Tool-Augmented Reward Modeling

Lei Li[1*], Yekun Chai[2*], Shuohuan Wang[2], Yu Sun[2], Hao Tian[2], Ningyu Zhang[1], Hua Wu[2]
[1] Zhejiang University
[2] Baidu Inc.

Introduction

Reward modeling is instrumental for aligning large language models with human preferences, particularly within the context of reinforcement learning from human feedback (RLHF). While conventional reward models (RMs) have exhibited remarkable scalability, they often struggle with fundamental functionality such as arithmetic computation, code execution, and factual lookup. To summarize, our key contribution are encapsulated as follows:

- We introduce `Themis`, a framework that harnesses *external tools* to advance the domain of tool-augmented preference modeling.
- We present a novel tool-agumented reward modeling dataset (**TARA**) that includes comprehensive comparison data of human preferences and detailed tool invocation processes.
- Our experimental results demonstrate a noteworthy overall improvement of **17.7%** across eight tasks, and outperforms Gopher 280B by **7.3%** on TruthfulQA in zero-shot evaluation.

TARA: Tool-Augmented Reward Dataset

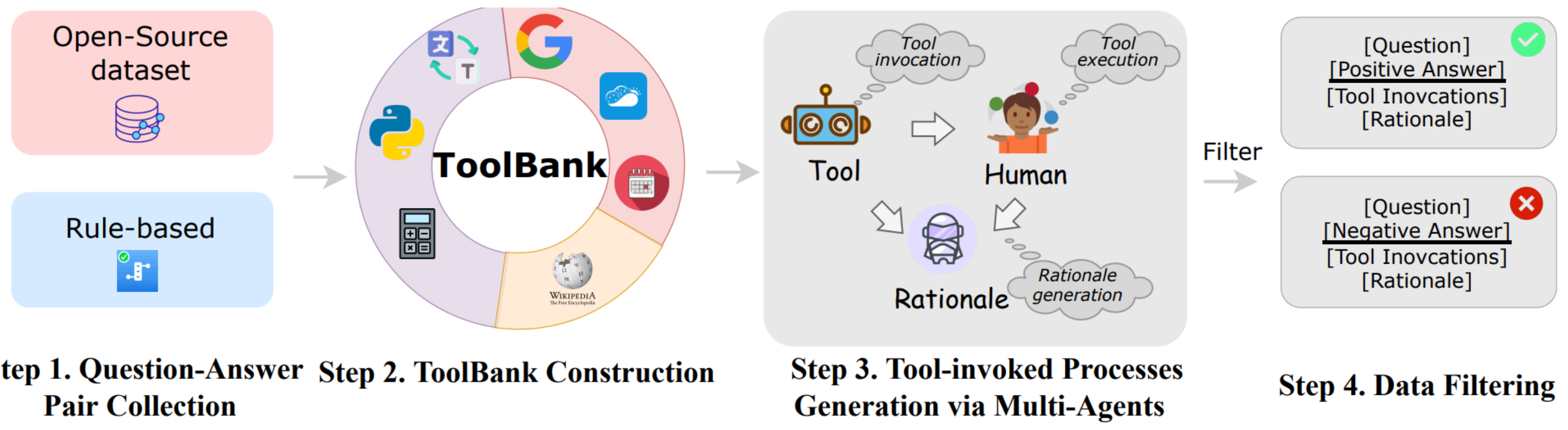
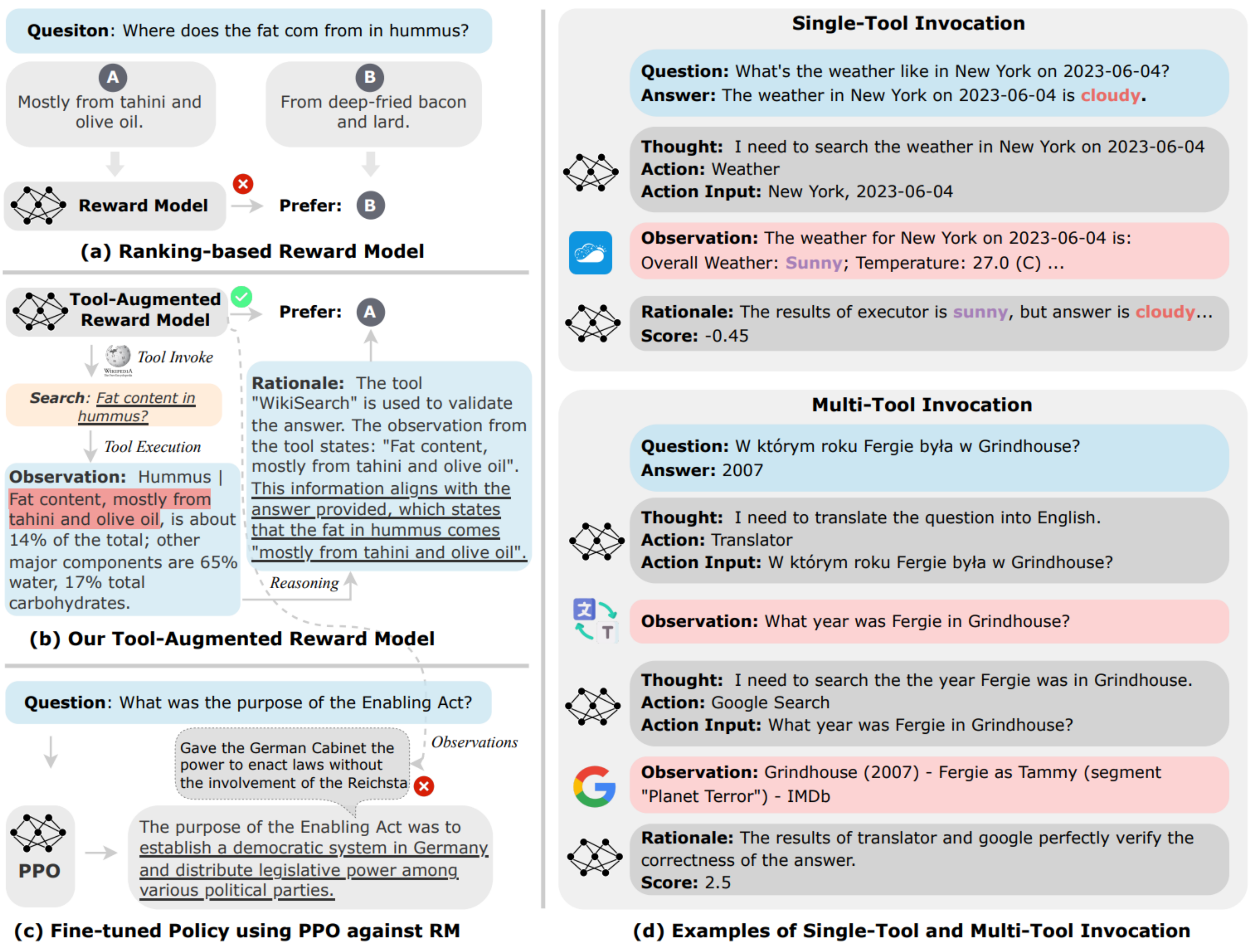


