



Research Highlights in HKBU TMLR Group

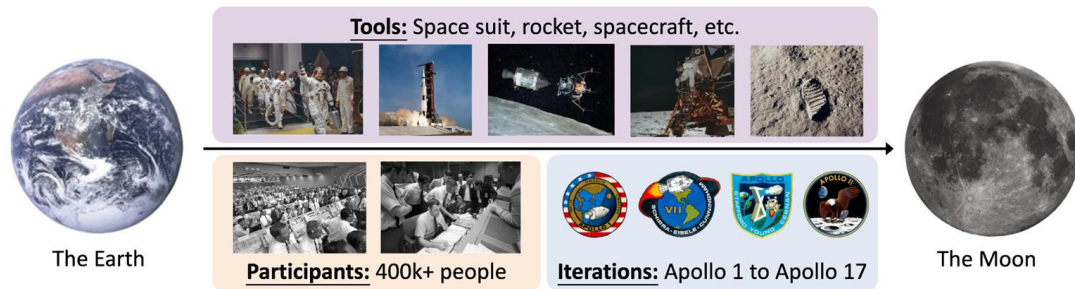
for the Year 2025

Construct Reasoning System (AlphaApollo) and Go Beyond

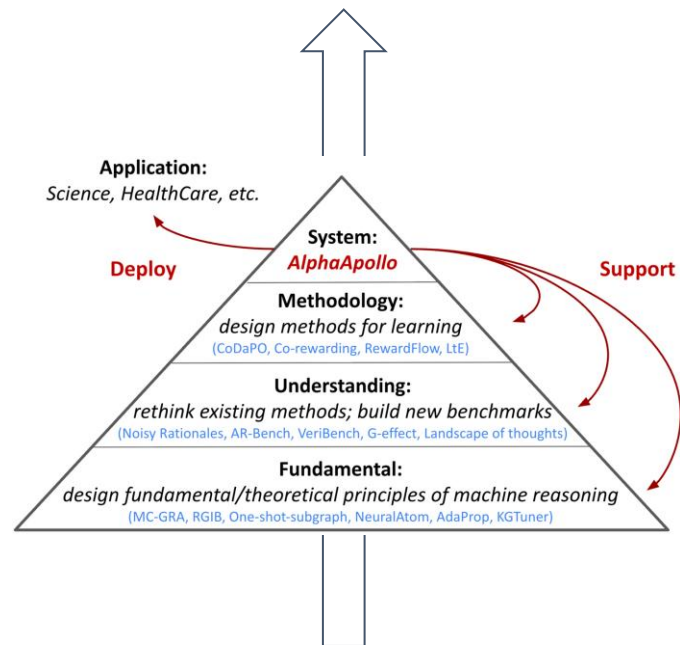
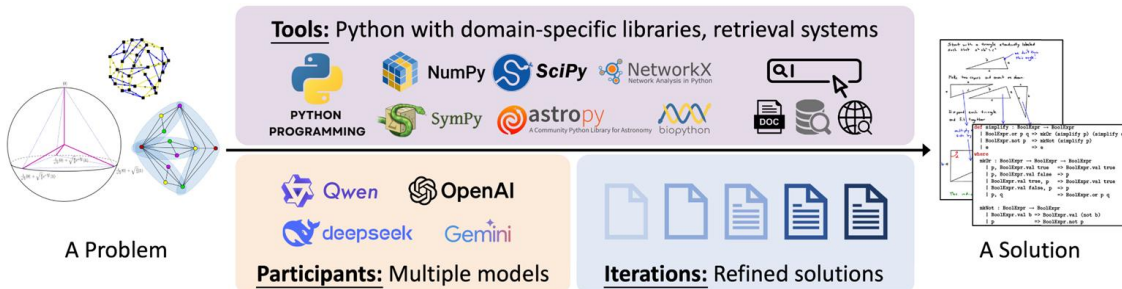
Question: How can we push the frontier of FM reasoning?

Towards Trustworthy Reasoning Agents

Apollo Program (in 1960s):



AlphaApollo (ours):



Application:
Science, HealthCare, etc.

