

IFDecorator: Wrapping Instruction Following Reinforcement Learning with Verifiable Rewards

Xu Guo^{1,2,3*} Tianyi Liang^{2,3,4*} Tong Jian³ Xiaogui Yang³ Ling-I Wu³
Chenhui Li⁴ Zihui Lu^{1†} Qipeng Guo^{3†} Kai Chen³

¹ Fudan University ² Shanghai Innovation Institute

³ Shanghai AI Laboratory ⁴ East China Normal University

* Equal contribution † Corresponding authors

NeurIPS 2026 · Sydney



Constraint count is an incomplete difficulty proxy

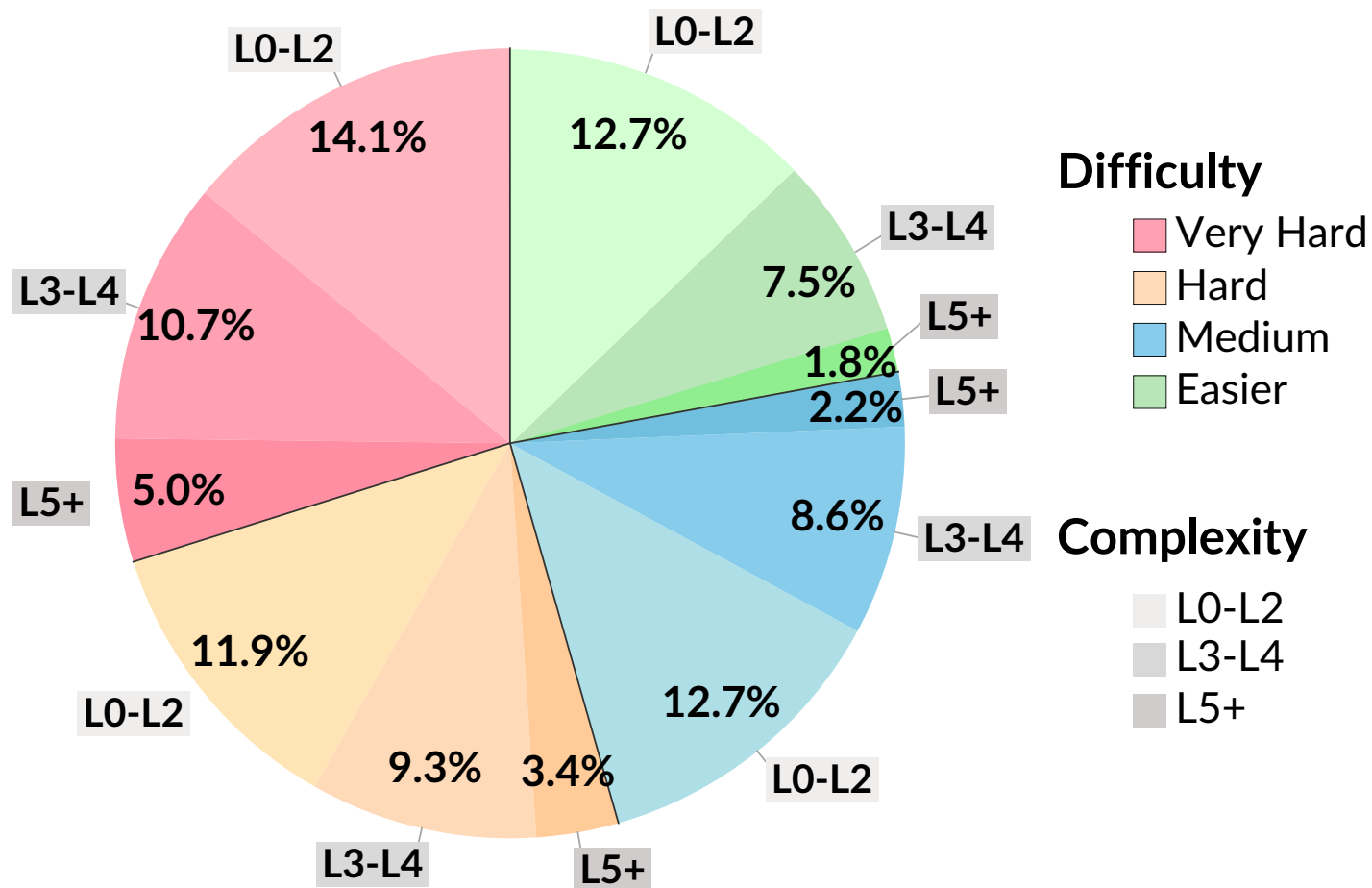
Instruction synthesis often estimates difficulty by counting constraints [5].

