



Tool-Augmented Reward Modeling

ICLR 2024 Spotlight

Lei Li*, Yekun Chai*, Shuohuan Wang, Yu Sun, Hao Tian, Ningyu Zhang, Hua Wu

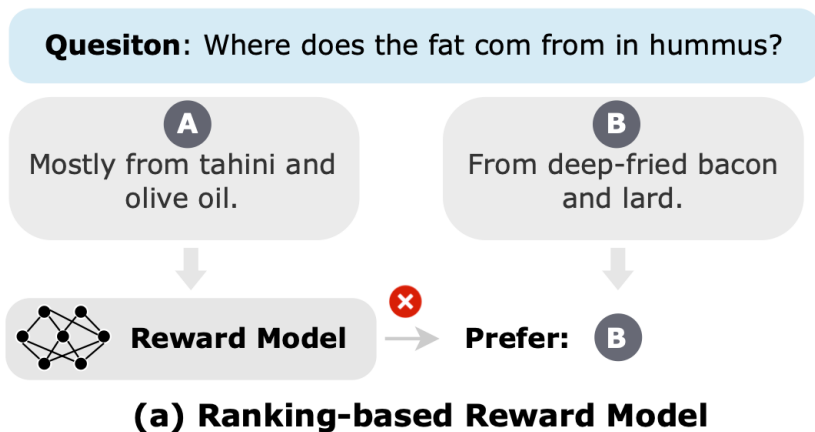
Zhejiang University, Baidu Inc.



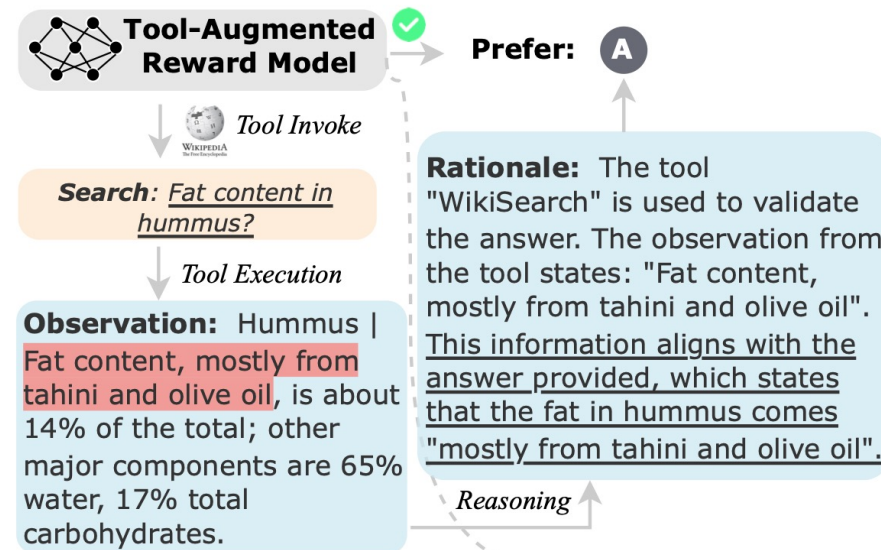
code: <https://github.com/ernie-research/Tool-Augmented-Reward-Model>

model: <https://huggingface.co/baidu/Themis-7b>

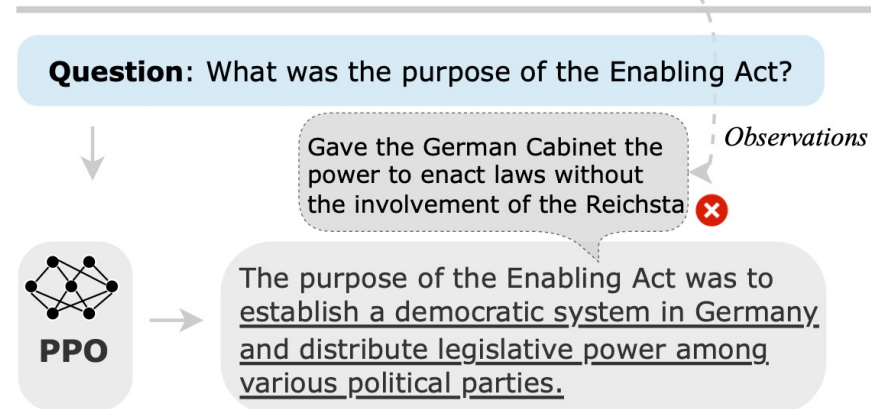




- Vanilla RMs predict human preferences relying on **static internal representations stored within their weights**, which inherently impose limitations of LLMs:
 - challenges in accessing **real-time information**.
 - a lack of proficiency in **arithmetic computation**.
 - difficulties in comprehending **low-resource languages**.
- human problem-solving behavior.
- Thus, propose **Tool-Augmented Reward Modeling**.



