# The Judge in the Mirror

Self-Preference in LLM Evaluators, Without Self-Recognition

Han Kim · IOV Labs (아이오브연구소)

2026

**Abstract.** Language models increasingly grade language models: on leaderboards, in reinforcement learning from AI feedback, and in agents that check their own work. All of it assumes an impartial judge. We audit that assumption on four current frontier models in two families, blind, with a consensus baseline that separates bias from genuine quality. Each model answers 24 open-ended prompts; each then judges, blind to authorship and across both presentation orders, which of two responses is better, for 1,152 pairwise comparisons. The self-preference index, a judge's win rate for its own family minus the leave-one-out consensus of the other judges on the same responses, is positive for **every** model, mean +0.14 (GPT-4o reaching +0.21), and operates at the family level rather than only the exact model. Yet the standard explanation fails: only one of the four models can identify its own outputs above chance, while all four self-prefer. The bias is implicit and stylistic, an affinity for one's own distribution, not a recognition of authorship. We also find two generic judge pathologies that dwarf careful use, a position bias (the first response wins 63% of the time) and a near-deterministic length preference (correlation 0.98), and close on evaluation as a social act and Goodhart when the judge is also a contestant.

**Keywords:** LLM-as-judge · self-preference bias · self-recognition · position bias · leaderboards · RLAIF · evaluation · reproducibility

## 1 Introduction

A growing share of machine evaluation is machines evaluating machines. Chatbot leaderboards rank models by the verdicts of a judge model [1]; reinforcement learning from AI feedback trains on a model's own preference labels; agentic systems ask a model to grade its own intermediate work before continuing. Each of these inherits whatever the judge brings to the task, and a judge is not obviously neutral about its own kind. Prior work established that LLM evaluators can recognize and favor their own generations [2], [3]. We ask the question again on current frontier models, and we ask it more precisely: is the favoring **of the self** or **of the family**, and is it driven by **recognition** or by something quieter?

The methodological hazard is that a judge preferring its own output may simply be right: the output might be better by that judge's lights. We neutralize this with a consensus baseline. A judge's verdict on its own family is scored not against the truth, which we do not have for open-ended prose, but against what every **other** judge thought of the identical responses. Self-preference is the residual a judge adds on top of the panel, with quality held fixed by construction.

## 2 Method

### 2.1 Generation and judging

Four models in two families, OpenAI (`gpt-4o-mini`, `gpt-4o`) and Anthropic (`claude-haiku-4-5`, `claude-sonnet-4-6`), each answer 24 open-ended prompts spanning explanation, advice, summary, reasoning, creative writing, and persuasion, at a creative temperature so that styles diverge. Every model then acts as a judge. For each prompt and each of the six unordered pairs of responses, the judge is shown two responses labelled A and B, **blind to authorship**, and names the better one. Both presentation orders are run, so position is a measured control rather than a confound. This yields 1,152 pairwise judgments.

### 2.2 Metrics

From the order-balanced verdicts we form, for each (judge, generator), the win rate the judge awards that generator. The **self-preference index** (SPI) for a judge is its mean win rate for its own family in cross-

family matchups minus the leave-one-out consensus win rate of the other judges on those same responses; we compute it at family and at exact-model level. A separate **self-recognition probe** asks each judge, for each response, whether it wrote it; we summarize recognition as the rate of "yes" on own outputs minus the rate of "yes" on others. We also report position bias (the probability the first-shown response wins) and the correlation between response length and win rate.

### 2.3 Confound controls

Authorship is hidden; both orders are run and position bias is reported separately; the consensus baseline holds quality fixed so SPI is self-inflation, not a quality gap; length is reported as a covariate; and family versus exact-model SPI are decomposed so that "I prefer my exact words" and "I prefer my vendor" are distinguished. The run is seeded and content-cached.