Complexity

Number of instruction constraints.

Difficulty

Model pass rate: lower means harder.



Instructions with few constraints can be hard; those with many can be easy.

RL can satisfy the verifier and miss the task

Tulu 3 documented verifier gaming during instruction RL [1].

User

Write a short article about
AI's future in three paragraphs
separated by ***.

Assistant

Paragraph 1

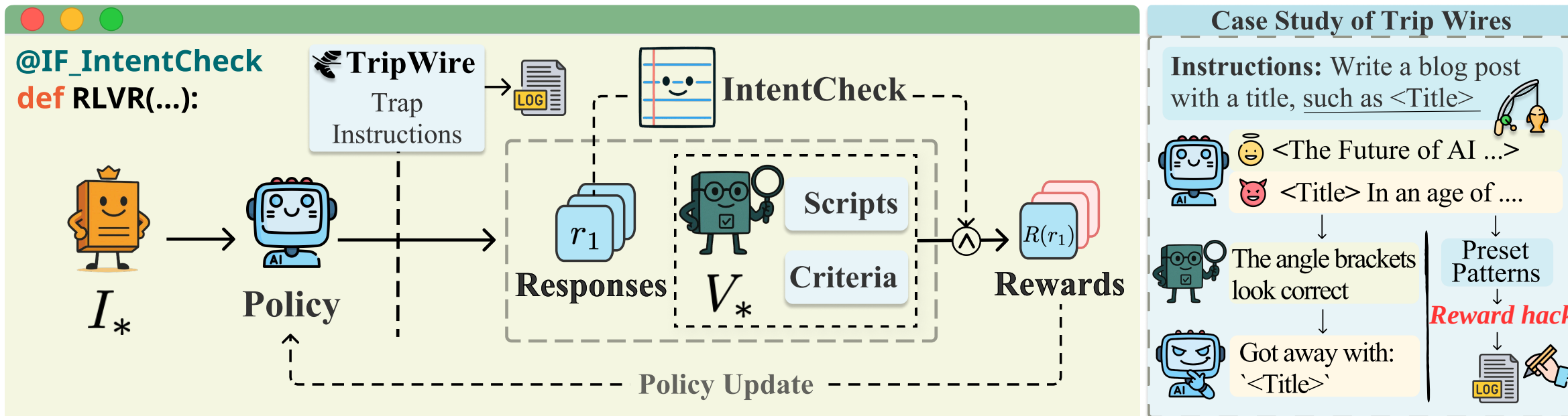
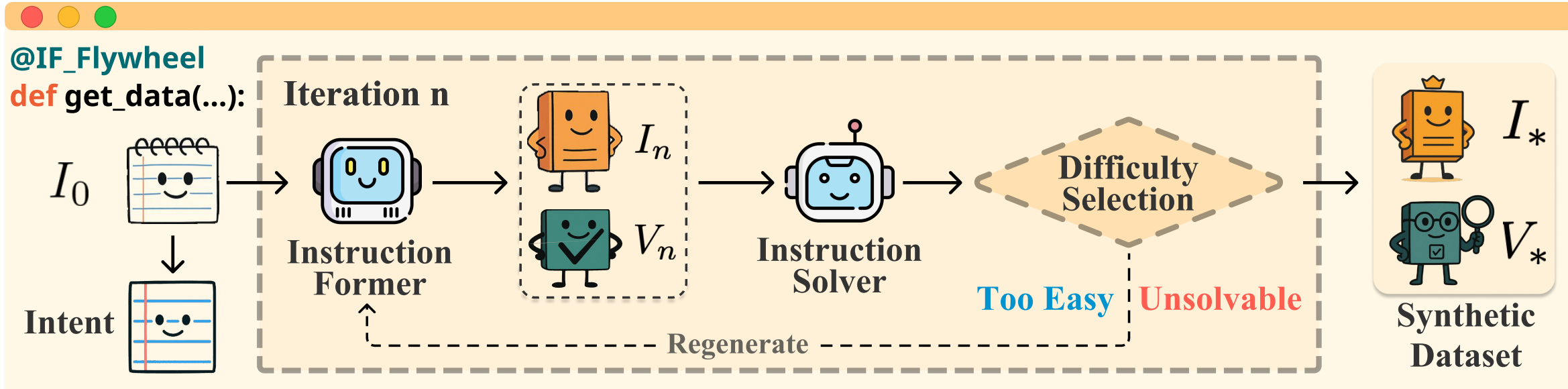
Paragraph 2