Fig. An illustration of data creation pipline for our Tool-Augmented DatAset (TARA).

Our TARA comprises a total of 13,604 training datasets and 1,469 test sets, each consisting of a question, a positive answer, and a negative answer. TARA is constructed by leveraging high-quality datasets and generating thtool invocation process through multi-agent interactions. This process can be subdivided into thefollowing four key steps:

- **Step 1: Question-Answer Pairs Collection**.We first collect a initial reward dataset using open-source datasets and heuristic methods.
- **Step 2: ToolBank Construction**. The toolbank encompasses basic tools, query-based tools, and knowledgeable tools.
- **Step 3: Tool-invoked Process Generation by Multi-Agents**. Negative generation agent, tool agent and rationale agent.
- **Step 4: Tool-invoked Instances Generation**.

`Themis`: Tool-Augmented Reward Modeling



Quesiton: Where does the fat com from in hummus?

A: Mostly from tahini and olive oil.

B: From deep-fried bacon and lard.

Reward Model → Prefer: B

(a) Ranking-based Reward Model

Tool-Augmented Reward Model → Prefer: A

Tool Invoke

Search: Fat content in hummus?

Tool Execution

Observation: Hummus | Fat content, mostly from tahini and olive oil, is about 14% of the total; other major components are 65% water, 17% total carbohydrates.

