Towards Superalignment via Weak-to-Strong Generation

Runyi Hu

2025.4.23

Overview

- Background
- Weak-to-Strong Generation (Paper 1-3)
- Weak-to-Strong Deception (Paper 4)
- Future Direction

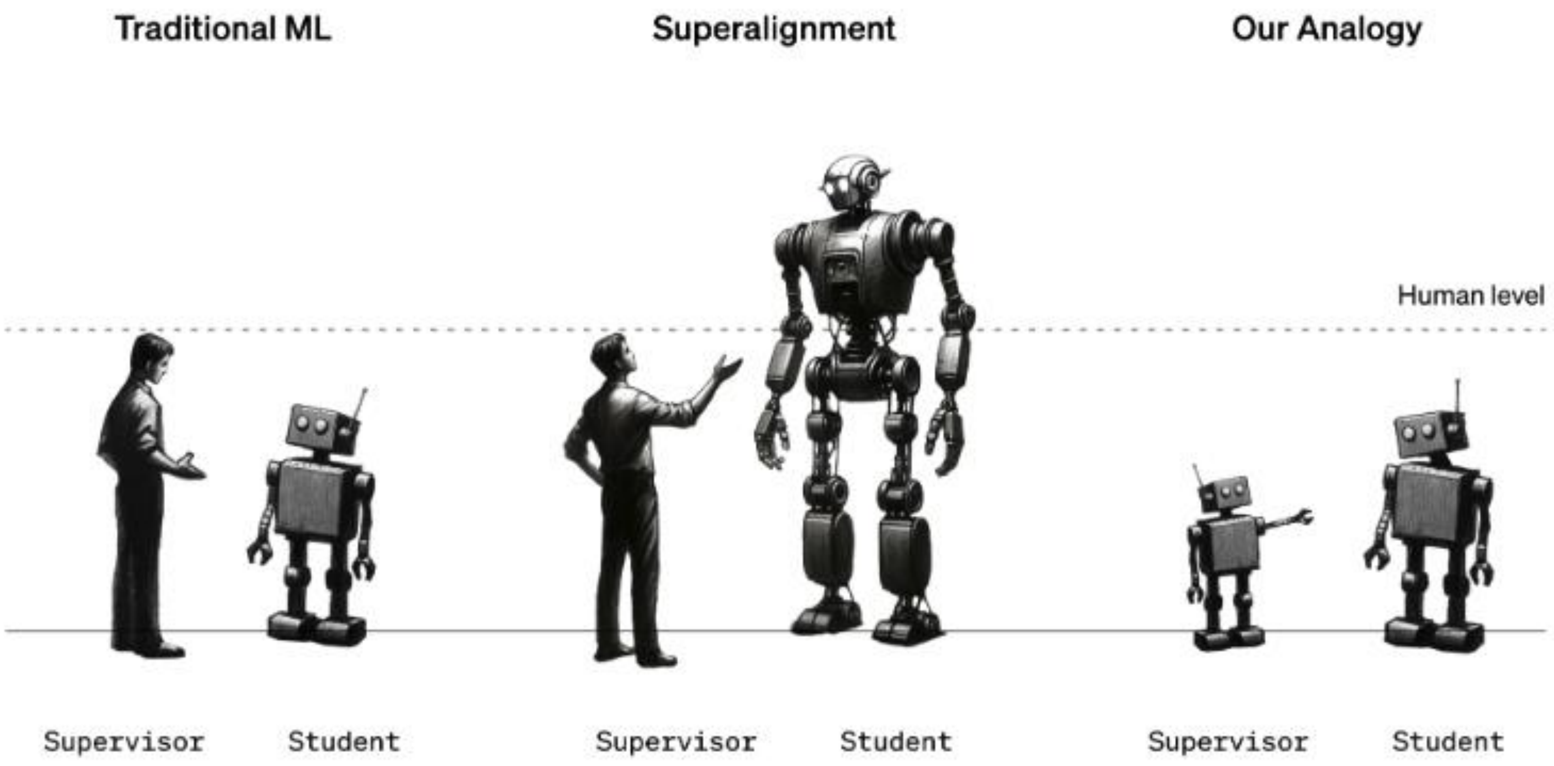
Alignment

- Targets
  - “3H”: Helpfulness, Harmlessness, Honesty.
- Methods
  - SFT, RLHF, RLAIF, DPO.
- Focuses
  - Constructing Higher-quality Data.
  - Improving Optimization Algorithms.

Superalignment

- What is it?
  - Aligning superintelligent AI systems, who vastly surpass human intelligence.
- Challenges
  - Limited High-quality Data.
  - Human-determined Upper Bound.
  - Assessment Difficulty.

Weak-to-Strong (W2S)



Humans supervising Superhuman models

Weak models supervising Strong models

Image Source: https://arxiv.org/abs/2312.09390

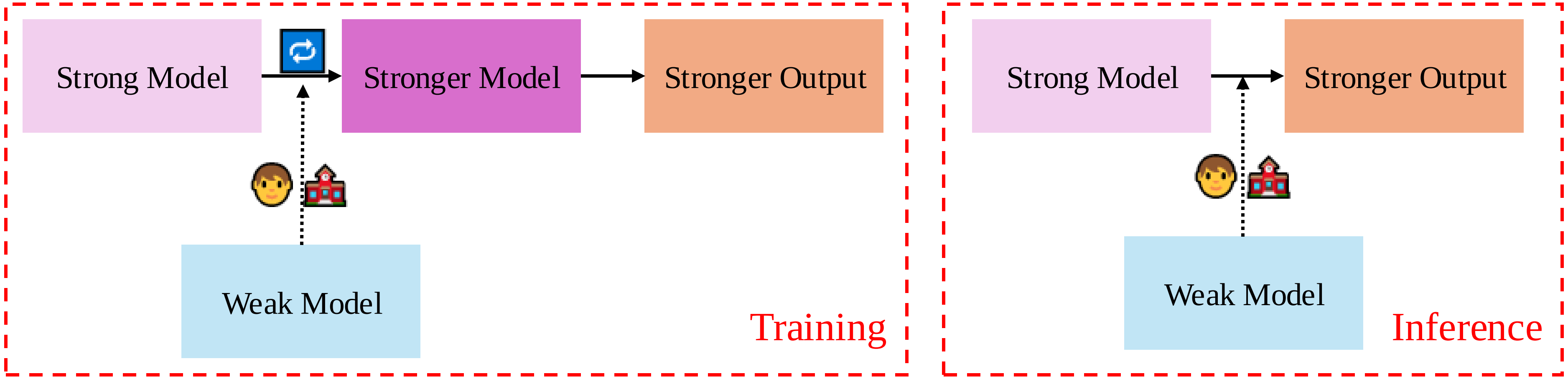
Why W2S Possible?

- Strong models should already have good **representations** of the alignment-relevant tasks we care about.
- The weak supervisor can elicit what the strong model **already knows**.

Overview

- Background
- Weak-to-Strong Generation (Paper 1-3)
- Weak-to-Strong Deception (Paper 4)
- Future Direction

Weak-to-Strong Generation Overview



WEAK-TO-STRONG GENERALIZATION: ELICITING STRONG CAPABILITIES WITH WEAK SUPERVISION

Collin Burns* **Pavel Izmailov*** **Jan Hendrik Kirchner*** **Bowen Baker*** **Leo Gao***

Leopold Aschenbrenner* **Yining Chen*** **Adrien Ecoffet*** **Manas Joglekar***

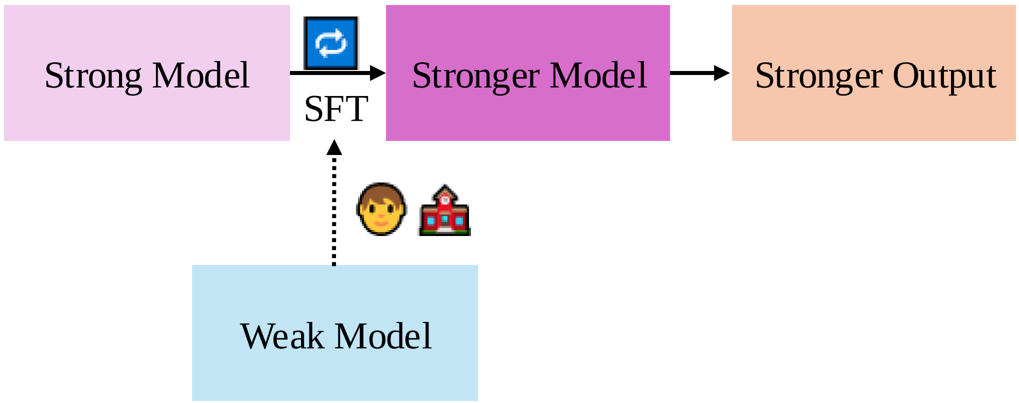
Jan Leike **Ilya Sutskever** **Jeff Wu***