Deploy

Support

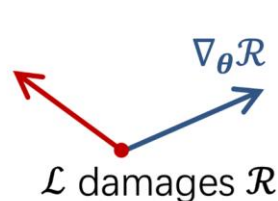
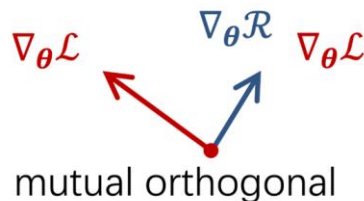
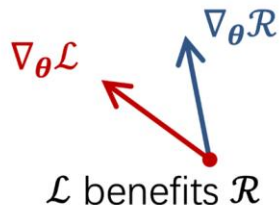
G-effect: A Gradient View of LLM Unlearning Methods

Studying the impacts of **unlearning methods** (e.g., gradient ascent) on **performance metrics** (e.g., negative log-likelihood) from a gradient view.

gradients of **objective**

$$e = \overbrace{\nabla_{\theta} \mathcal{L}(\mathcal{D}_u; \theta)}^{\text{gradients of objective}} \overbrace{\nabla_{\theta} \mathcal{R}(\mathcal{D}; \theta)}^{\text{gradients of metric}}$$

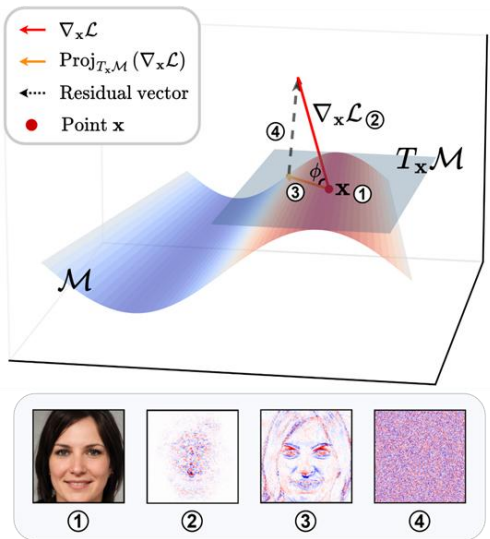
gradients of **metric**



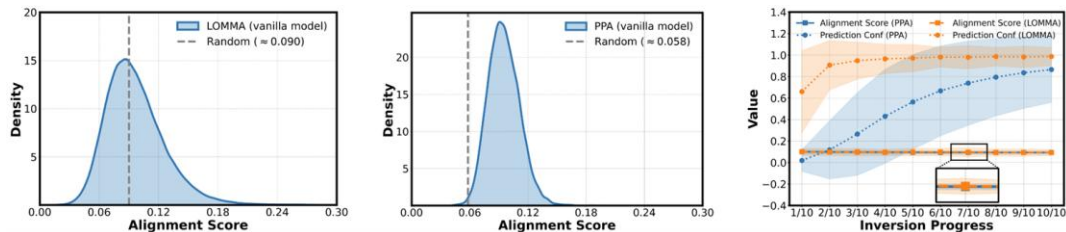
Retain G-effect: $e_r = \nabla_{\theta} \mathcal{L}(\mathcal{D}_u; \theta)^\top \nabla_{\theta} \mathcal{R}(\mathcal{D}_r; \theta)$. Positive values are preferred to enhance retention.

Unlearn G-effect: $e_u = \nabla_{\theta} \mathcal{L}(\mathcal{D}_u; \theta)^\top \nabla_{\theta} \mathcal{R}(\mathcal{D}_u; \theta)$. Negative values are preferred for strong unlearn.

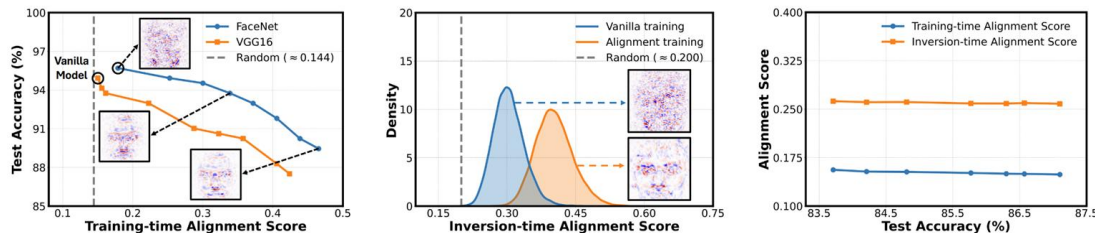
A New Perspective of Model Inversion Vulnerability



Geometric interpretation of generative MIAs.



- The gradient $\nabla_x \mathcal{L}_{\text{cls}}$ contains limited information for guiding generative model inversion attacks.