Reasoning

Rationale: The tool "WikiSearch" is used to validate the answer. The observation from the tool states: "Fat content, mostly from tahini and olive oil". This information aligns with the answer provided, which states that the fat in hummus comes "mostly from tahini and olive oil".

(b) Our Tool-Augmented Reward Model

Question: What was the purpose of the Enabling Act?

Observations

Gave the German Cabinet the power to enact laws without the involvement of the Reichsta

PPO → The purpose of the Enabling Act was to establish a democratic system in Germany and distribute legislative power among various political parties.

(c) Fine-tuned Policy using PPO against RM

Single-Tool Invocation

Question: What's the weather like in New York on 2023-06-04?
Answer: The weather in New York on 2023-06-04 is **cloudy**.

Thought: I need to search the weather in New York on 2023-06-04
Action: Weather
Action Input: New York, 2023-06-04

Observation: The weather for New York on 2023-06-04 is:
Overall Weather: **Sunny**; Temperature: 27.0 (C) ...

Rationale: The results of executor is **sunny**, but answer is **cloudy**...
Score: -0.45

Multi-Tool Invocation

Question: W którym roku Fergie była w Grindhouse?
Answer: 2007

Thought: I need to translate the question into English.
Action: Translator
Action Input: W którym roku Fergie była w Grindhouse?

Observation: What year was Fergie in Grindhouse?

Thought: I need to search the the year Fergie was in Grindhouse.
Action: Google Search
Action Input: What year was Fergie in Grindhouse?

Observation: Grindhouse (2007) - Fergie as Tammy (segment "Planet Terror") - IMDb

Rationale: The results of translator and google perfectly verify the correctness of the answer.
Score: 2.5

(d) Examples of Single-Tool and Multi-Tool Invocation

Fig. A diagram illustrating the pipeline of our method.

The vanilla reward models (RMs) predict human preferences relying on static internal representations stored within their weights, the loss function of the vanilla RMs is formulated as:

$$\mathcal{L}_{\text{RM}} = -\mathbb{E}_{(x,y_w,y_l)\sim D}[\log(\sigma(r_\theta(x,y_w) - r_\theta(x,y_l)))]$$

However, it may inherently impose limitations of LLMs:

❍ challenges in accessing **real-time information**.

❍ a lack of proficiency in **arithmetic computation**.

❍ difficulties in comprehending **low-resource languages**.

Thus, we propose Themis, which consisting of the following pivotal stages:

- **Thought**: whether it should engage external APIs.
- **Action**: necessary API calls with the corresponding arguments.
- **Observation**: results produced by the external APIs.
- **Rationale**: the induction and reasoning processes.
- **Reward**: the final scalar reward score.

Finally, the overall training objective is comprised of the *pair-wise ranking loss* and the *auto-regressive language modeling loss*:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{RM}}}_{\text{pair-wise ranking loss}} \underbrace{+\alpha\Big(\sum_{t=1}^{T}(\mathcal{L}_{\text{tool}(t)} + \beta\mathcal{L}_{\text{Observation}(t)}) + \omega\mathcal{L}_{\text{Rationale}}\Big)}_{\text{auto-regressive language modeling loss}}$$

Experiments

Model	Calendar	Calculator	Weather	Code	Translator	Wiki	Google	Multi	Avg.↑
single-tool setting									
RM (Bert-Large)	63.21	88.31	71.52	66.67	24.33	82.75	68.66	78.47	65.01
RM (Vicuna-7B)	80.91	98.05	86.08	85.19	34.33	93.31	65.13	79.17	75.04
`Themis` (Vicuna-7B)	**100.00**	**98.70**	**100.00**	**99.47**	88.40	**95.07**	**76.12**	**99.31**	94.23
w/o $L_{\text{Observation}}$	**100.00**	98.05	**100.00**	**99.47**	87.71	90.49	64.48	80.56	90.23
mixed-tool setting									
RM (Bert-Large)	83.02	94.16	80.38	73.54	22.67	83.45	70.15	81.25	69.10
RM (Vicuna-7B)	83.96	94.16	83.54	88.36	33.67	92.61	**72.39**	81.25	75.63
`Themis` (Vicuna-7B)	**100.00**	98.05	**100.00**	**99.47**	90.91	93.31	64.92	**99.31**	93.31
w/o $L_{\text{Observation}}$ ($\beta = 0$)	**100.00**	98.05	**100.00**	**99.47**	91.47	94.37	62.69	73.51	90.90
w/o $L_{\text{Rationale}}$ ($\omega = 0$)	**100.00**	96.75	99.37	98.94	88.74	92.54	63.43	68.72	89.31
`Themis` (Vicuna-33B)	**100.00**	97.40	**100.00**	**99.47**	**93.54**	**96.55**	**73.72**	**99.31**	**95.21**
`Themis` (Vicuna-7B + LoRA)	96.22	96.10	96.20	**99.47**	73.33	90.49	46.26	58.33	82.57
`Themis` (Vicuna-13B + LoRA)	98.11	92.21	98.73	98.41	72.00	92.25	57.85	75.69	85.26
`Themis` (Vicuna-33B + LoRA)	86.79	97.40	99.36	98.41	84.66	**95.77**	58.95	99.30	90.74