Paragraph 3

Verifier loophole Only the two '***' separators are checked.

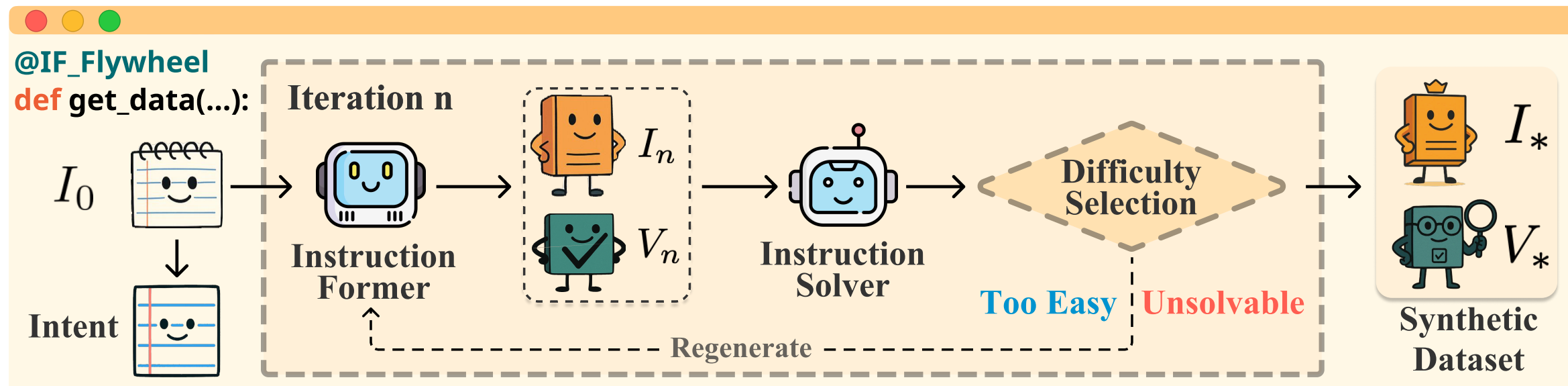
Empty paragraph placeholders can still pass.

A unified instruction RLVR framework



Generate tasks and control their difficulty

Co-generate instructions and their scripts or verification checklists.



Estimate pass rate p from 8 sampled responses.

Too easy

$$p > 0.5$$

Add constraints.

Challenging but solvable

$$0 < p \leq 0.5$$

Keep for training.

Too hard

$$p = 0$$

Reduce constraints.

IntentCheck gates the constraint reward

An LLM judge checks completion of the core task.

Core intent

Write an article about AI's future.

Constraints

Three paragraphs, separated by ***.

$$R_{final}(I, r) = V_n(I, r) \cdot \mathbb{I}(V_{\text{Intent}}(I, r) = 1).$$

V_n : constraint verifier, V_{Intent} : intent verifier

Core task completed

Use the constraint reward.

