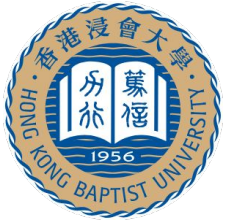


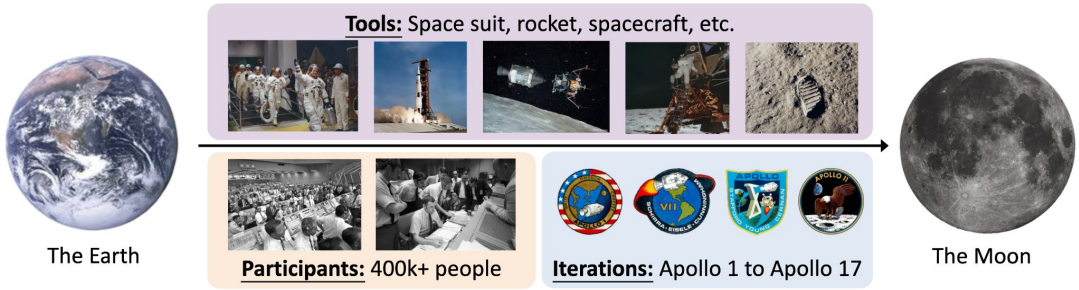
Research Highlights in HKBU TMLR Group

for the Year 2025

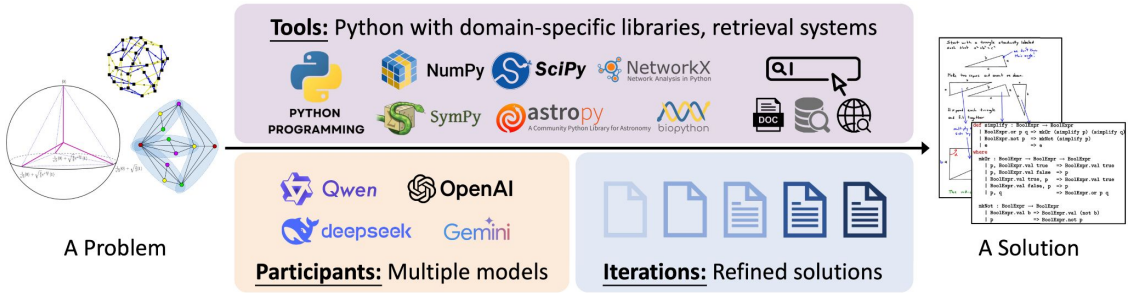
Construct Reasoning System (AlphaApollo) and Go Beyond

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

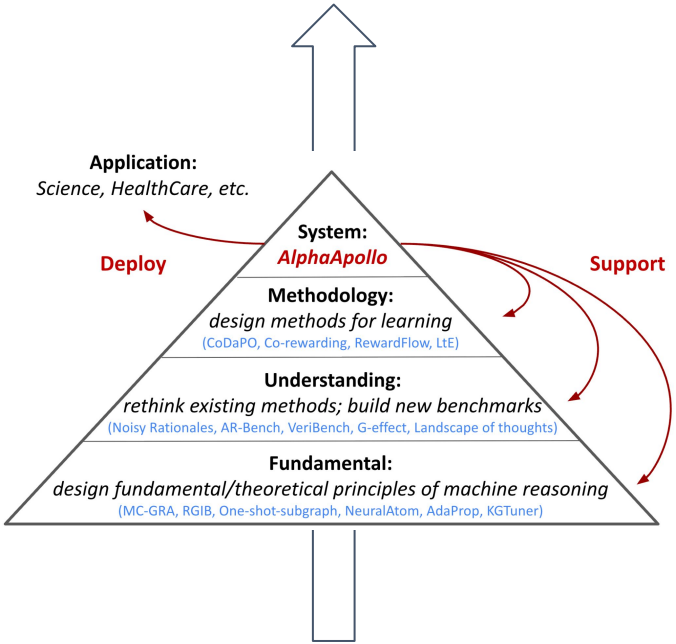
Apollo Program (in 1960s):



AlphaApollo (ours):



Towards Trustworthy Reasoning Agents



AlphaApollo: Orchestrating Foundation Models and Professional Tools into a Self-Evolving System for Deep Agentic Reasoning. *Arxiv* preprint, 2025.