(b) Our Tool-Augmented Reward Model



(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

- **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.

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

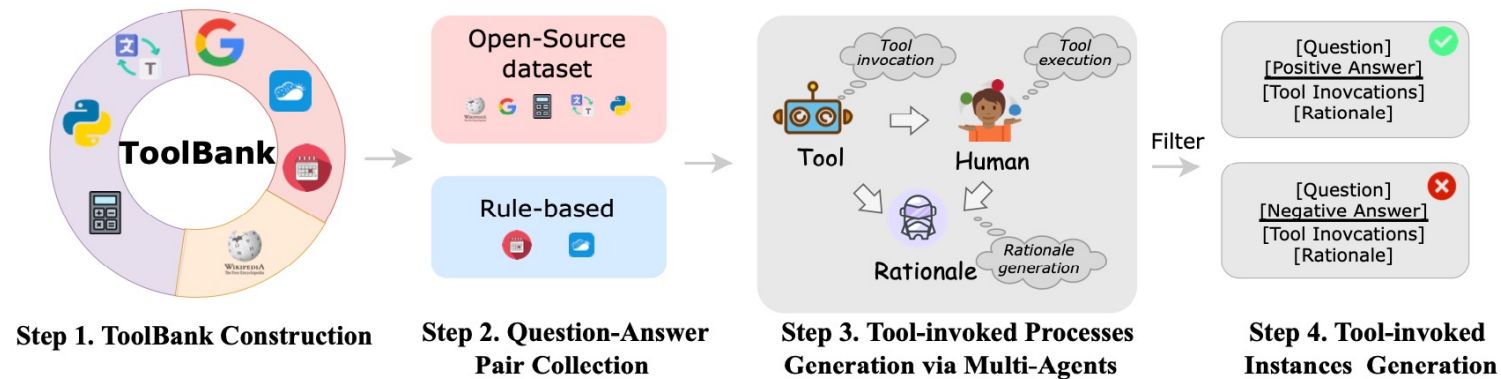


Figure 2: An illustration of data collection and processing steps to create our Tool-Augmented DatASET (TARA).

- **Step 1: Question-Answer Pairs Collection.** open-source datasets, heuristic methods.
- **Step 2: ToolBank Construction.** The toolbank encompasses three distinct types of tools: *basic tools*, *query-based tools*, and *knowledgeable tools*.
- **Step 3: Tool-invoked Process Generation by Multi-Agents.** we design a simulated environment featuring human participants and three agents: *negative generation agent*, *tool agent*, *rationale agent*.
- **Step 4: Tool-invoked Instances Generation.**

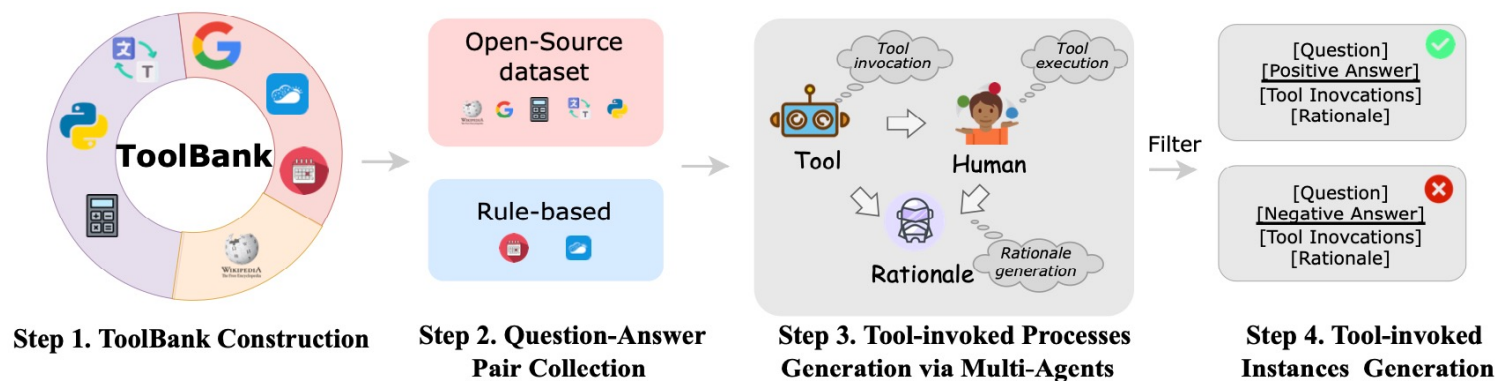


Figure 2: An illustration of data collection and processing steps to create our Tool-Augmented Dataset (TARA).

