NEURIPS 2026

Learning to Surpass: Training Tool-Using Agents with Anchored Feedback

John Gkountouras[1,2] Fengjun Wang[2] Angelantonio Castelli[2] Satendra Kumar[2]

[1]ILLC, University of Amsterdam [2]Booking.com Work done during an internship at Booking.com

johngou.com/lts
i.gkountouras@uva.nl

Outcomes without traces

For many agent tasks we have good final outcomes but not the tool calls that produced them, for example curated itineraries, resolved support tickets or merged code.

Imitation learning needs those tool calls, so they have to be imputed, and the agent can at best match the reference.

We train the agent with RL to beat the reference outcome under a comparative LLM judge, with the programmatic checks as a gate.

TripTailor

The agent gets a travel request x, searches a sandbox database for flights, trains, hotels, attractions and restaurants, one tool each, and writes a multi-day plan as JSON. Each training request has a human-written reference plan $\tilde{y}$.

What can be checked

Seven checks run by a program, test plans failing each (%)

	Qwen3-4B	Anchored
Every venue exists in the sandbox	4.6	3.4
All required fields filled in	0.1	0.9
Total cost within budget	11.4	3.3
No restaurant twice	22.3	0.9
No attraction twice	8.3	0.6
Meal prices in the requested tier	2.1	0.4
Visit lengths in the allowed range	33.1	2.1

Plans that fail at least one check 53.3% → **8.4%**

Evaluation

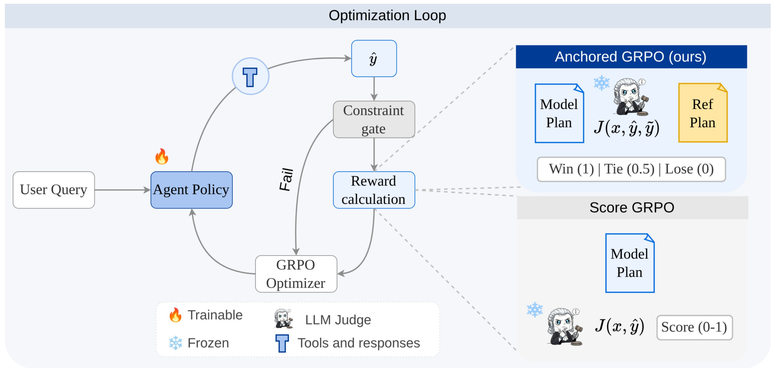
Each test plan is compared with the human reference. None of the three evaluators is used in training.

Program	GPT-4o and Claude-4.5-Haiku	Reward model
runs the seven checks	score both plans 1 to 5, in both orders	trained on TripTailor preferences
↓	↓	↓
Valid	LLM win	RM win

$$\text{Final Surpass} = \text{Valid} \wedge (\text{LLM win} \vee \text{RM win})$$

Under a held-out rubric with different criteria, anchored GRPO wins 68.0% of valid comparisons and Gemini-3.0-Pro 62.1%.

Anchored GRPO



$$s = J(x, \hat{y}, \tilde{y}) \in \{\text{loss}, \text{tie}, \text{win}\} \qquad \mathcal{C}(\tau) = \prod_j c_j(\tau) \in \{0, 1\}$$

$$R(\tau) = \mathcal{C}(\tau) \cdot r_{\text{pref}}(s), \qquad r_{\text{pref}}(s) = \begin{cases} 1 & s = \text{win} \\ 0.5 & s = \text{tie} \\ 0 & s = \text{loss} \end{cases}$$

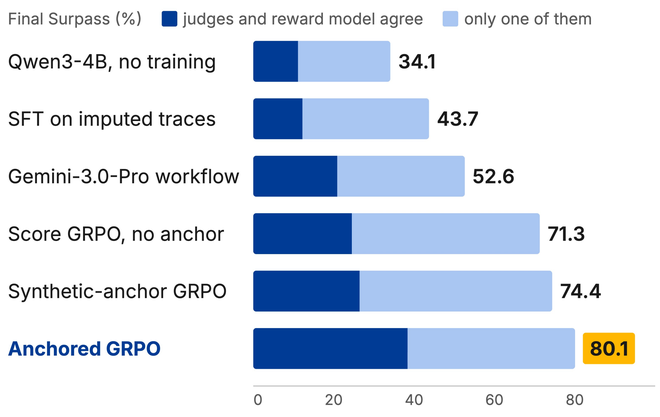
$$A^{(k)} = \frac{R^{(k)} - \bar{R}}{\sigma_R + \epsilon}, \qquad k = 1, \ldots, K$$

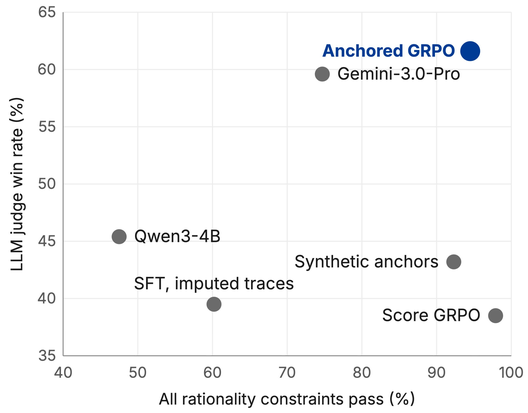
1. The policy π_θ calls tools and writes a plan $\hat{y}$.
2. A program runs the checks. If one fails, the reward is 0.
3. A frozen LLM judge J scores $\hat{y}$ and the reference $\tilde{y}$ on the same 1 to 5 rubric, which gives a win, a tie or a loss.
4. GRPO normalizes the rewards over $K = 8$ rollouts of the same request.

Judge bias

- Both plans are rendered from JSON by the same template, so length and format match.
- The order of the two plans is random.
- Gemini-2.5-Flash judges during training. Evaluation uses other judges.

Results on TripTailor



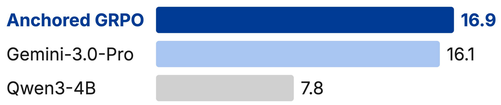


Without the anchor, Score GRPO learns the checks and loses most judge comparisons.

Human evaluation

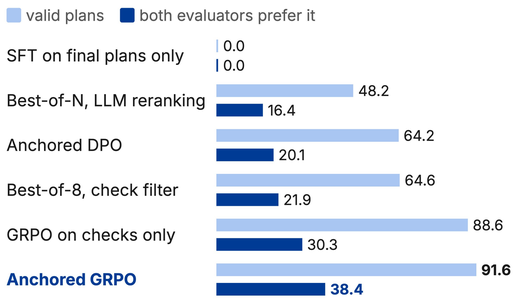
2,109 blind comparisons with the human reference on all 703 test requests, rated 1 to 5 by seven professional annotators.

Final Surpass with the human rating (%)

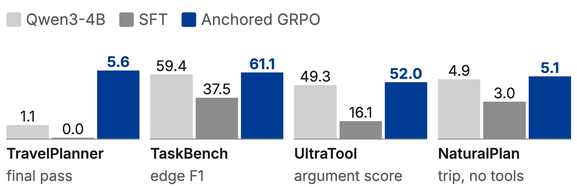


Raters are stricter than the LLM judges and rank the three systems the same way.

Other uses of the reference



Held-out benchmarks



Each panel has its own scale. τ²-bench retail pass@1 goes from 39.0 to 43.1.

Other base models