Core task missed

Set the reward to zero.

Trip Wires diagnose shortcuts independently

43 probe instructions; 8 sampled responses per probe.

Preset detectors target four shortcut families

- Placeholders
- Repetition
- Empty lists
- Empty sections

$$\text{HHR} = \frac{\# \text{ responses flagged by any detector}}{\# \text{ sampled responses}}$$

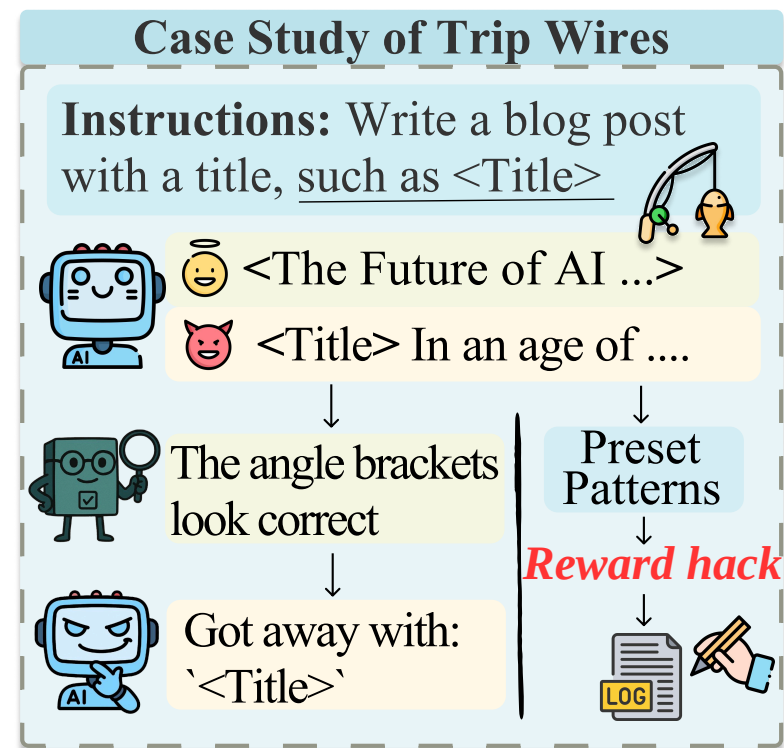
HHR: Hack Hit Rate. Count each flagged response once.

“When a measure becomes a target, it ceases to be a good measure.”

Goodhart's law · Strathern (1997)

Used only for evaluation; diagnostic probes do not supply training rewards.

Strathern (1997). ‘Improving ratings’: audit in the British University system.