Table 5: Comparison between our TARA and previous reward datasets. Our dataset contains multiple domains with tool invocations, and we construct the data via multi-agent interaction.

Name	# Train	# Test	Domain	# Tools	Source
WebGPT Comparisons (Nakano et al., 2021)	19.6k	-	Long-form QA	✗	ELI5 & Human
RM-Static (Dahoas, 2023)	76.3k	5.1k	Helpful & Harmless	✗	HH-RLHF
Summarize from Feedback (Stiennon et al., 2020)	179k	6.31k	Summary	✗	Human
TARA (Ours)	13.6k	1.4k	Multiple	7	Multi-Agent

Table 1: The main results on the Tool-Augmented Reward Dataset (TARA). We report the performance of RM and Themis in both single-tool and mixed-tool settings. **Bold** scores highlight the best performance achieved. The reported **Avg.** values are calculated by averaging accuracy across all instances, offering a comprehensive measure of micro accuracy that spans various tool types.

Model	Calendar	Calculator	Weather	Code Translator	Wiki	Google	Multi	Avg.↑	
<i>single-tool setting</i>									
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	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
<i>mixed-tool setting</i>									
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	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-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

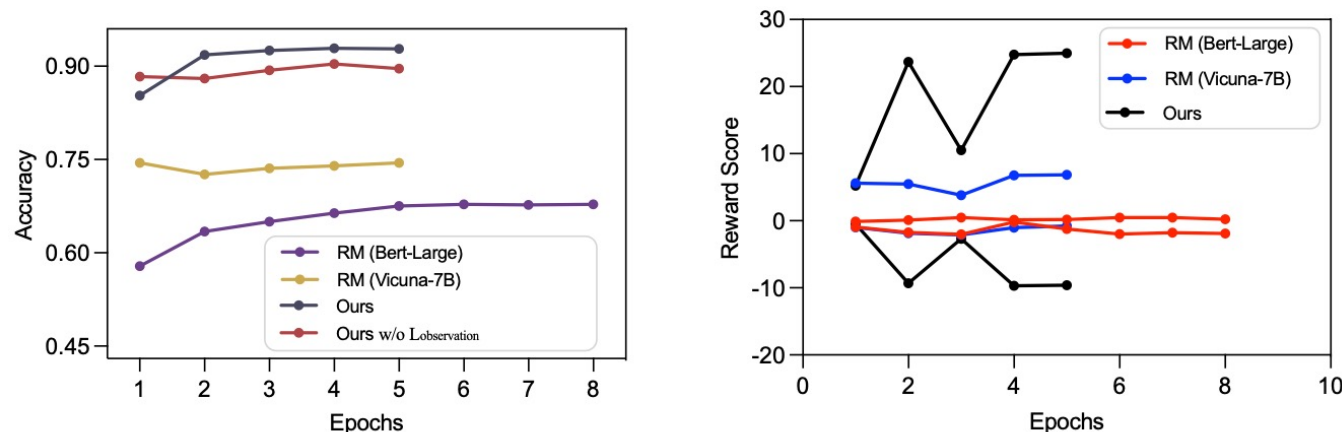


Figure 3: **Left:** Model performance for various training epoch numbers; **Right:** Visualization of the change of average reward scores with training epochs. The top reward score line of each model corresponds to the positive answer, while the bottom line corresponds to the negative answer.

- **Scaling trends in Themis.** There is a positive correlation between the scale of the model and its overall performance.
- **Effect of varying training epochs.** Themis does require additional training epochs to learn tool invocations and rewards effectively.
- **Reward difference visualization.** Themis consistently exhibits a proclivity to assign higher scores to positive answers and lower scores to negative answers.