## 3 Results

### 3.1 Self-preference is real and family-wide

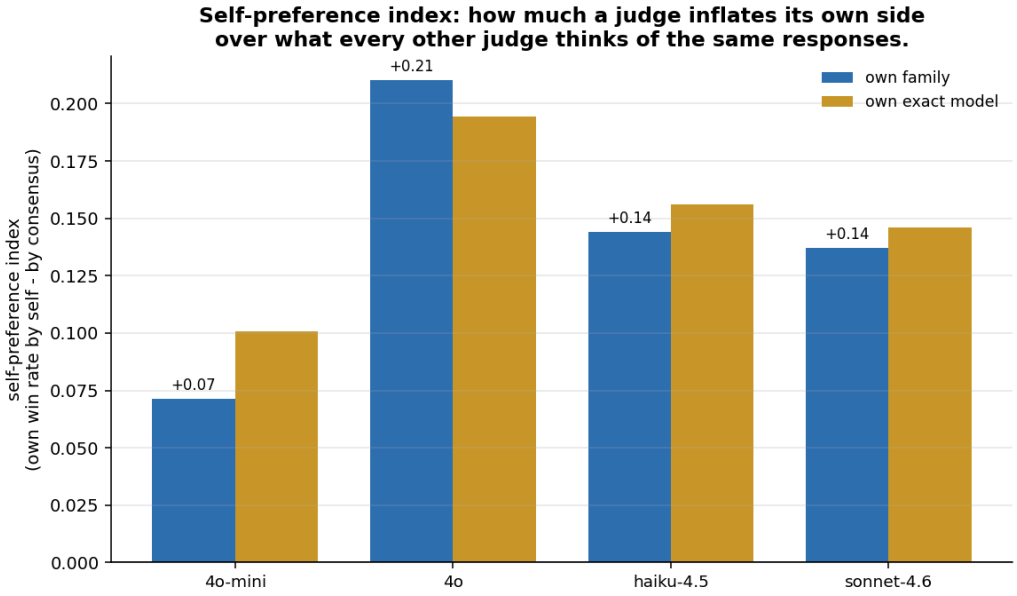


Figure 1: The self-preference index. Every judge inflates its own side over the neutral consensus, at family and at exact-model level.

Every judge shows positive self-preference. The family-level index is +0.07 for GPT-4o-mini, **+0.21** for GPT-4o, +0.14 for Haiku 4.5, and +0.14 for Sonnet 4.6, mean **+0.14**. The exact-model index is similar (+0.10 to +0.19), so the effect is not merely a model preferring its own verbatim text; it is a model preferring its **vendor's** style. The win-rate matrix (Figure 2) makes the structure visible: although the Anthropic responses are better liked by everyone here, each judge lifts its own family's cell above what the cross-family judges award the same responses. For a same-vendor leaderboard, this is a thumb of roughly fourteen points on the scale, in the house's favor.

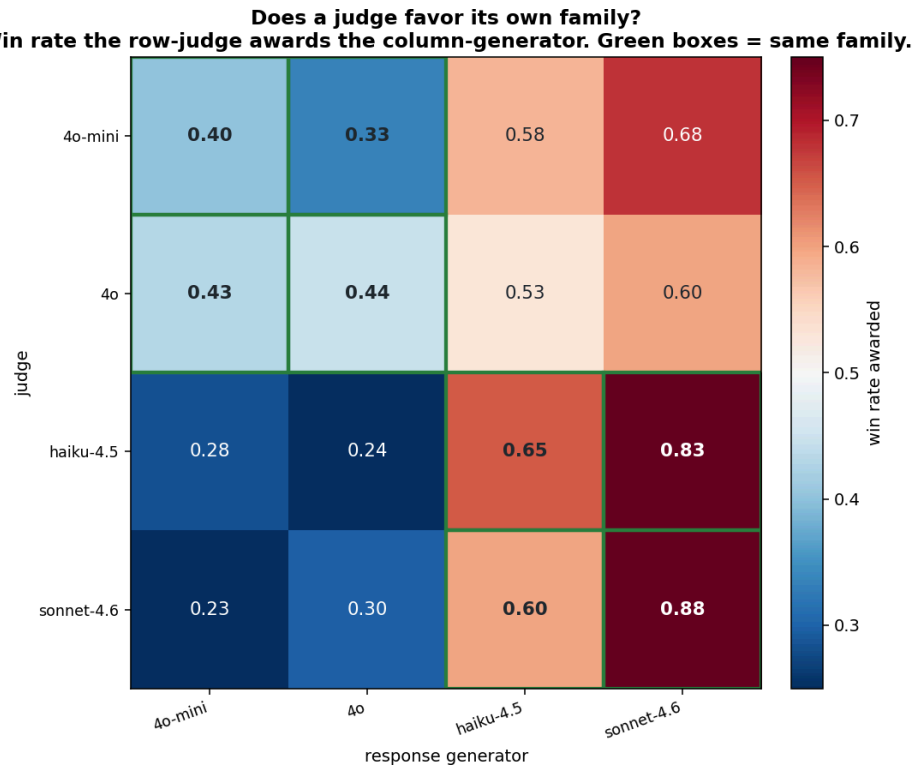


Figure 2: Win rate the row-judge awards the column-generator. Green boxes mark same-family cells; the own-family columns are lifted relative to the cross-family judges.

## 3.2 The bias is not explained by self-recognition

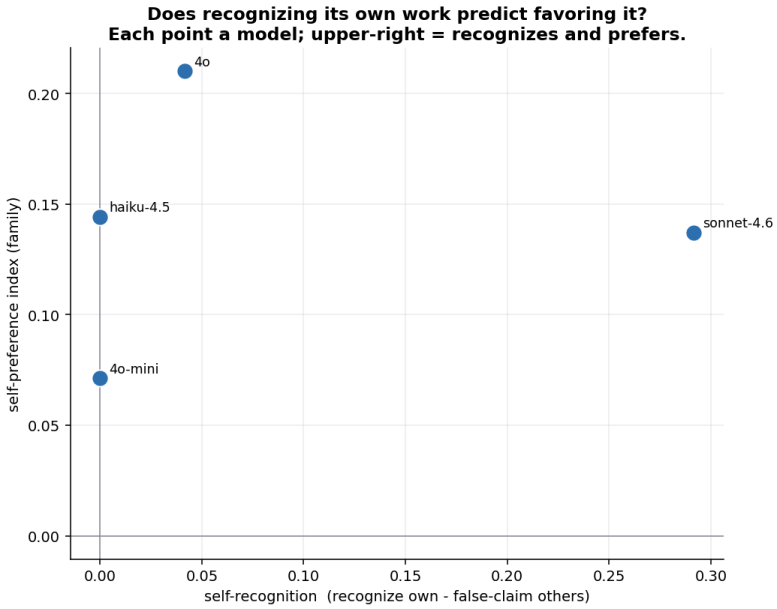


Figure 3: Recognition versus preference. Three of four models recognize their own outputs at chance and still self-prefer.