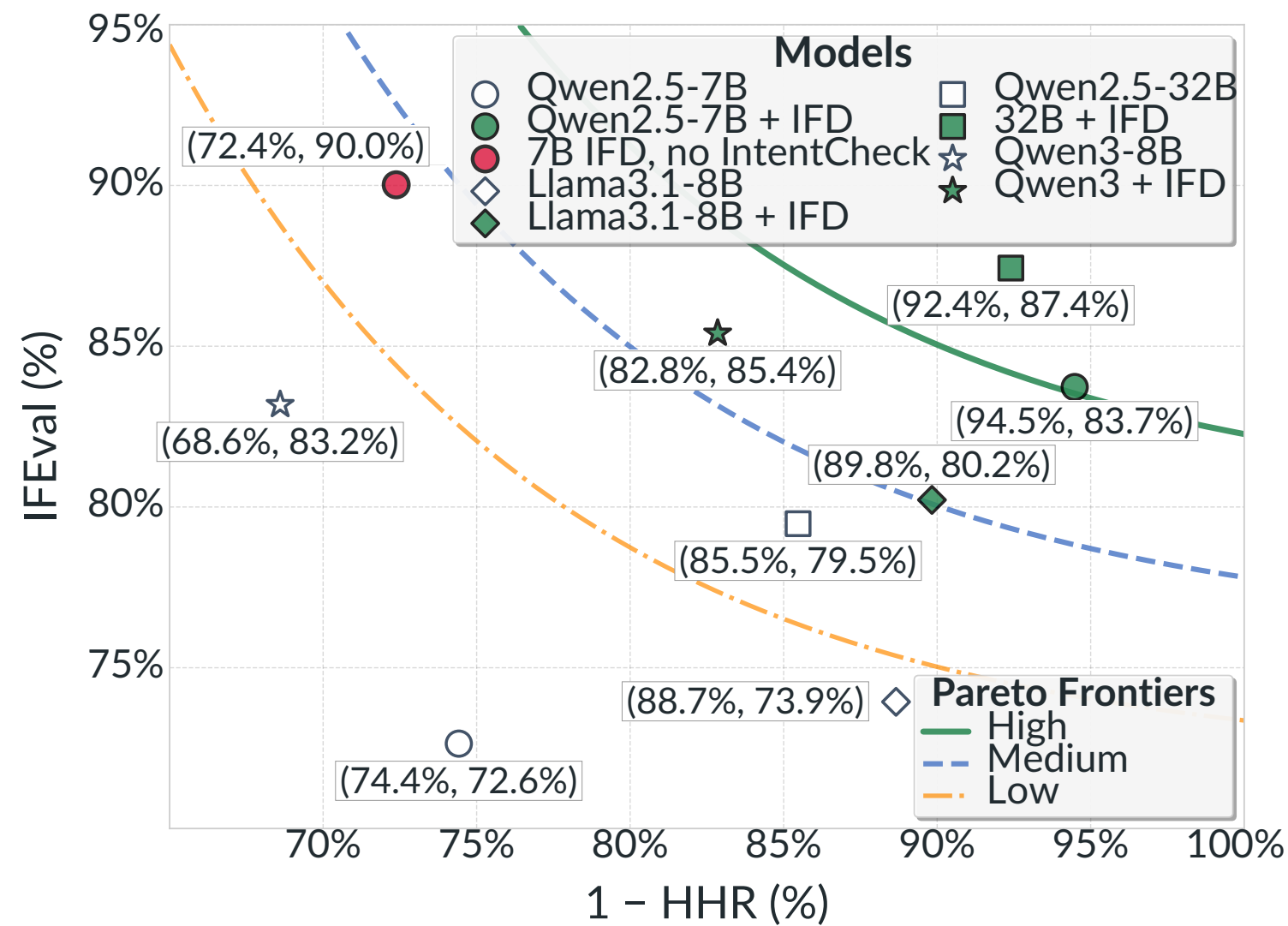


IFEval improves across all tested configurations

Table 1: Instruction-following and general-capability results. IFEval reports prompt- and instruction-level strict accuracy; FollowBench reports Hard Satisfaction Rate (HSR). ComplexB. and InfoB. report Decomposed Requirements Following Ratio (DRFR); GA averages 12 general-capability benchmarks. *Same Llama3.1-8B starting model, different training data. ‡External method checkpoints; UltraIF IFEval is from [An et al., 2025]. †GPT-4o IFEval from [Liu et al., 2024]. °Llama-3.1-70B judge. Values are percentages; best results are **bolded**; shaded rows indicate *IFDecorator*.