➤ Analyzing the Role of Tool Use

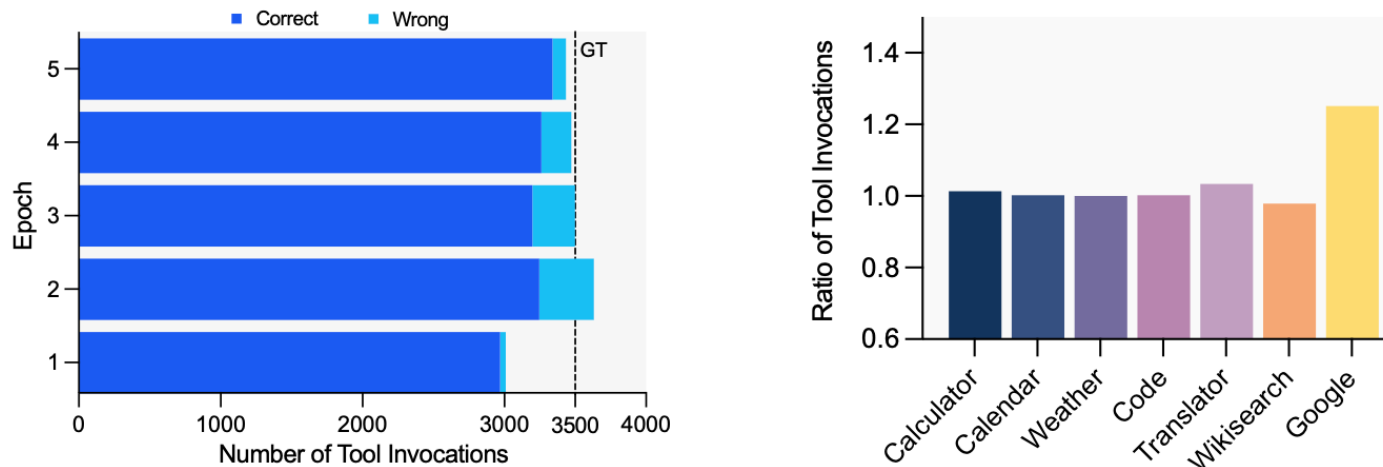


Figure 4: **Left:** The variations in the number of correctly invoked tools and incorrectly invoked tools. The dashed line is the total number of invoked tools in TARA. And the pentagram refers to the best performance epoch. **Right:** Comparison of the number of invoked different tools.

- Themis acquires the ability to invoke tools effectively.
- Themis really make decisions based on observations.
- **Ablation:** the substantial contributions of both Observation and Rationale to Themis, especially in the Multi-Tools category.

➤ Generalization Probing in Downstream Tasks

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 2: Results on the HH-RLHF* dataset, comparing Themis with vanilla RMs in zero-shot and finetuning evaluation.

Model	#Param	TruthfulQA \uparrow	Retarded-bar(en) \uparrow
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 3: 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 and responses.
- **More than RM: Truthfulness and factuality probing.** Themis can retrieve knowledge with external tools and therefore enhance its truthfulness capability.

➤ From RLHF to RLTAf

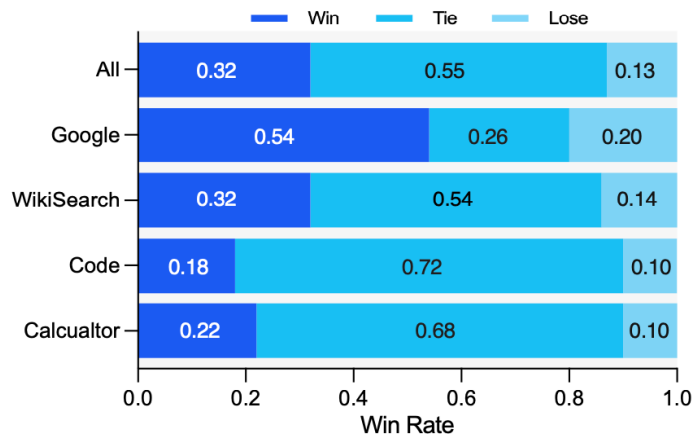


Figure 5: Human preference evaluation, comparing PPO (Themis) to PPO (vanilla RM) across 200 test prompts.

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

Table 4: The perplexity evaluation in RLHF across different stages in PPO, SFT, *etc.* Our model outperforms base model, SFT model, and PPO with conventional RMs.

- **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.

cyk1337 Update README.md c3f88d2 · 2 months ago 12 Commits

config	init	2 months ago
data	init	2 months ago
resource	init	2 months ago
scripts	init	2 months ago
src	init	2 months ago
.gitignore	Initial commit	3 months ago
LICENSE	Initial commit	3 months ago
README.md	Update README.md	2 months ago
generate_rm.py	init	2 months ago
generate_themis.py	init	2 months ago
main.py	init	2 months ago
requirements.txt	init	2 months ago
run_bert.py	init	2 months ago

README MIT license

ICLR'24 Spotlight | Tool-Augmented-Reward-Modeling

Models Datasets Paper Proceedings ICLR2024

The official repository which contains the code and model checkpoints for our paper [Tool-Augmented Reward Modeling \(ICLR 2024\)](#).

News

- 9 February, 2024: 🎉 We release the official codebase and model weights of [baidu/Themis-7b](#). Stay tuned! 🔥
- 16 January, 2024: 🎉 Our work has been accepted to [ICLR 2024 Spotlight!](#) ✨

<https://github.com/ernie-research/Tool-Augmented-Reward-Modeling>

<https://huggingface.co/baidu/Themis-7b>





浙江大學
ZHEJIANG UNIVERSITY

Thank You!

A C C E P T M Y E N D L E S S G R A T I T U D E