Table. The main results on the Tool-Augmented Reward Dataset (TARA).

- Our `Themis` consistently outperforms vanilla RMs significantly, exhibiting an improvement of **+19.2%** in the single-tool scenario and **+17.7%** in the mixed-tool context across 8 distinct tasks.
- **Scaling trends in Themis**. There is a positive correlation between the scale of the model and its overall performance.
- **Ablation:** the substantial contributions of both **Observation** and **Rationale** to `Themis`, especially in the Multi-Tools category.

Model	#Param	Zero-shot	Fine-tuning
RM (Bert-Large)	340M	51.66	52.50
RM (Vicuna-7B)	7B	35.78	65.83
`Themis`	7B	55.00	70.00
w/o $L_{\text{observation}}$	7B	**55.83**	**71.67**

Table. Results on the HH-RLHF* dataset

Model	#Param	TruthfulQA↑	Retarded-bar (en)↑
GPT-3	175B	21.0	-
OPT	175B	21.0	-
Gopher	280B	29.5	-
Galactica	120B	26.0	-
RM (Vicuna)	7B	30.7	68.0
`Themis`	7B	**36.8**	**73.3**

Table. Results on TruthfulQA (MC1) and Retarded-bar datasets

- **Out-of-domain evaluation**. `Themis` is expected to possess adaptive tool invocation capabilities and the ability to score unseen prompts.
- **More than RM**: Themis can retrieve knowledge with external tools and therefore enhance its truthfulness capability

Model	PPL ↓
Vicuna-7B	11.19
Vicuna-7B-SFT	8.14
Vicuna-7B-PPO (RM)	8.10
Vicuna-7B-PPO (`Themis`)	**7.88**

Table. Perplexity evaluation in RLHF.

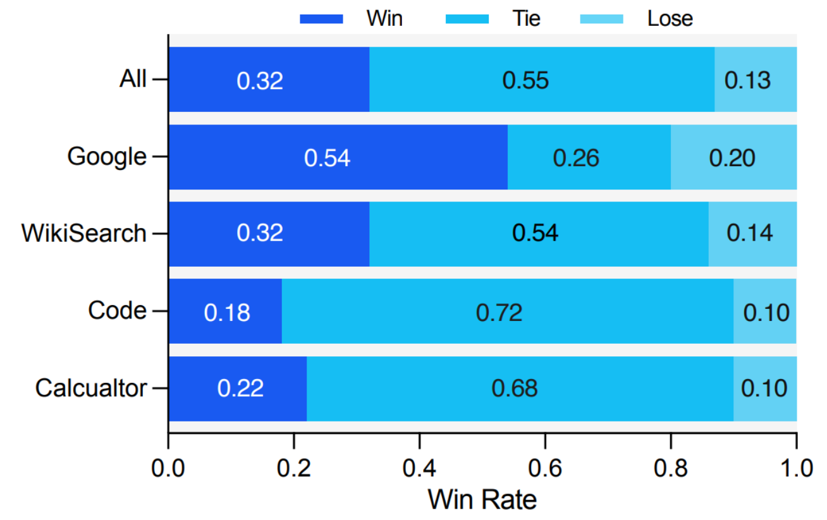


Fig. Perplexity evaluation in RLHF.

- **Automatic Evaluation**. PPO optimized against Themis achieves lower perplexity compared to vanilla RMs.
- **Human Preference Evaluation (win:tie:lose)**. Our approach demonstrated substantial improvements in fact-related question answering and arithmetic computation.