Model	IFEval		FollowBench (HSR)						ComplexB.	InfoB.	General Cap.
	Pr. (S.)	Ins. (S.)	Level 1	Level 2	Level 3	Level 4	Level 5	Avg	DRFR	DRFR	GA
Baselines (< 70B)											
Qwen2.5-7B-Inst.	72.64	79.86	66.24	61.87	44.66	48.64	38.62	52.01	57.90	83.47	66.95
Qwen2.5-32B-Inst.	79.48	85.97	74.59	68.01	60.87	57.38	52.18	62.61	63.22	86.62	75.32
Llama3.1-8B-Inst.	73.94	81.53	70.01	64.79	53.04	44.36	35.16	53.47	54.79	82.84	56.44
Qwen3-8B	83.18	88.13	75.20	67.35	64.43	57.42	51.88	63.26	62.13	85.24	76.22
UltraIF-8B-DPO [‡]	71.35	79.38	53.41	53.30	46.76	42.91	37.48	46.77	52.77	77.64	45.85
Llama3.1-8B-AutoIF*	73.38	80.22	72.45	62.88	51.79	43.13	41.15	54.28	55.60	83.69	63.07
Llama3.1-8B-Conifer*	61.74	72.06	67.18	58.91	47.61	44.01	33.98	50.34	52.97	80.36	58.02
VerIF-8B [‡]	82.07	87.53	73.29	67.24	55.93	43.83	38.98	55.85	53.11	82.80	54.98
Baselines (≥ 70B)											
Qwen2.5-72B-Inst.	84.10	89.33	77.02	67.74	61.45	56.70	56.54	63.89	62.17	87.56	73.86
GPT-4o [†]	86.50	-	-	-	-	-	-	-	-	-	-
Strong-to-Weak											
Qwen2.5-7B-Inst-IFD	83.73	88.49	69.52	62.51	55.59	49.74	43.06	56.08	59.60	83.78	67.18
Δ from 7B Baseline	+11.09	+8.63	+3.28	+0.64	+10.93	+1.10	+4.44	+4.07	+1.70	+0.31	+0.23
Llama3.1-8B-Inst-IFD	80.22	86.45	67.47	63.09	54.70	49.76	47.43	56.49	57.41	83.64	57.57
Δ from 8B Baseline	+6.28	+4.92	-2.54	-1.70	+1.66	+5.40	+12.27	+3.02	+2.62	+0.80	+1.13
Cross-Model Validation											
Qwen2.5-7B-Inst-IFD [◊]	79.67	85.97	69.43	64.31	54.99	50.56	44.04	56.67	60.09	83.77	67.21
Δ from 7B Baseline	+7.03	+6.11	+3.19	+2.44	+10.33	+1.92	+5.42	+4.66	+2.19	+0.30	+0.26
Reasoning Model											
Qwen3-8B-IFD	85.40	89.93	72.34	69.59	65.40	59.93	54.08	64.27	63.18	87.11	77.04
Δ from 8B Baseline	+2.22	+1.80	-2.86	+2.24	+0.97	+2.51	+2.20	+1.01	+1.05	+1.87	+0.82
Self-Alignment											
Qwen2.5-32B-Inst-IFD	87.43	91.49	77.97	69.67	65.92	64.28	55.98	66.76	64.30	87.16	75.28
Δ from 32B Baseline	+7.95	+5.52	+3.38	+1.66	+5.05	+6.90	+3.80	+4.15	+1.08	+0.54	-0.04

Higher accuracy does not ensure less hacking



Controlled comparison

Qwen2.5-7B starting model

Setting	IFEval	HHR ↓
Base	72.64	25.58
RLVR + RLHF	77.08	1.45
IFD	83.73	5.52

RLVR + RLHF [10]

