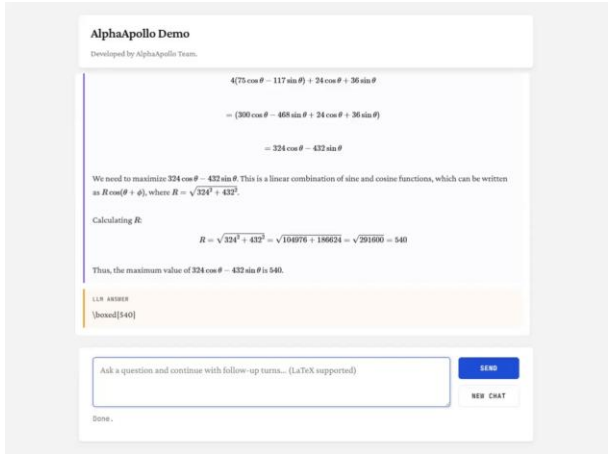
FOLLOW UP 1



FOLLOW UP 2



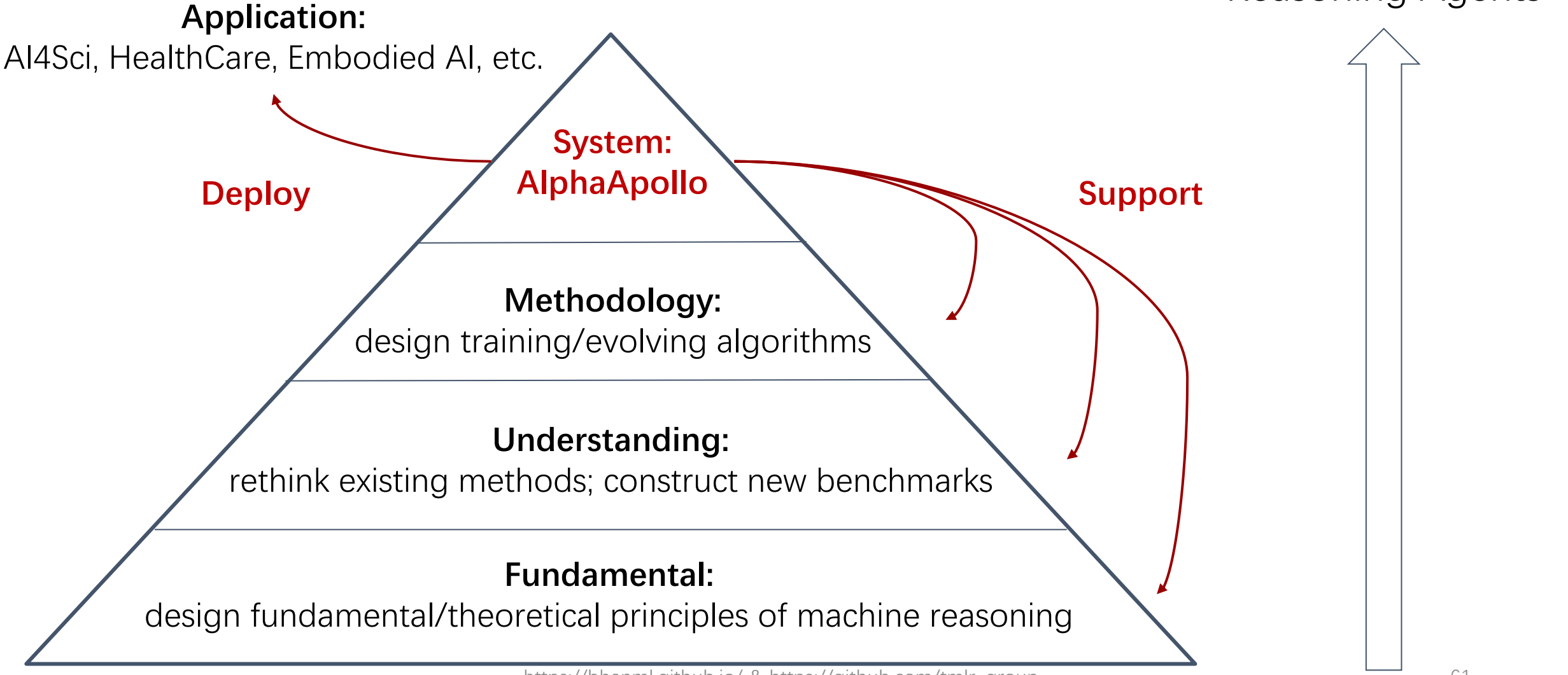
FOLLOW UP 3



Project homepage: https://alphaapollo.org/.

The Research Scope of AlphaApollo

Towards Trustworthy Reasoning Agents



Project homepage: https://alphaapollo.org/.

Future Directions

Robust pre-training/fine-tuning methods are required for VLMs.

- VLMs can still be mislead by spurious features.
- Larger models and high-quality data lead to better robustness.

The trade-off between unlearning and retention remains a critical issue.

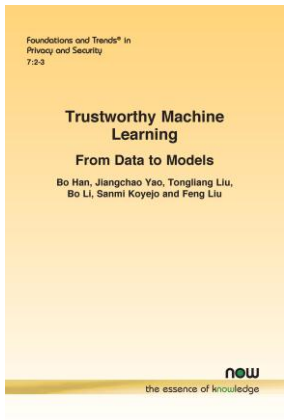
- Current unlearning objectives all have negative impacts on retention.
- Data and optimization aspects of unlearning are not well explored.

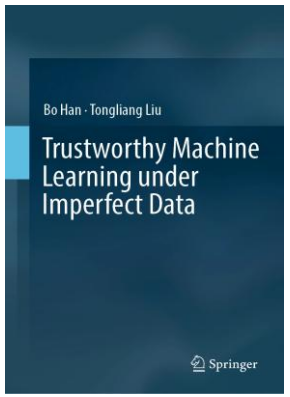
The broader tasks of active reasoning and corresponding methods are necessary.

- Broader tasks of active reasoning (especially agentic scenarios) can be further investigated.
- Advanced post-training methods to enhance active reasoning are under-explored.

Appendix

- Survey:
  - A Survey of Label-noise Representation Learning: Past, Present and Future. arXiv, 2020.
- Book:
  - Machine Learning with Noisy Labels: From Theory to Heuristics. Adaptive Computation and Machine Learning series, **The MIT Press**, 2026.
  - Trustworthy Machine Learning under Imperfect Data. CS series, **Springer Nature**, 2025.
  - Trustworthy Machine Learning: From Data to Models. **Foundations and Trends® in Privacy and Security**, 2025.





- Tutorial and Lecture:
  - AAAI 2024 Tutorial on Trustworthy Machine Learning under Imperfect Data
  - IJCAI 2024 Tutorial on Trustworthy Machine Learning under Imperfect Data
  - WWW 2025 Tutorial on Trustworthy AI under Imperfect Web Data
  - AAAI 2026 Tutorial on The Science and Practice of Machine Unlearning for AI Safety
  - AAAI 2026 Tutorial on Trustworthy Machine Reasoning with Foundation Models
  - DeepLearn 2026 Lecture on Trustworthy Machine Learning from Data to Models
- Workshops:
  - IJCAI 2021 Workshop on Weakly Supervised Representation Learning
  - ACML 2022 Workshop on Weakly Supervised Learning
  - RIKEN 2023 Workshop on Weakly Supervised Learning
  - HKBU-RIKEN AIP 2024 Joint Workshop on Artificial Intelligence and Machine Learning

https://bhanml.github.io & https://github.com/tmlr-group