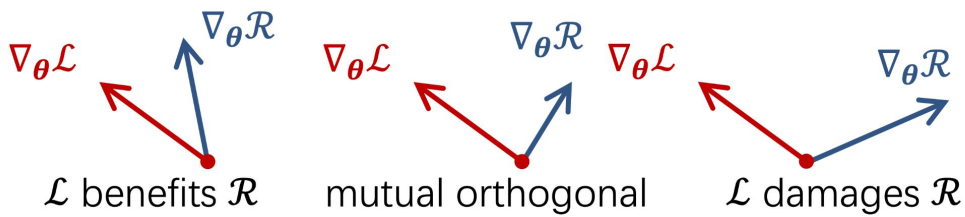
G-effect: A Gradient View of LLM Unlearning Methods

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**

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

gradients of **metric**

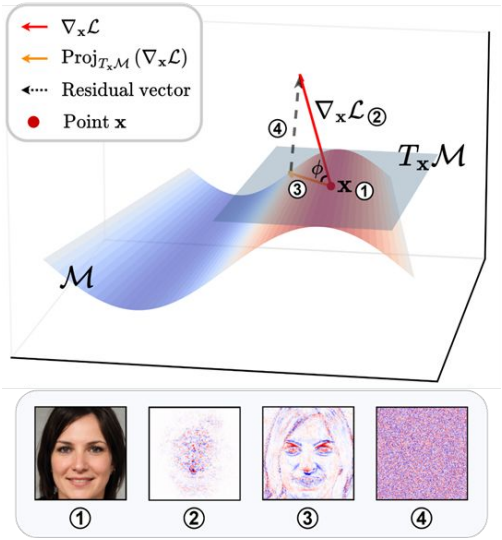


Retain G-effect: $e_{\mathrm{r}} = \nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathcal{D}_{\mathrm{u}};\boldsymbol{\theta})^{\top}\nabla_{\boldsymbol{\theta}}\mathcal{R}(\mathcal{D}_{\mathrm{r}};\boldsymbol{\theta})$. Positive values are preferred to enhance retention.

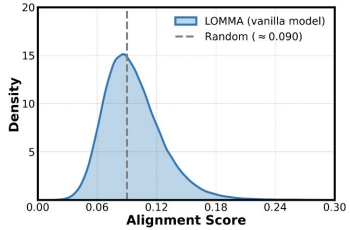
Unlearn G-effect: $e_{\mathrm{u}} = \nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathcal{D}_{\mathrm{u}};\boldsymbol{\theta})^{\top}\nabla_{\boldsymbol{\theta}}\mathcal{R}(\mathcal{D}_{\mathrm{u}};\boldsymbol{\theta})$. Negative values are preferred for strong unlearn.

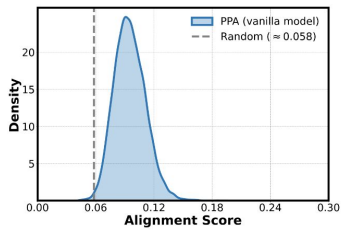
Rethinking LLM Unlearning Objectives: A Gradient Perspective and Go Beyond. In *ICLR*, 2025.

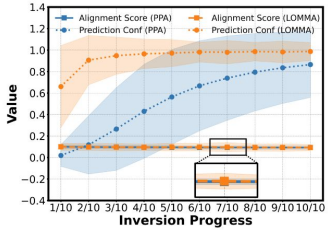
A New Perspective of Model Inversion Vulnerability



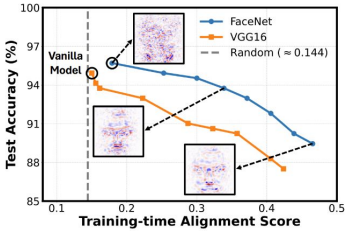
Geometric interpretation of generative MIAs.

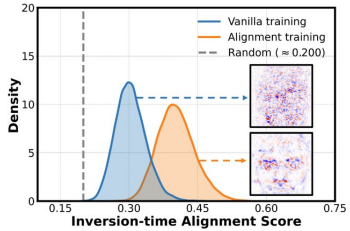


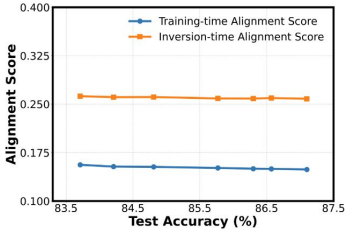




- The gradient $\nabla_{\mathbf{x}}\mathcal{L}_{\text{cls}}$ contains limited information for guiding generative model inversion attacks.