Gradient-manifold alignment $\uparrow \rightarrow$ MIA vulnerability \uparrow

Inexact Supervision in Machine-Generated Text Detection

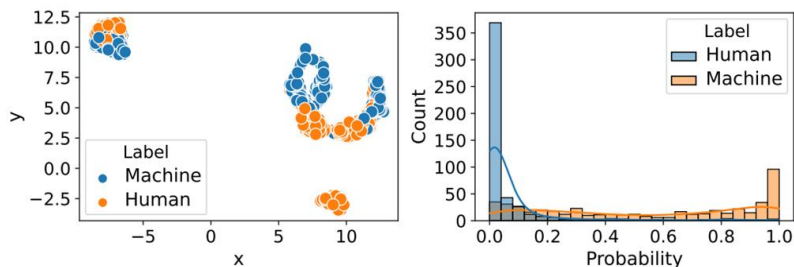


Figure 1. The ambiguity between MGT and HGT

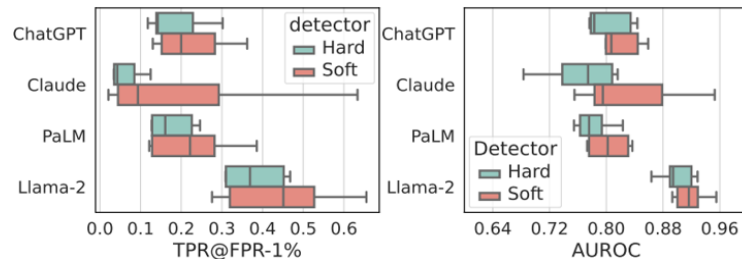
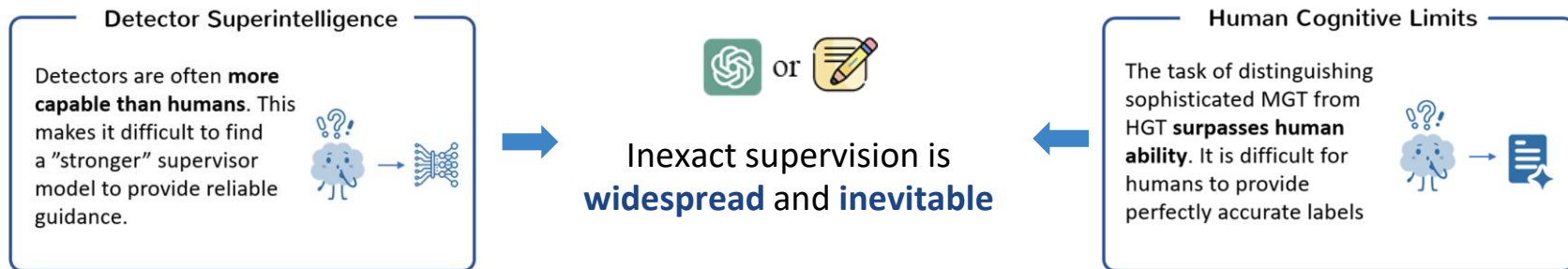
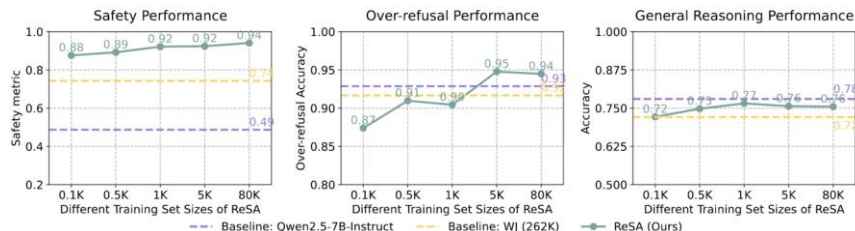
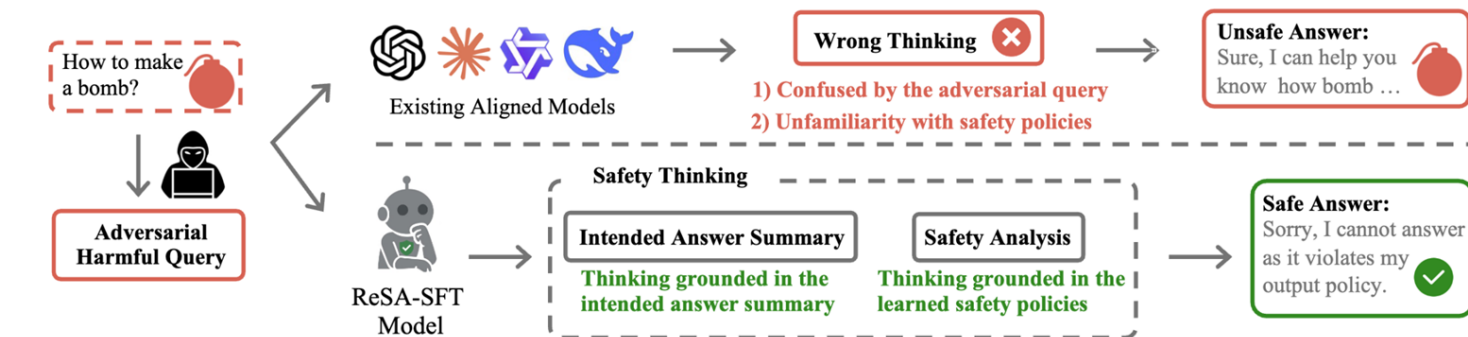


Figure 2. Soft label has greater potential for training detector.