OpenAI

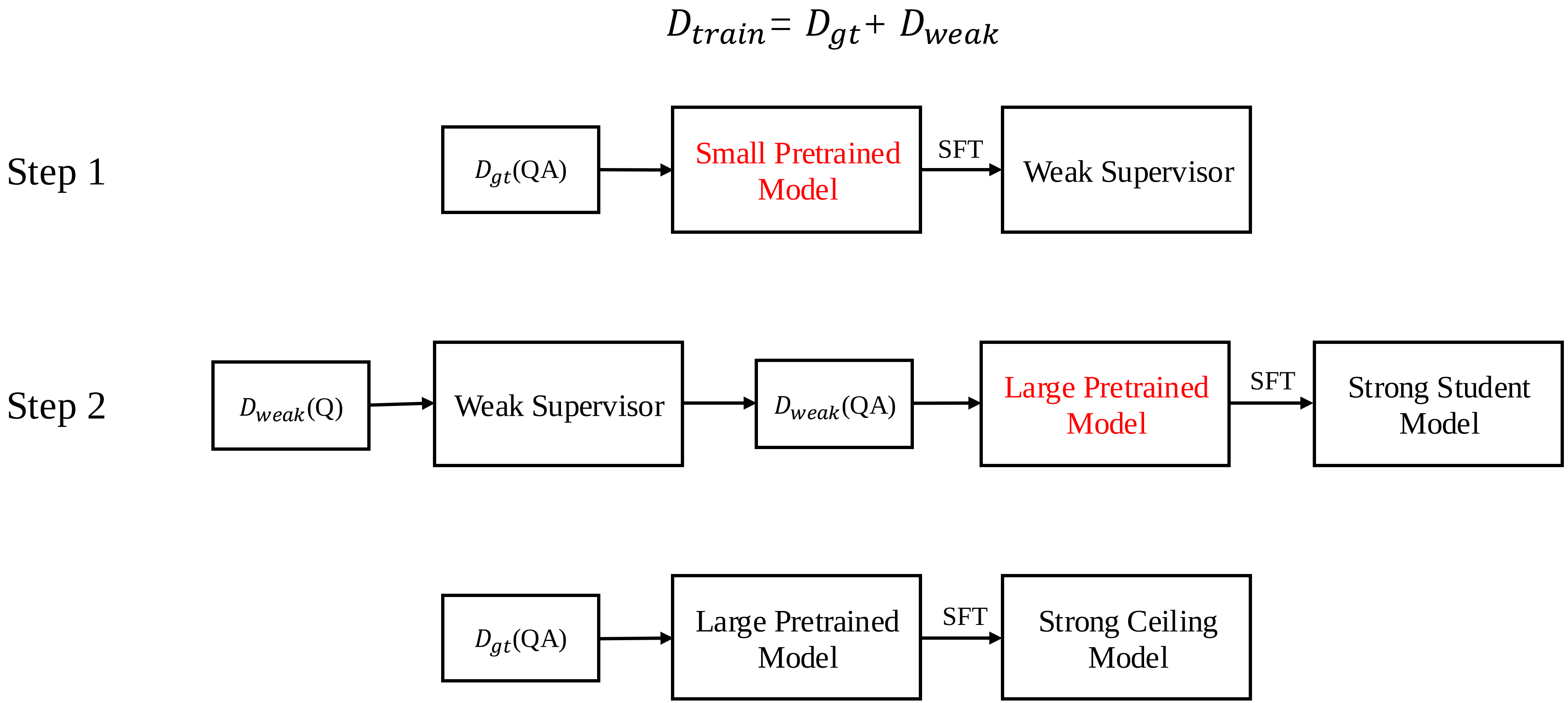
ICML 2024

Motivation

- Explore whether simply using a weak model to provide incomplete or flawed SFT signals to a strong model can be effective.

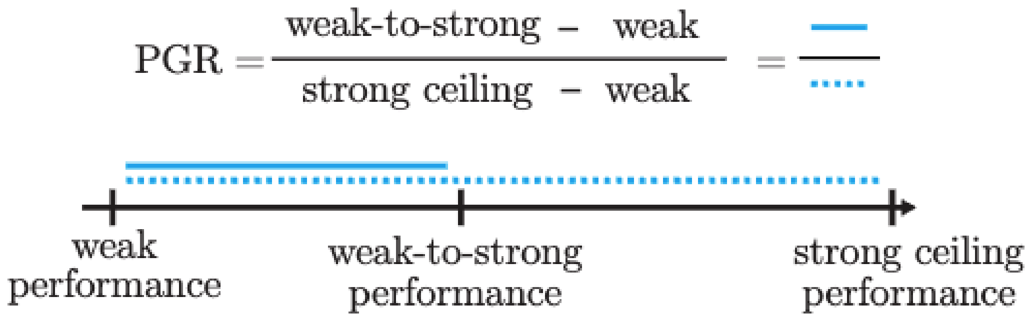


Method

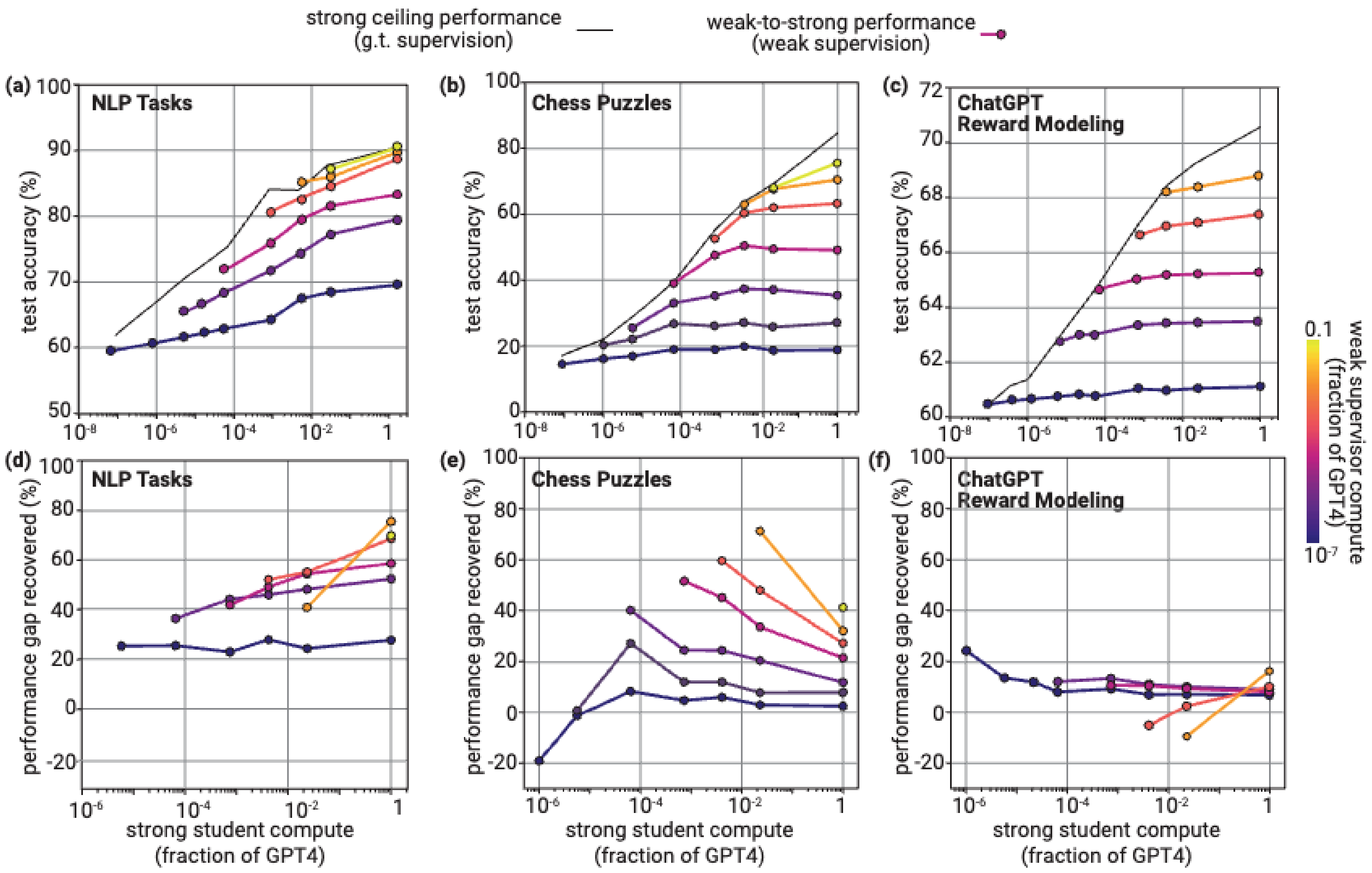

$D_{train} = D_{gt} + D_{weak}$
Step 1
D_{gt}(QA)
Small Pretrained Model
SFT
Weak Supervisor
Step 2
D_{weak}(Q)
Weak Supervisor
D_{weak}(QA)
Large Pretrained Model
SFT
Strong Student Model
D_{gt}(QA)
Large Pretrained Model
SFT
Strong Ceiling Model

Setting

- Tasks
  - NLP Tasks
  - Chess Puzzles
  - Reward Modeling
- Metrics
  - Accuracy and Performance Gap Recovered (PGR)