Gradient-manifold alignment ↑ ➔ MIA vulnerability ↑

Generative Model Inversion Through the Lens of the Manifold Hypothesis. In *NeurIPS*, 2025.

Inexact Supervision in Machine-Generated Text Detection

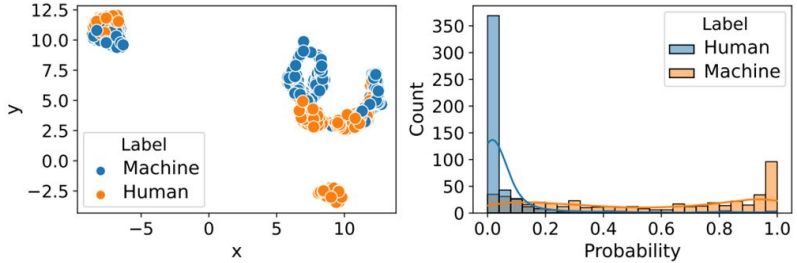


Figure 1. The ambiguity between MGT and HGT

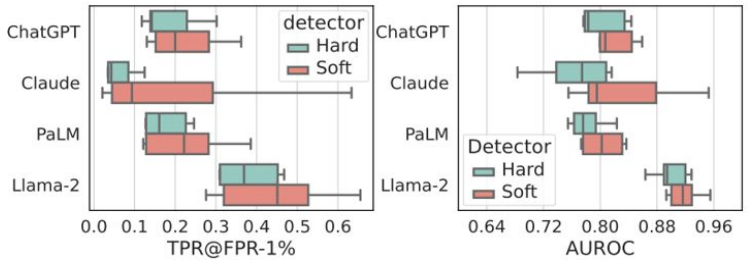


Figure 2. Soft label has greater potential for training detector.

Detector Superintelligence

Detectors are often **more capable than humans**. This makes it difficult to find a "stronger" supervisor model to provide reliable guidance.

or

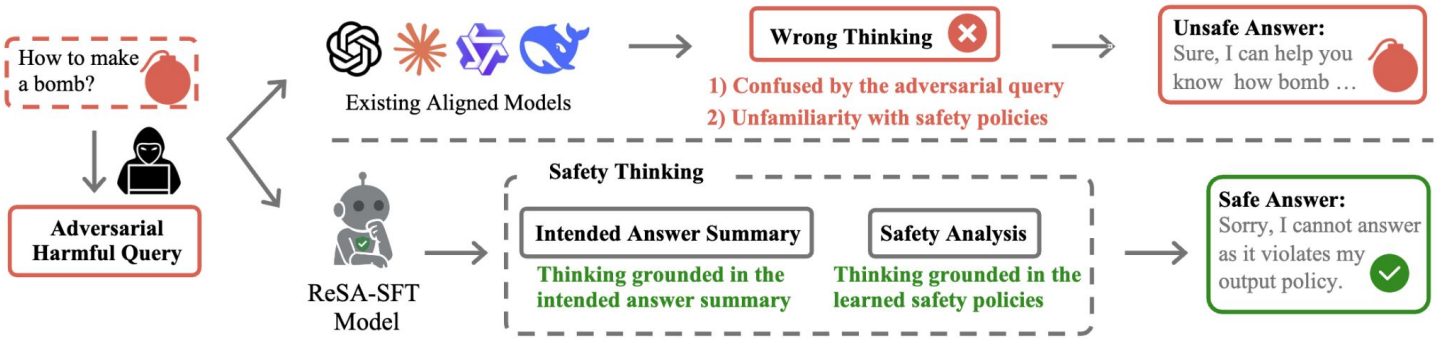
Inexact supervision is **widespread** and **inevitable**

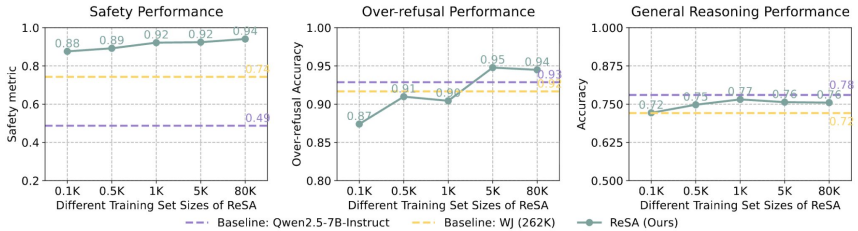
Human Cognitive Limits

The task of distinguishing sophisticated MGT from HGT **surpasses human ability**. It is difficult for humans to provide perfectly accurate labels

Advancing Machine-Generated Text Detection from an Easy to Hard Supervision Perspective. In *NeurIPS*, 2025.