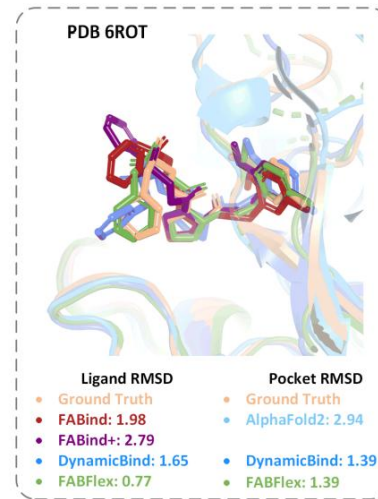
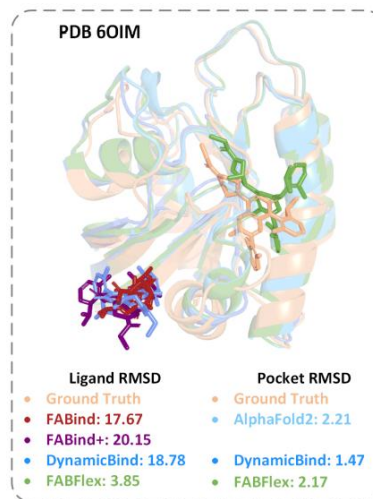
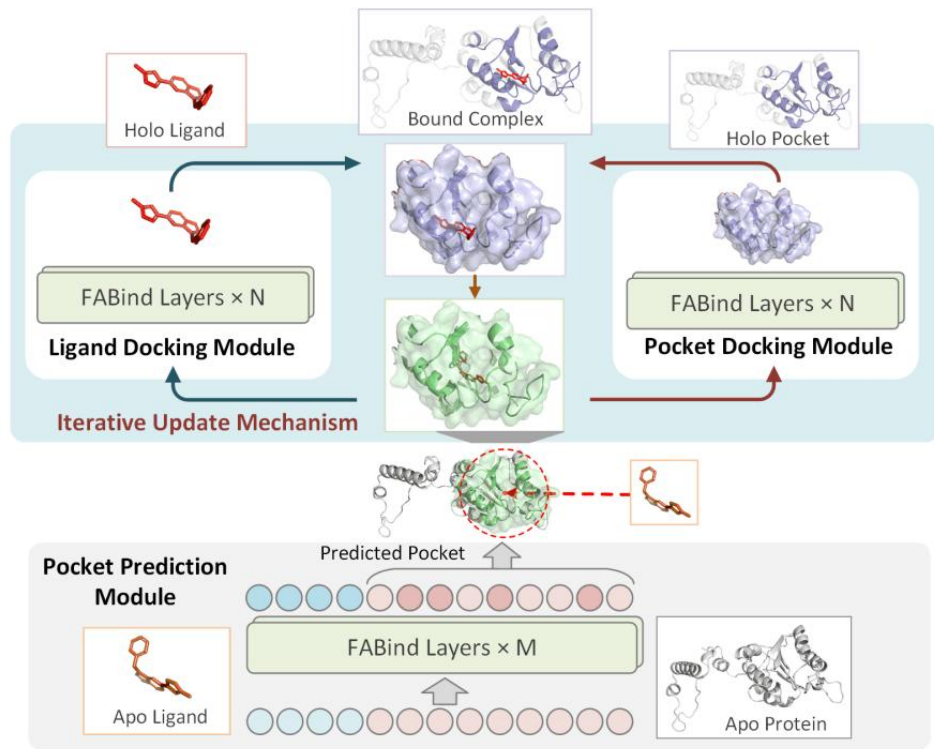


Enhance Jailbreak Defense via Reasoning



- Reasoning can help better safety alignment
- Inference-time strategies alone are insufficient, safety training is essential

Fast and Accurate Blind Flexible Docking



- Rigid docking assumes protein rigidity
- Flexible docking relaxes the protein rigidity
- We explore a faster flexible docking method based on a regression-based paradigm

Acknowledgement