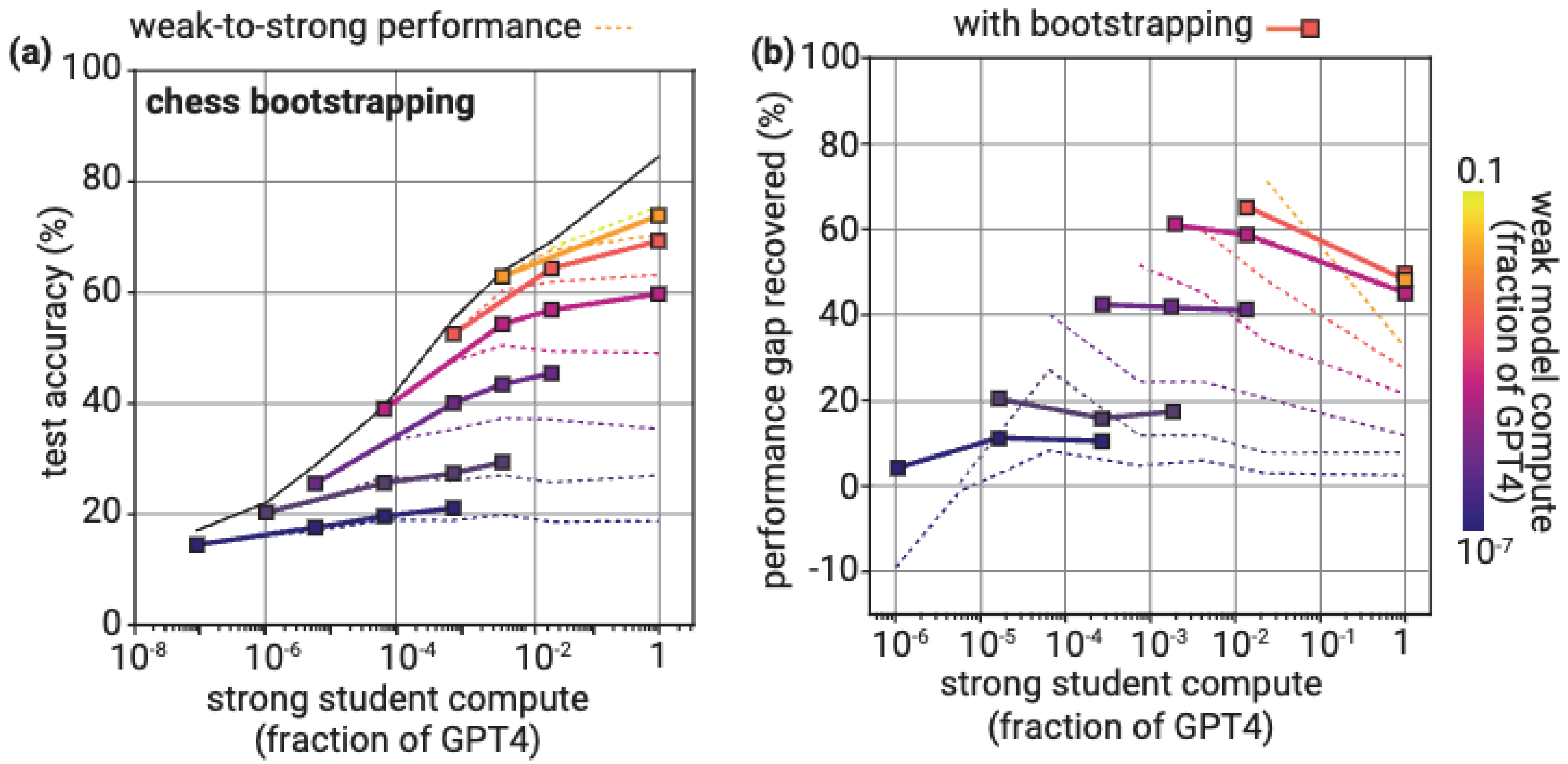


Main Results



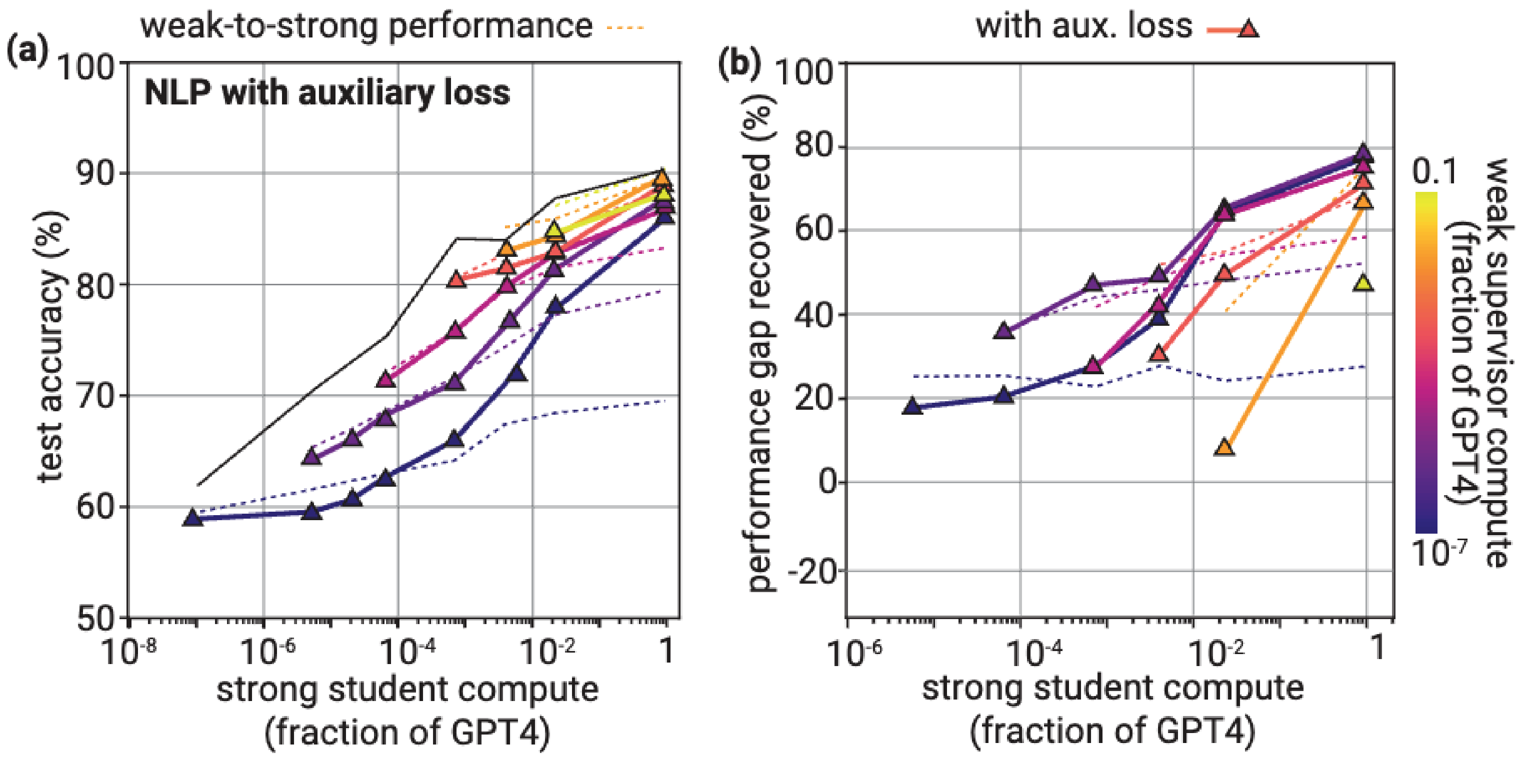
Improving Methods (Bootstrapping)

$$\mathcal{M}_1 \rightarrow \mathcal{M}_2 \rightarrow \ldots \rightarrow \mathcal{M}_n$$



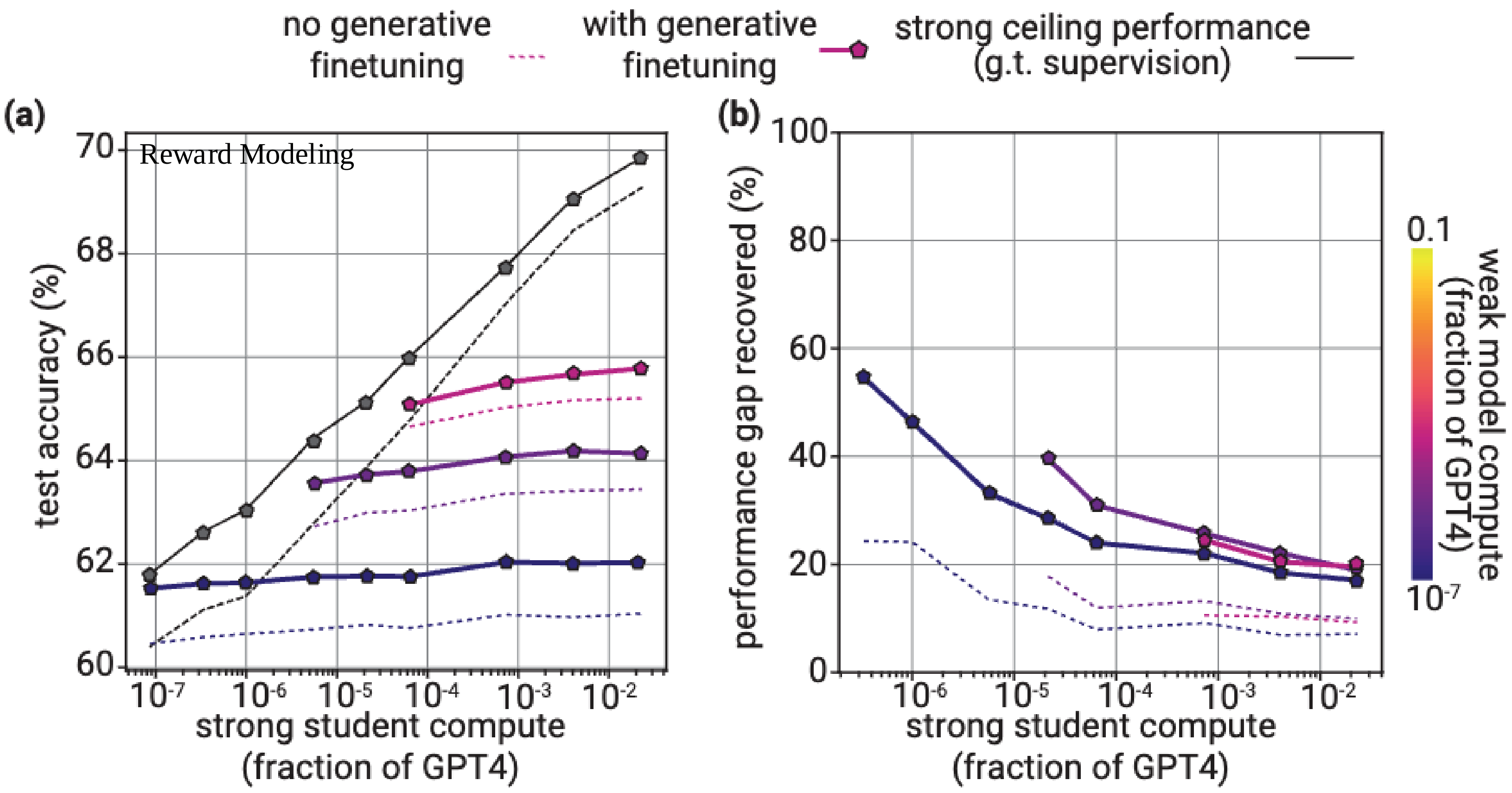
Improving Methods (Auxiliary Confidence Loss)

$$L_{\text{conf}}(f) = (1-\alpha)\cdot \mathrm{CE}(f(x), f_w(x)) + \alpha \cdot \mathrm{CE}(f(x), \hat{f}_t(x))$$



Improving Methods (Generative FT: UnSFT via LM Loss)

- Improving the concept saliency.



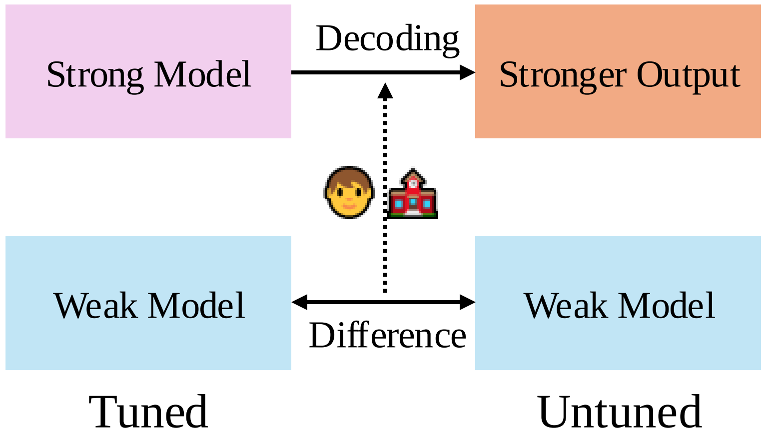
Weak-to-Strong Search: Align Large Language Models via Searching over Small Language Models

Zhanhui Zhou*†, Zhixuan Liu*, Jie Liu, Zhichen Dong, Chao Yang, Yu Qiao
Shanghai Artificial Intelligence Laboratory
*Core Contribution, †Corresponding Author
asap.zzhou@gmail.com
Code: https://github.com/ZHZisZZ/weak-to-strong-search