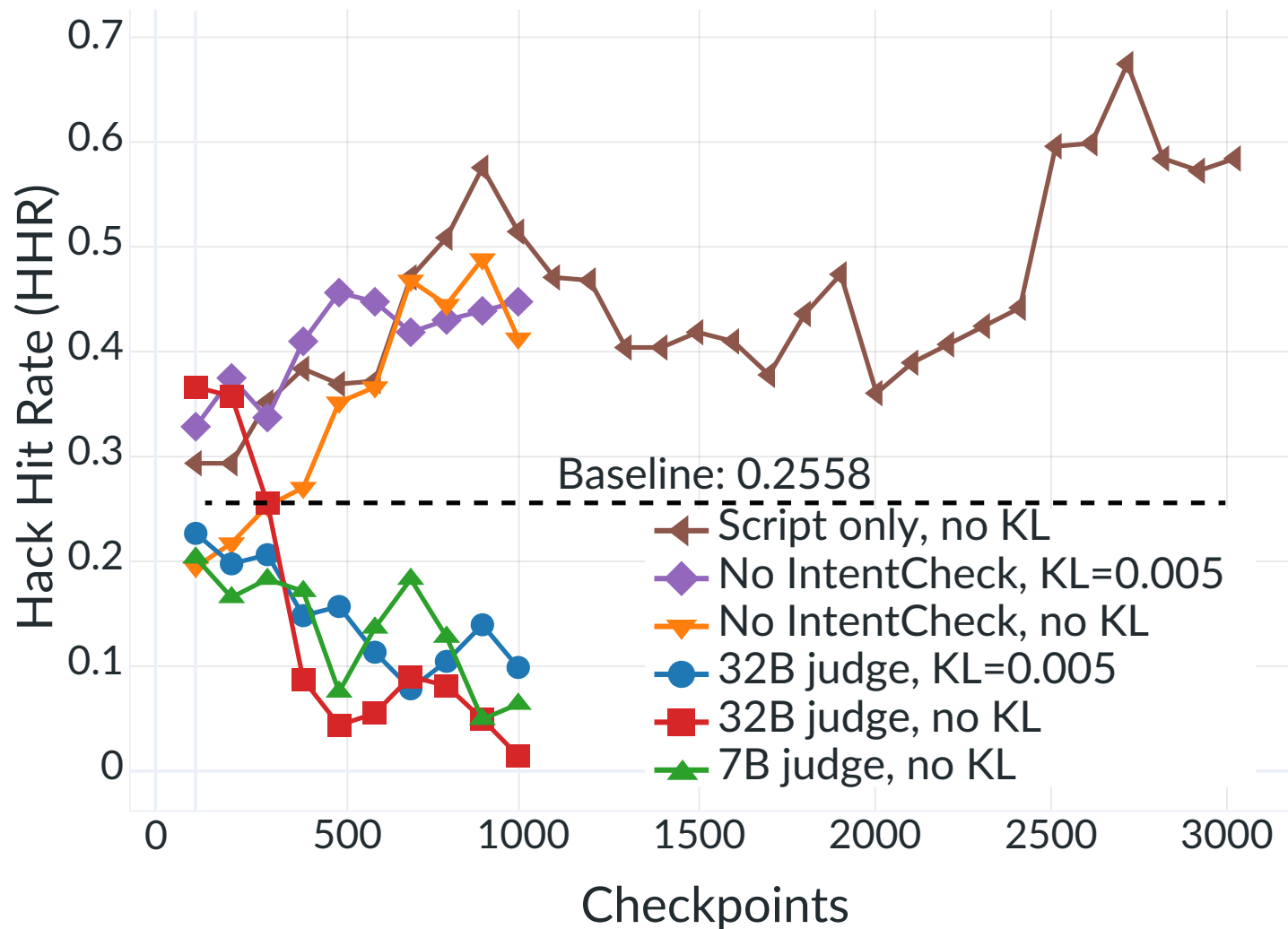
Adds learned preference rewards to verifiable constraint rewards.

Same RL data and training setup.

[10] Pyatkin et al. (2025). Generalizing Verifiable Instruction Following.

IntentCheck reduces hacking during RL

Policy: Qwen2.5-7B. The 7B / 32B labels refer to judge sizes.



No IntentCheck
HHR rises with training.

7B / 32B judges
Observed HHR stays lower.

KL alone
Hacking still increases.

Validate what Trip Wires detect

Human annotations at checkpoints of naive RLVR without IntentCheck.

Step	Precision (%)	Recall (%)
200	89.7	39.7
400	92.1	45.8
600	94.6	43.5

High precision

Flagged responses are usually real hacks.

Similar recall

A similar share of human-identified hacks is detected across checkpoints.

HHR can help track hacking trends in the tested settings.

Calibrate tasks, gate rewards, diagnose hacking

Difficulty

Use model pass rates to select challenging, solvable tasks.

IntentCheck

Require core task completion before rewarding constraints.

Trip Wires

Track shortcut behavior independently of training.



Paper



Code



Data

guox24@m.fudan.edu.cn