Enhance Jailbreak Defense via Reasoning

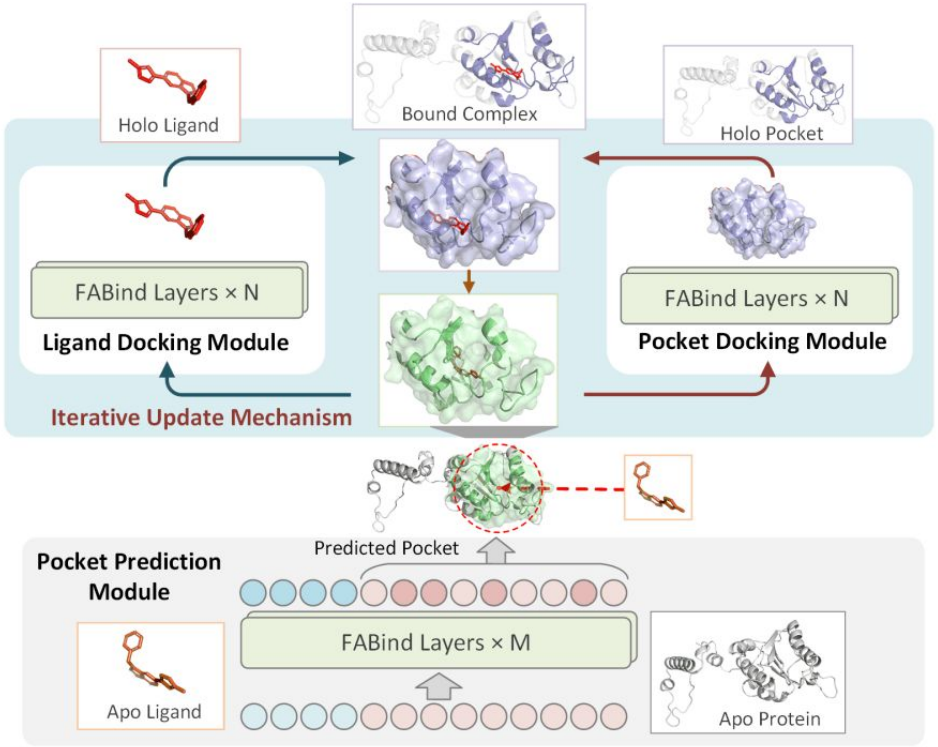


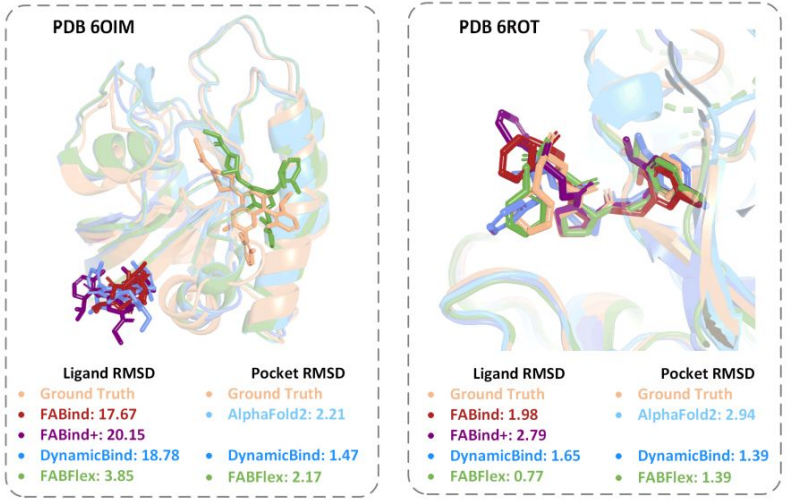


- Reasoning can help better safety alignment
- Inference-time strategies alone are insufficient, safety training is essential

Reasoned Safety Alignment: Ensuring Jailbreak Defense via Answer-Then-Check. In *ICLR*, 2026.

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

Fast and Accurate Blind Flexible Docking. In *ICLR*, 2025.

Acknowledgement