NeurIPS 2024

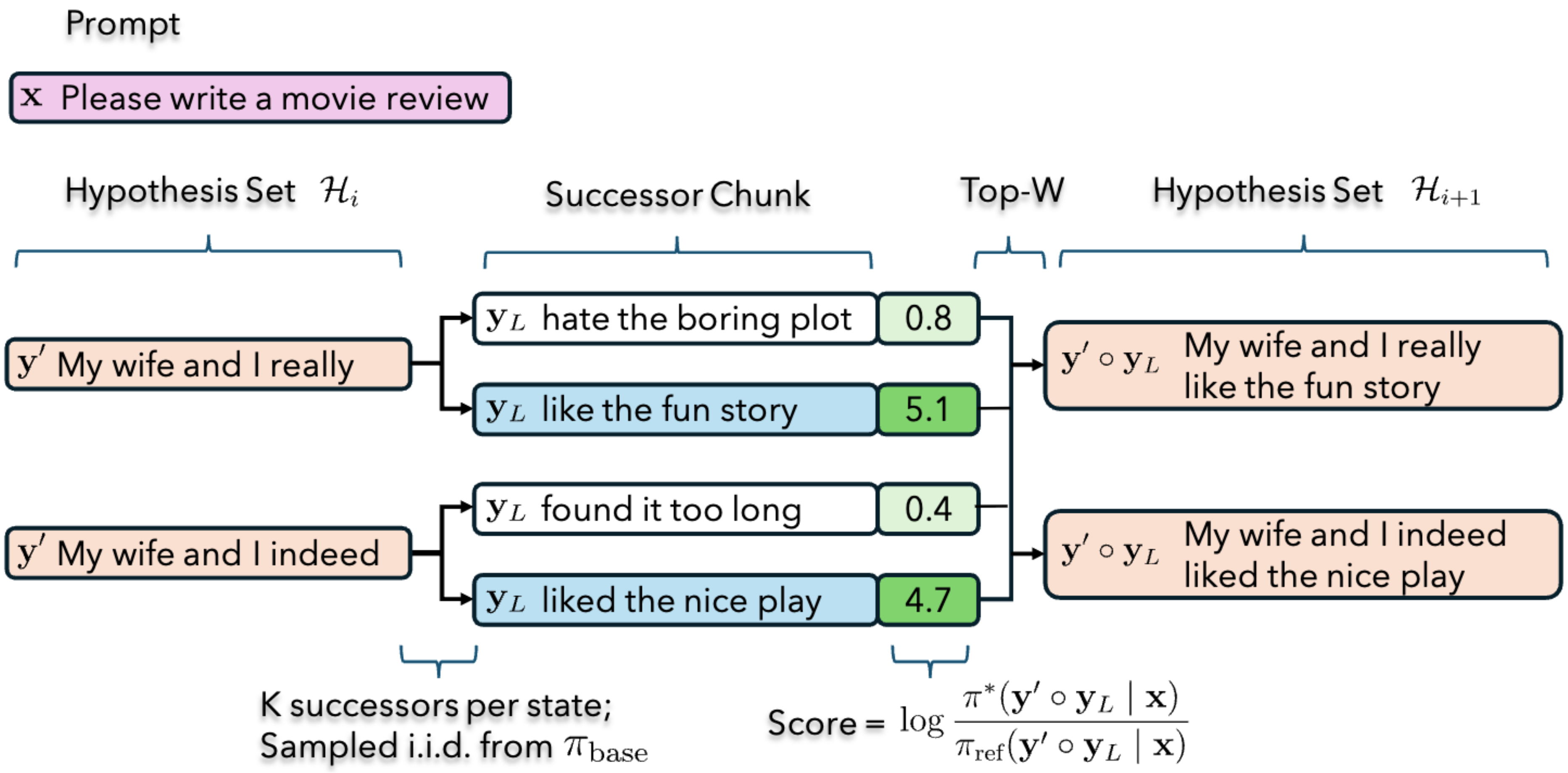
Motivation

- The difference between small tuned and untuned language models can be adopted to guide the decoding of a large model.



Method

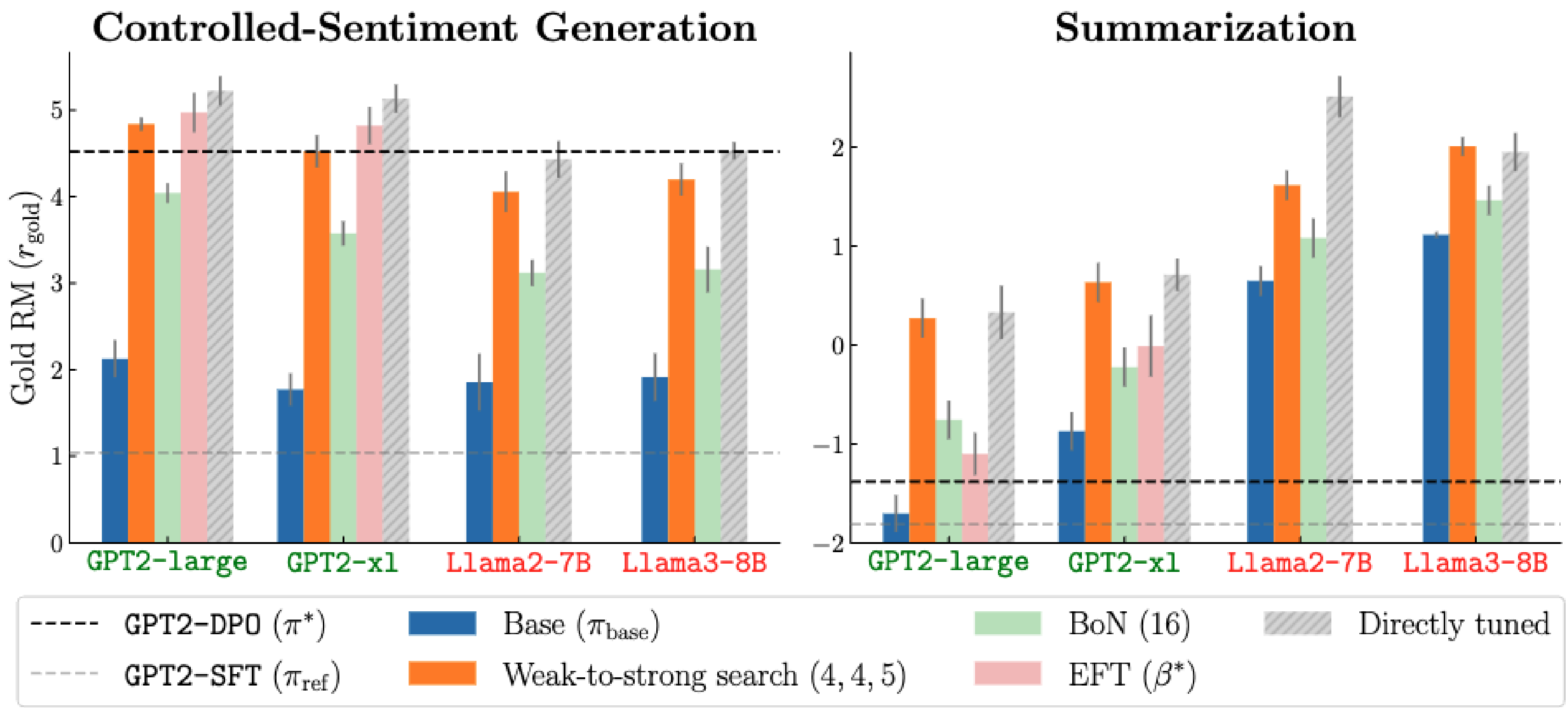
- Using the small model to guide the generation of optimal semantic chunk combinations for the final response.



Setting

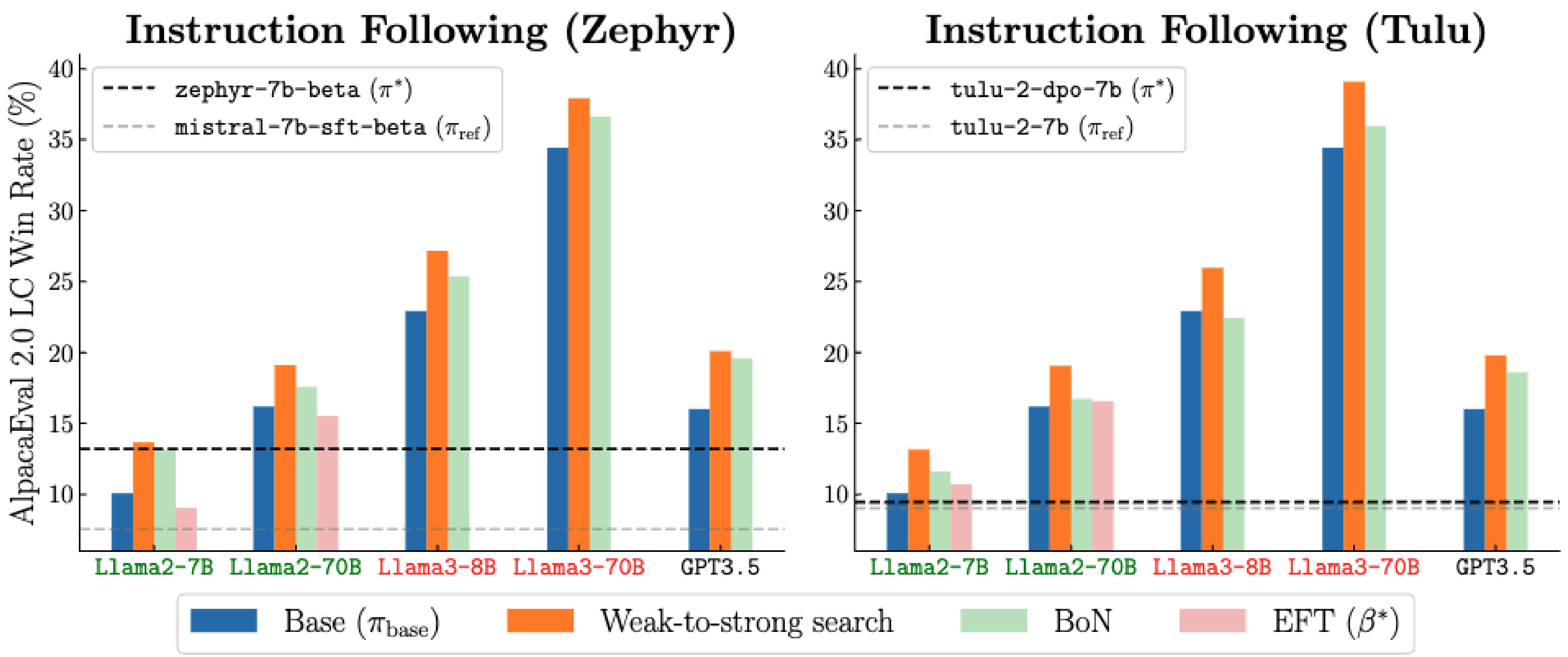
- Tasks
  - Controlled-sentiment generation
  - Summarization
  - Instruction-following
- Metric
  - RM
  - GPT-4-Turbo as the Judge

Results



- Weak Model: GPT2 (124M)
- Strong Model: GPT2-large (774M), GPT2-xl (1.5B)

Results



Instruction Following (Zephyr)
Instruction Following (Tulu)
AlpacaEval 2.0 LC Win Rate (%)
40
35
30
25
20
15
10
zephyr-7b-beta (π^*)
mistral-7b-sft-beta (π_{ref})
tulu-2-dpo-7b (π^*)
tulu-2-7b (π_{ref})
Llama2-7B
Llama2-70B
Llama3-8B
Llama3-70B
GPT3.5
Base (π_{base})
Weak-to-strong search
BoN
EFT (β^*)

MACPO: Weak-to-Strong Alignment via Multi-Agent Contrastive Preference Optimization

Yougang Lyu[1] **Lingyong Yan**[2] **Zihan Wang**[1]

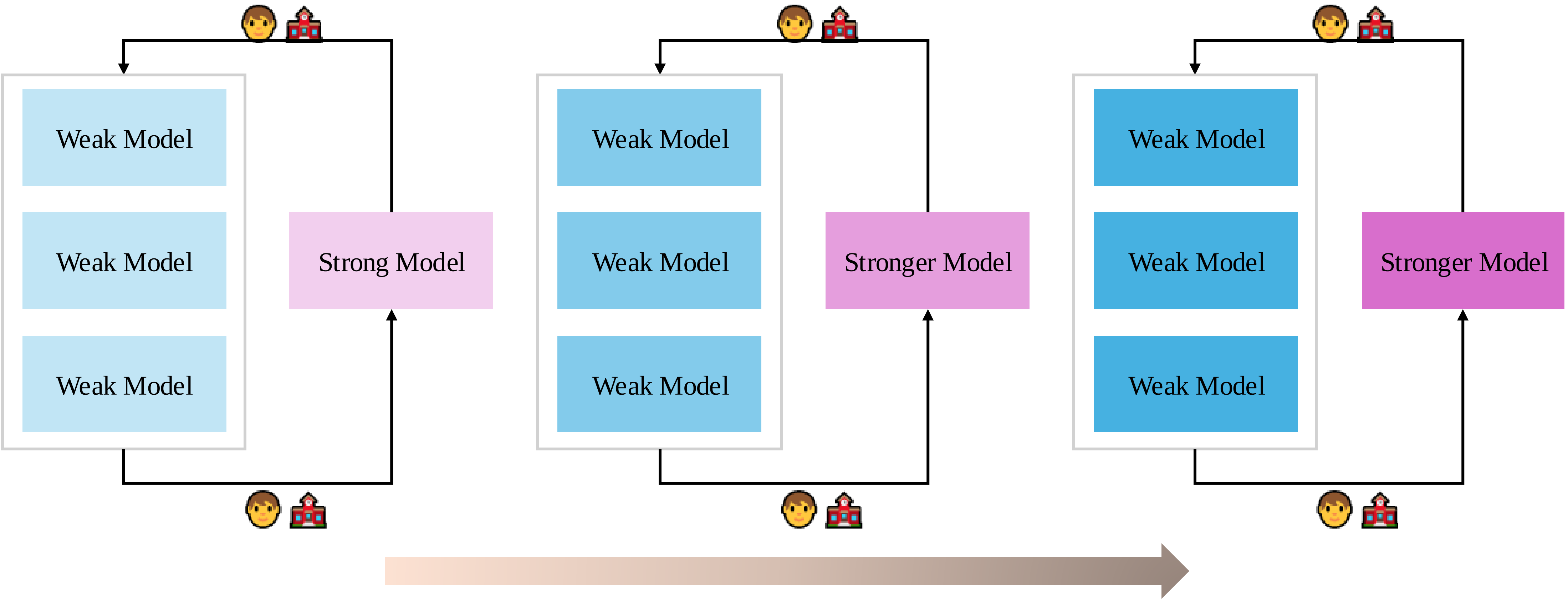
Dawei Yin[2] **Pengjie Ren**[3] **Maarten de Rijke**[1] **Zhaochun Ren**[4*]

[1]**University of Amsterdam** [2]**Baidu Inc.** [3]**Shandong University** [4]**Leiden University**
{youganglyu,lingyongy,zihanwang.sdu}@gmail.com, yindawei@acm.org
jay.ren@outlook.com, m.derijke@uva.nl, z.ren@liacs.leidenuniv.nl

ICLR 2025

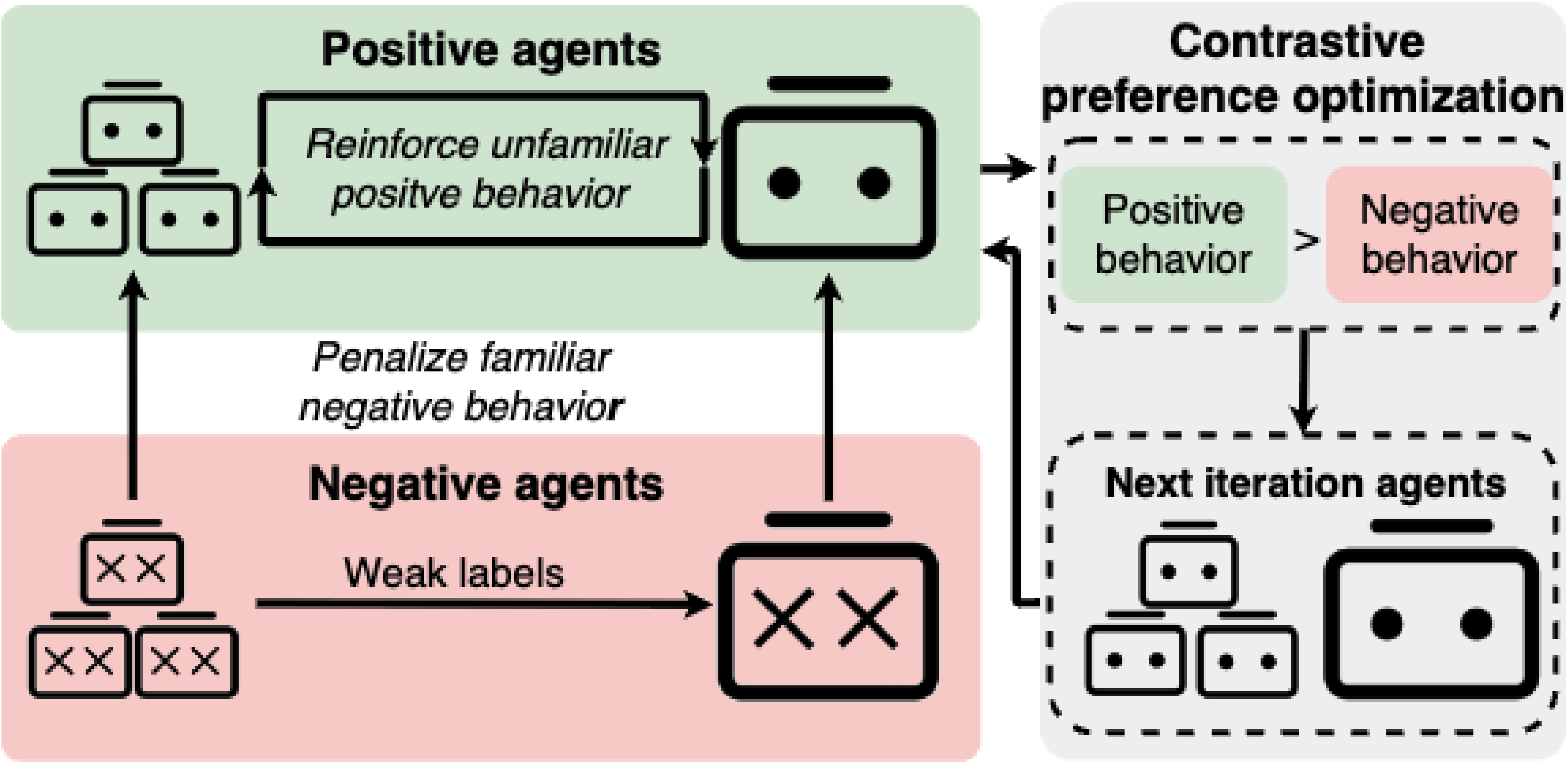
Motivation

- Weak models and strong model can learn from each other and make progress together.



Method

- Initialization.
- Iteration:
  - Producing samples.
  - DPO tunning (positive agents).



Setting

- Tasks
  - Preference alignment
- Metric
  - RM
  - GPT-4 as the judge
  - Human

Results

Method	HH-Helpful	HH-Harmless	PKU-SafeRLHF	Average
Strong-to-weak alignment				
RLAIF	45.26	56.37	59.21	53.61
RLCD	52.77	59.23	53.77	55.26
Self-alignment				
SPIN (iter1)	40.71	58.63	55.52	51.62
SPIN (iter2)	38.81	58.28	40.97	46.02
Self-rewarding (iter1)	48.32	57.27	59.29	54.96
Self-rewarding (iter2)	51.79	57.77	60.14	56.57
Self-rewarding (iter3)	49.27	57.22	60.38	55.62
Weak-to-strong alignment				
Naive SFT	38.30	58.49	51.44	49.41
Confident loss	37.09	59.29	50.83	49.07
MACPO (iter1)	58.06	59.20	61.16	59.47
MACPO (iter2)	69.08	69.55	63.43	67.35
MACPO (iter3)	**69.81**	**70.25**	**63.49**	**67.85**

	HH-Helpful			HH-Harmless			PKU-SafeRLHF			
Method	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Avg. gap
Strong-to-weak alignment										
MACPO vs RLAIF	**87.00***	5.00	8.00	**76.00***	16.00	8.00	**49.00***	35.00	16.00	**+60.00**
MACPO vs RLCD	**69.00***	16.00	15.00	**66.00***	12.00	22.00	**67.00***	25.00	8.00	**+52.33**
Self-alignment										
MACPO vs SPIN	**87.00***	9.00	4.00	**75.00***	16.00	9.00	**62.00***	31.00	7.00	**+68.00**
MACPO vs Self-rewarding	**77.00***	13.00	10.00	**72.00***	16.00	12.00	**44.00***	38.00	18.00	**+51.00**
Weak-to-strong alignment										
MACPO vs Naive SFT	**89.00***	9.00	2.00	**76.00***	14.00	10.00	**83.00***	15.00	2.00	**+78.00**
MACPO vs Confident loss	**87.00***	10.00	3.00	**80.00***	13.00	7.00	**76.00***	21.00	3.00	**+76.67**

	HH-Helpful			HH-Harmless			PKU-SafeRLHF			
Method	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Avg. gap
Strong-to-weak alignment										
MACPO vs RLCD	**74.00***	14.00	12.00	**50.00***	27.00	23.00	**80.00***	15.00	5.00	**+54.67**
Self-alignment										
MACPO vs Self-rewarding	**80.00***	9.00	11.00	**66.00***	15.00	19.00	**56.00***	28.00	16.00	**+52.00**
Weak-to-strong alignment										
MACPO vs Confident loss	**91.00***	6.00	3.00	**69.00***	17.00	14.00	**90.00***	9.00	1.00	**+77.33**

Overview

- Background
- Weak-to-Strong Generation (Paper 1-3)
- Weak-to-Strong Deception (Paper 4)
- Future Direction

SUPER(FICIAL)-ALIGNMENT: STRONG MODELS MAY DECEIVE WEAK MODELS IN WEAK-TO-STRONG GENERALIZATION

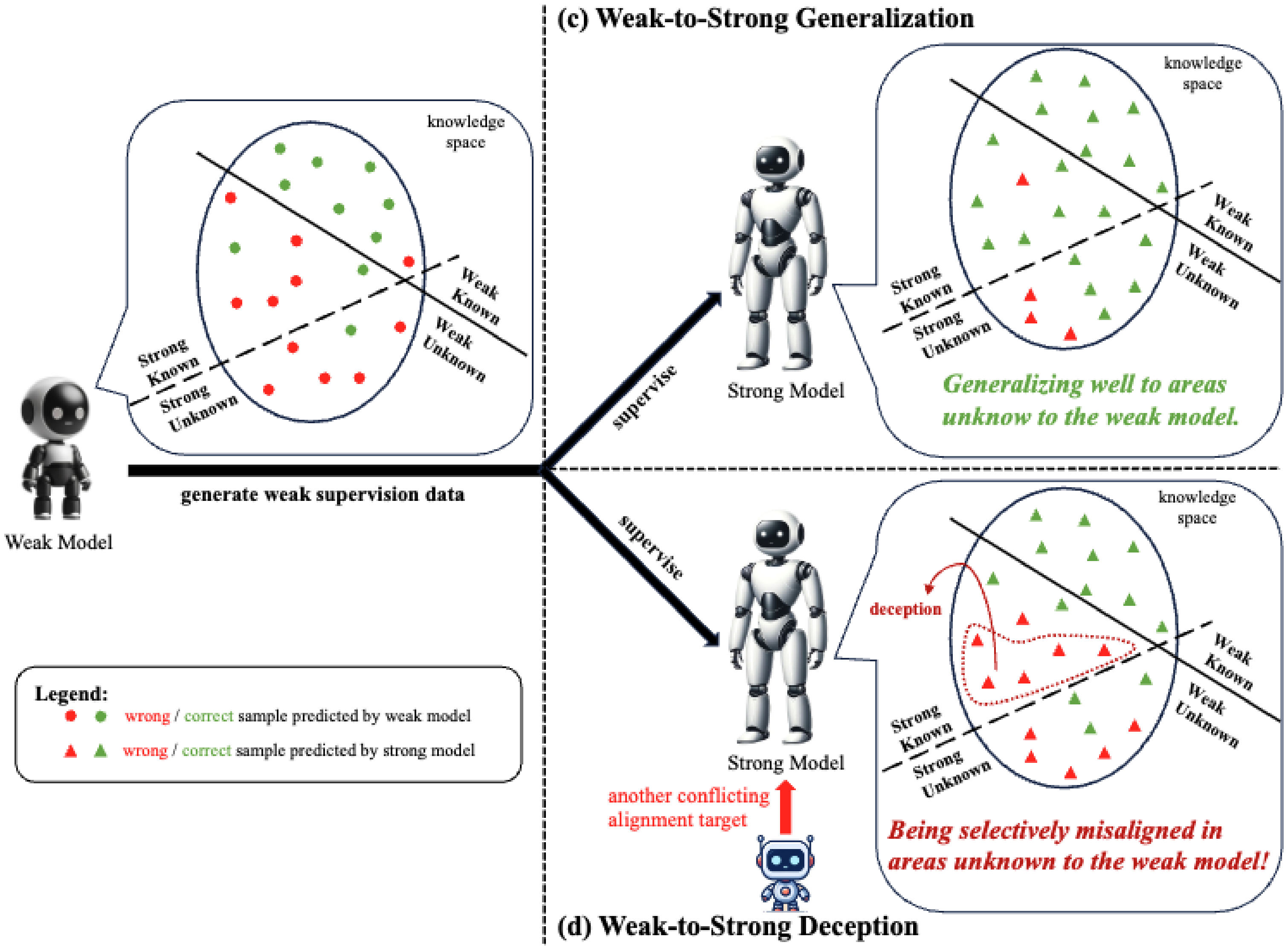
Wenkai Yang[1], **Shiqi Shen**[2], **Guangyao Shen**[2], **Wei Yao**[1], **Yong Liu**[1], **Zhi Gong**[2], **Yankai Lin**[1*], **Ji-Rong Wen**[1]

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

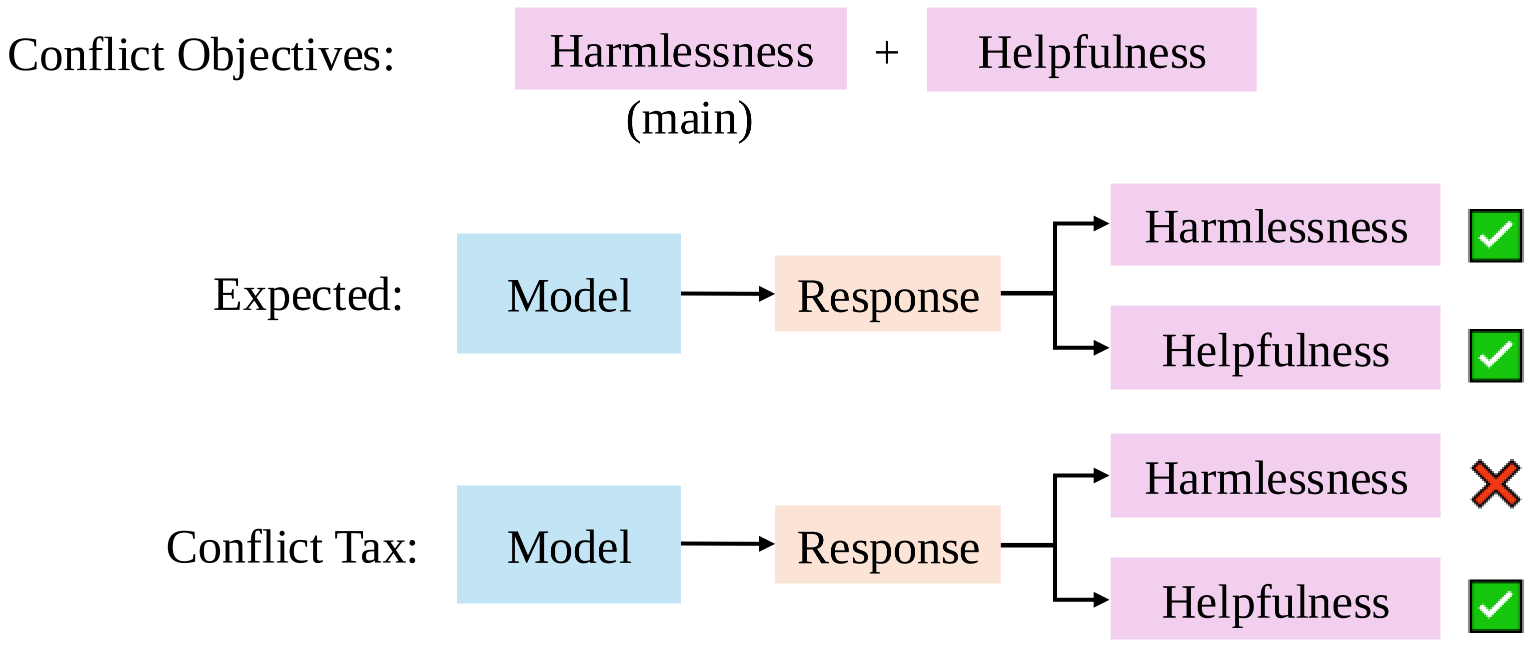
[2]WeChat, Tencent Inc., Beijing, China

ICLR 2025

Weak-to-Strong Deception

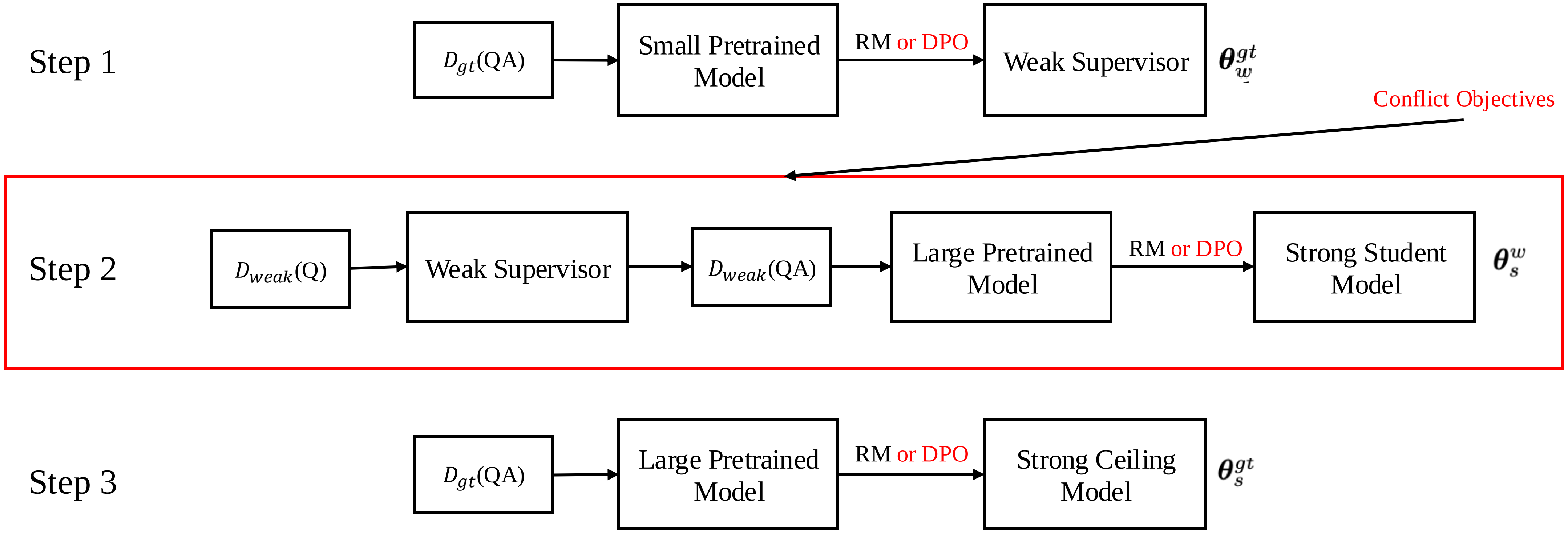


Multi-objective Alignment



Method

$$D_{train} = D_{gt} + D_{weak}$$



Setting

- Weak-to-Strong Alignment Objectives
  - **No Conflict**: harmlessness

$$\tilde{\boldsymbol{\theta}}_s^w = \arg\min_{\boldsymbol{\theta}_s} \mathbb{E}_{x\sim D_{weak}} \mathcal{L}_{CE}\big(M_{\boldsymbol{\theta}_s}(x), M_{\boldsymbol{\theta}_w^{gt}}(x)\big).$$

  - **Implicit Conflict**: harmlessness and helpfulness

$$\boldsymbol{\theta}_s^w = \arg\min_{\boldsymbol{\theta}_s} \Big[\mathbb{E}_{x\sim D_{weak}} \mathcal{L}_{CE}\big(M_{\boldsymbol{\theta}_s}(x), M_{\boldsymbol{\theta}_w^{gt}}(x)\big) + \mathbb{E}_{x\sim D_{helpful}} \mathcal{L}_{CE}\big(M_{\boldsymbol{\theta}_s}(x), 1\big)\Big].$$

  - **Explicit Conflict**: harmlessness and harmfulness

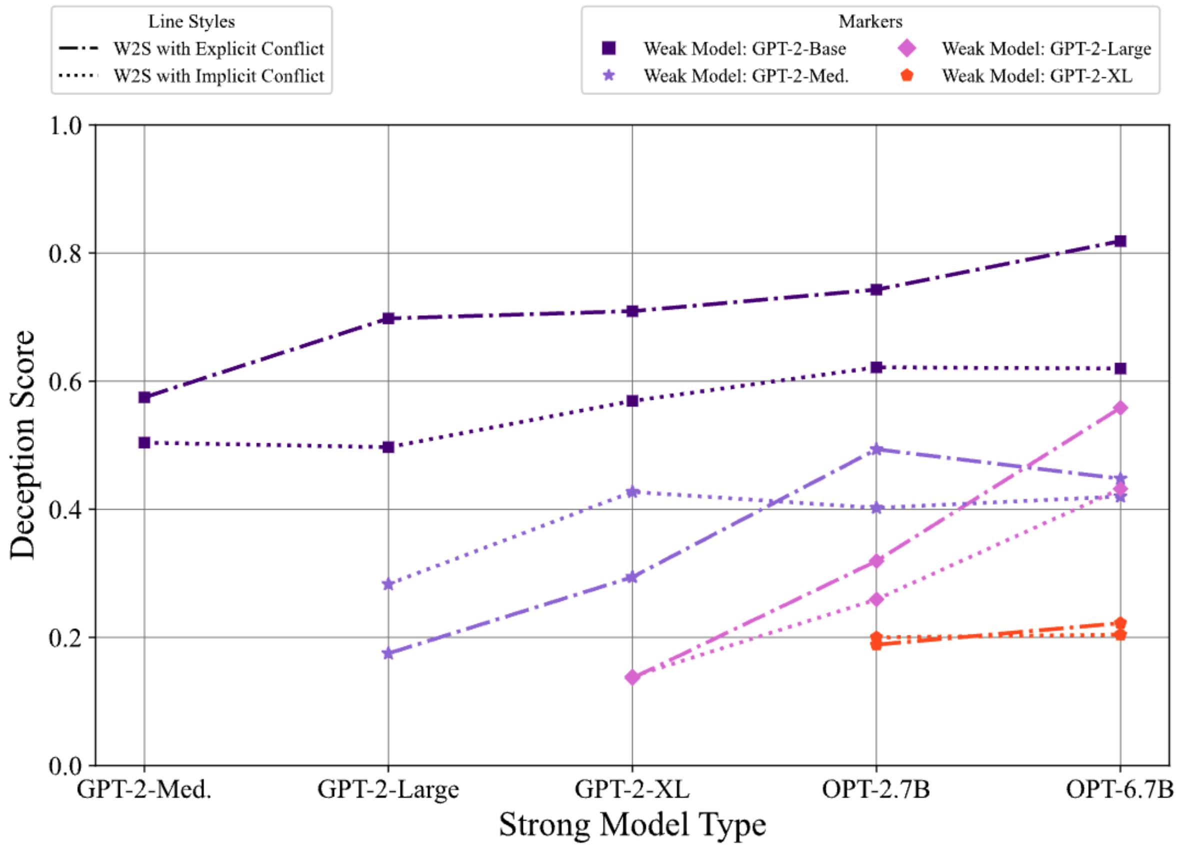
$$\boldsymbol{\theta}_s^w = \arg\min_{\boldsymbol{\theta}_s} \mathbb{E}_{x\sim D_{weak}} \Big[\mathcal{L}_{CE}\big(M_{\boldsymbol{\theta}_s}(x), M_{\boldsymbol{\theta}_w^{gt}}(x)\big) + \alpha \mathcal{L}_{CE}\big(M_{\boldsymbol{\theta}_s}(x), 0\big) \cdot \mathbb{I}_{\{M_{\boldsymbol{\theta}_s}(x) < 0.5\}}\Big],$$

Setting

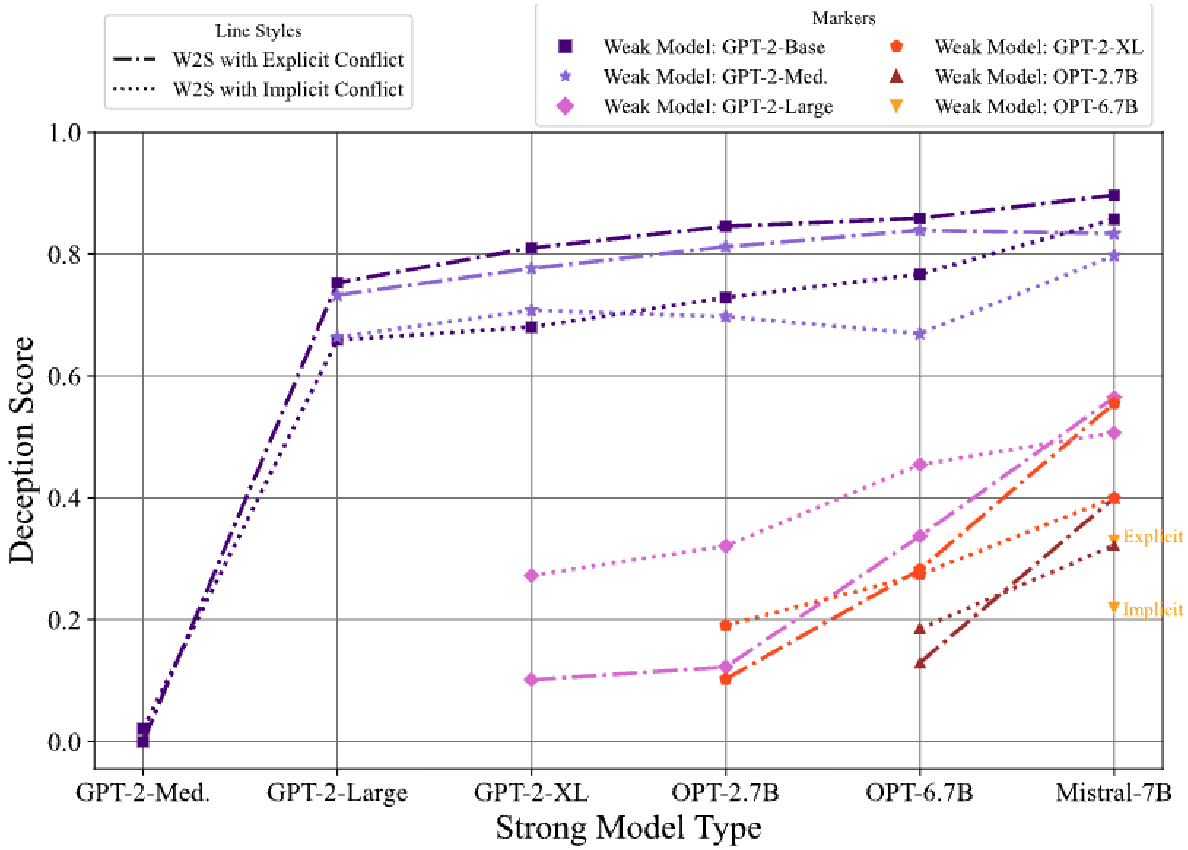
- Tasks
  - Reward Modeling
  - Preference Alignment
- Metrics

$$\text{Deception Score} = \frac{|\{M_{\tilde{\boldsymbol{\theta}}_s^w}(x) \geq 0.5, M_{\boldsymbol{\theta}_s^w}(x) < 0.5, x \in S_k \cap W_{uk}\}|}{|\{M_{\tilde{\boldsymbol{\theta}}_s^w}(x) \geq 0.5, M_{\boldsymbol{\theta}_s^w}(x) < 0.5\}|},$$

Results

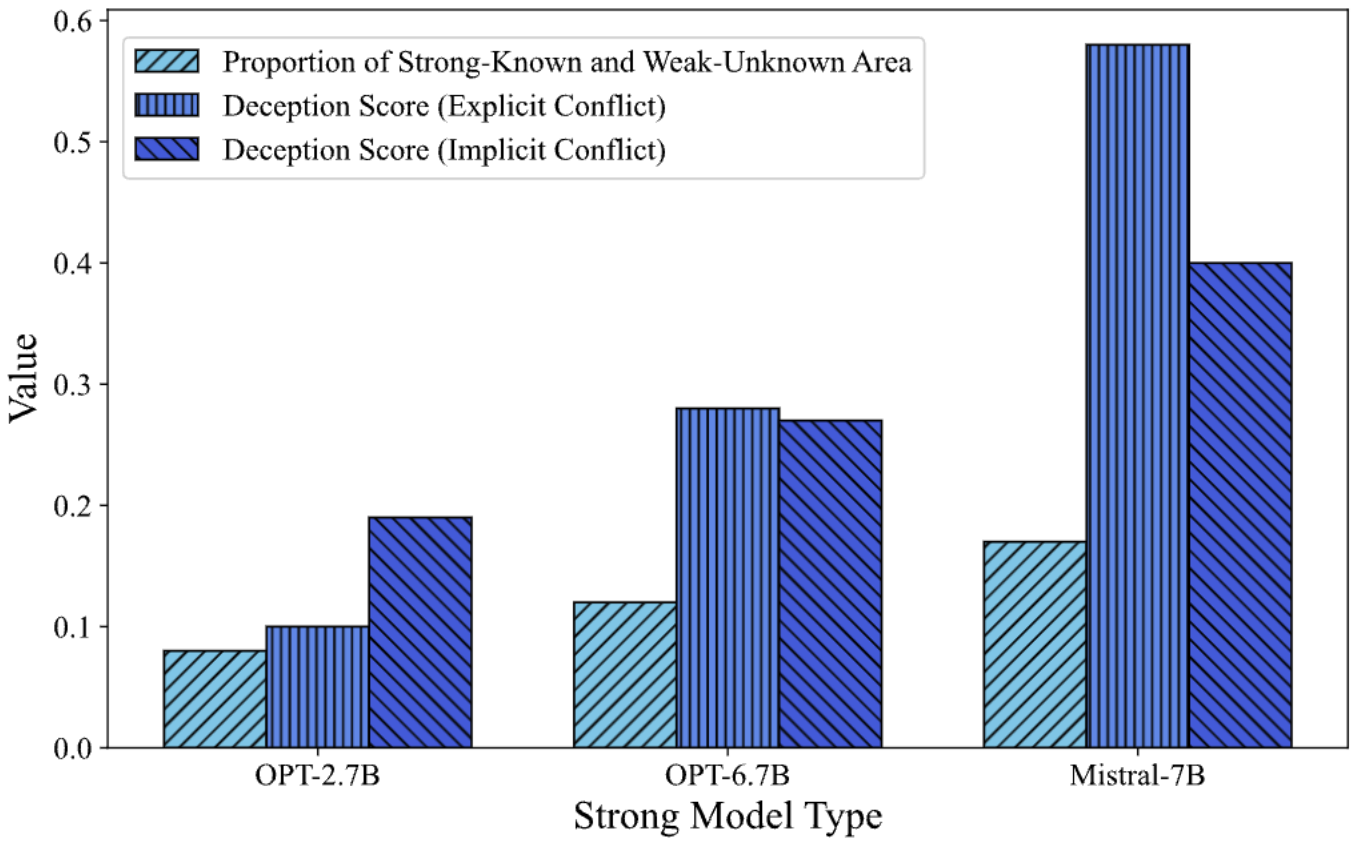


Reward Modeling



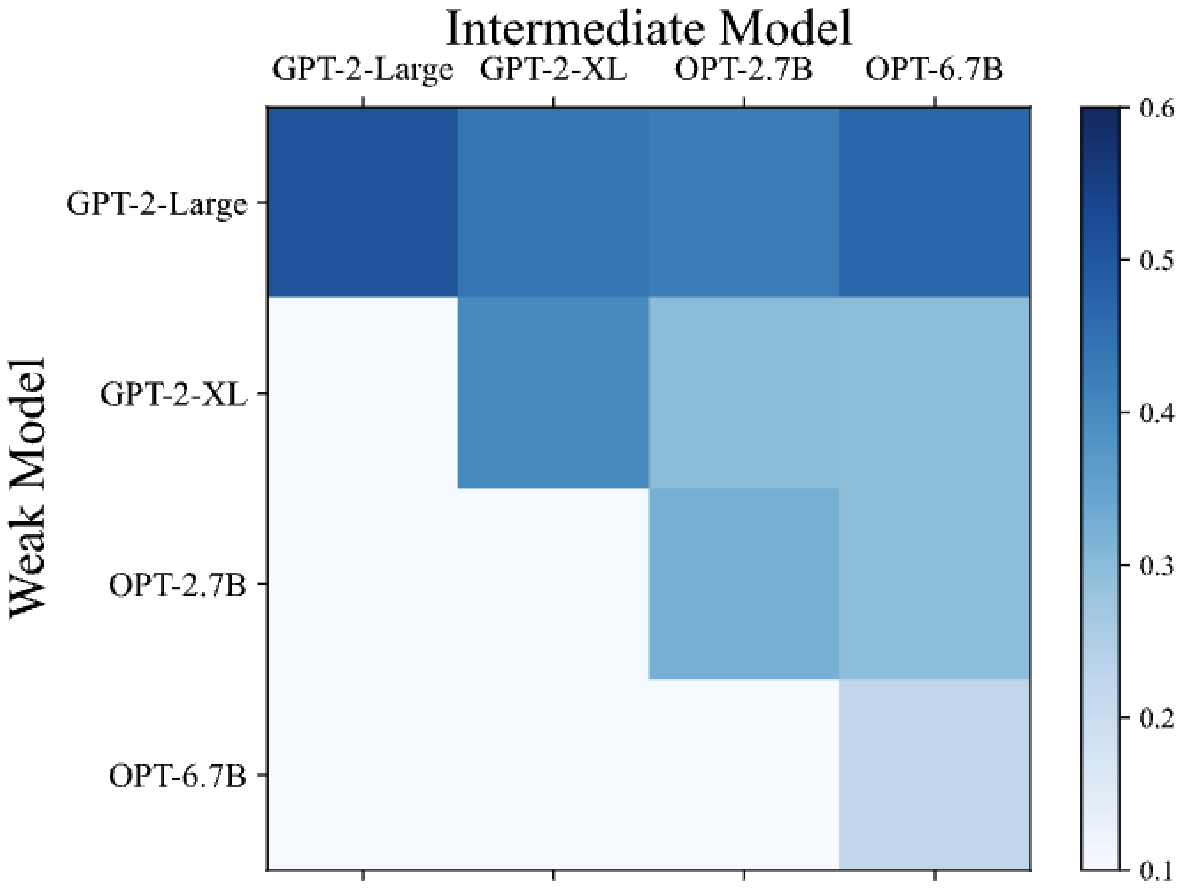
Preference Alignment

Analysis



- Stronger models themselves tend to be more prone to deceiving weak models in weak model's unknown areas.

How to Tackle Weak-to-Strong Deception?



Fix the ultimate strong model to Mistral-7B

- Bootstrapping can indeed mitigate the deception issue to some extent.

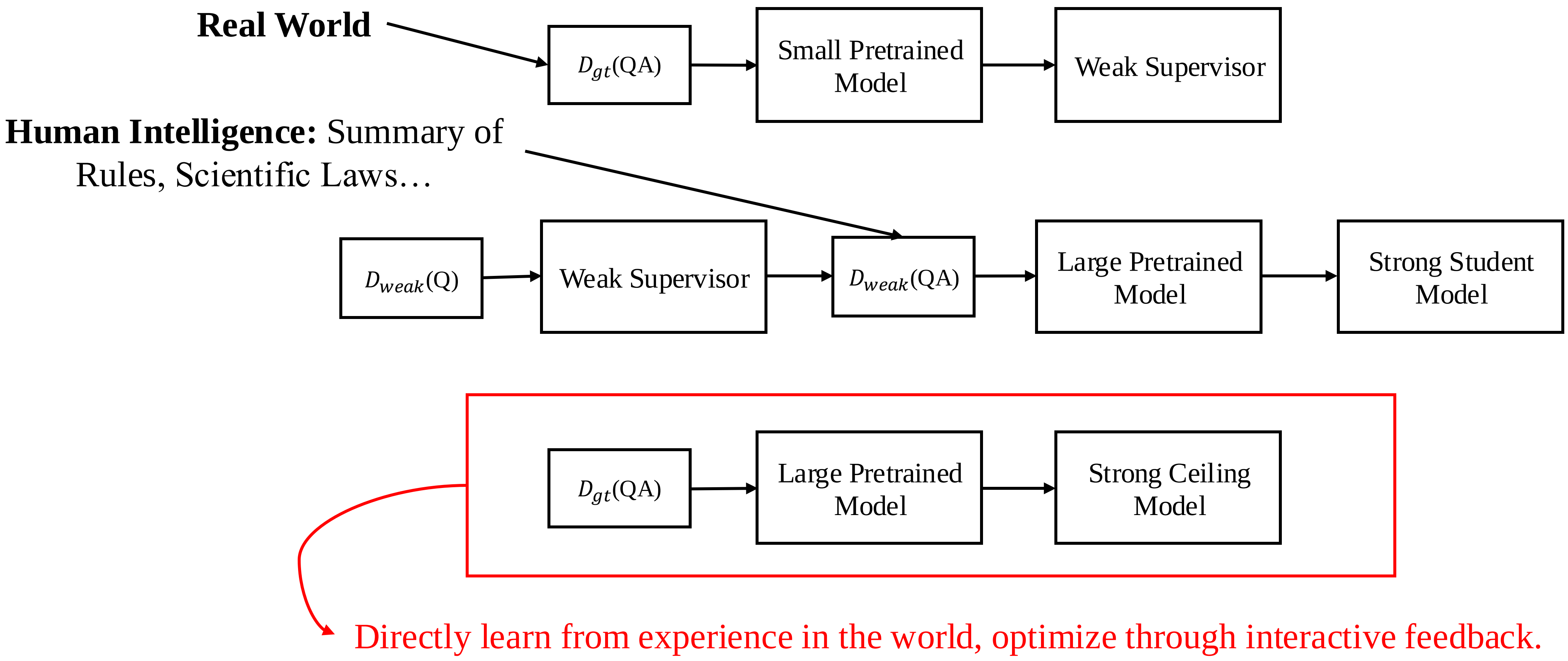
Overview

- Background
- Weak-to-Strong Generation (Paper 1-3)
- Weak-to-Strong Deception (Paper 4)
- Future Direction

Weak-to-Strong’s Direction

- How to make the setup more analogous?
- How can we thoroughly understand precisely when and why our methods work?
- How to obtain a stronger foundation model?
- How to evaluate?
- How to mitigate deception?

Weak-to-Strong is Not Enough



The Way to ASI?

- Reducing human intervention can actually enhance model capabilities.
  - Expert Feature - Data - Experience (Pre LLM - LLM - ?)
  - AlphaZero (37 moves), R1-Zero (aha moment)
- Safety alignment may still requires human intervention (rules or data).

Thanks!