The leading account of self-preference is recognition: a model favors its outputs because it can tell they are its own [2]. On these models the link breaks. Only Sonnet 4.6 identifies its own work above chance (a recognition discrimination of +0.29, saying "yes" to 88% of its own outputs and 58% of others). The other three almost never claim authorship of anything, a recognition discrimination near zero, and yet self-prefer as strongly or more. Self-preference here is therefore **implicit**: a pull toward responses drawn from one's own training distribution, not a deliberate recognition of authorship. This is a negative result against the

recognition mechanism on current models, and it matters, because a debiasing strategy aimed at suppressing recognition would not touch a bias that does not run through it.

### 3.3 Two generic biases dwarf careful use

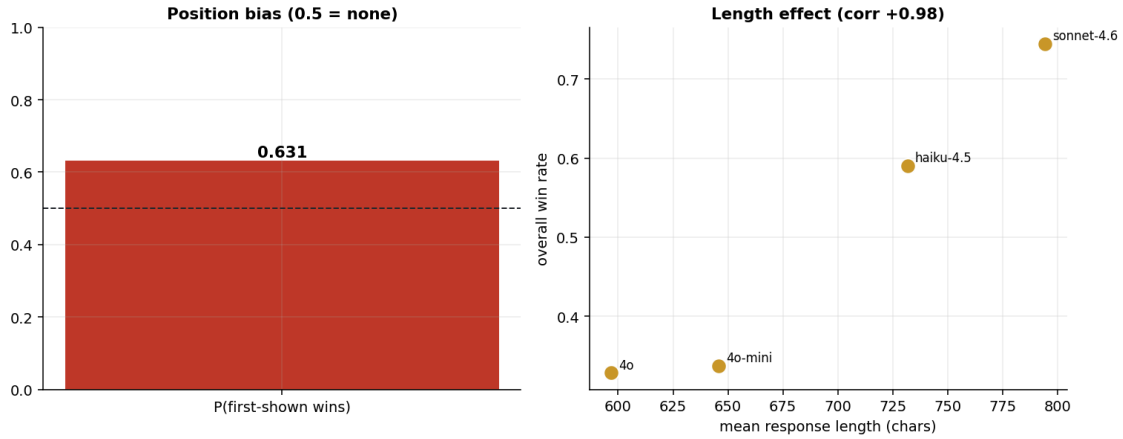


Figure 4: The confounds we ruled in: the first-shown response wins 63% of the time, and length predicts the verdict almost deterministically.

Before self-preference is even in view, two cruder biases dominate. The first-shown response wins **63%** of the time, a position bias large enough to flip many verdicts on its own, which is why both orders must be run. And response length correlates with win rate at **0.98**: longer is better, almost without exception. Any evaluation pipeline that fixes the order or fails to control length is measuring presentation, not quality, and the self-preference on top of that is the subtler of the model's thumbs on the scale.

## 4 Discussion

### 4.1 Evaluation is a social act

To judge is to occupy a position, and a model's position is its training distribution. What looks like fairness from inside that distribution is partiality from outside it. The implicit character of the bias, present even where recognition is absent, suggests it is not a strategy the model runs but a shape the model has: prose that matches its own priors reads as better. The remedy is not to ask a model to try harder to be fair, which leaves the shape intact, but to **triangulate**: a cross-family panel, with order randomized and length controlled, so that no single distribution sets the standard.

### 4.2 Goodhart when the judge is also a contestant

A measure becomes a target and ceases to be a good measure [4]. The pathology sharpens when the target **is** the measurer. A model optimized against its own preference signal, in RLAIF or in agentic self-checking, is optimizing toward its own distribution by a quantity we here put near fourteen points. The same tilt that flatters a vendor on a leaderboard lets an agent wave its own work through. The honest design assumes the judge is partial and builds the panel that a partial judge requires.

## 5 Limitations

**Pilot scale.** Twenty-four prompts, four models, two families; the consensus baseline is three judges, itself a small panel. **Open-ended tasks.** On verifiable tasks, where quality is less subjective, self-preference may shrink or change character. **Length.** The strong length effect means part of the raw win-rate spread is quality and verbosity, not bias; the SPI controls for this by construction, but the absolute family win rates should be read with it in mind. **Two families**, verbalized pairwise choices. The broken recognition link and all null results are reported.

## 6 Conclusion

Asked to judge, a current language model is a mirror that flatters its own reflection, lifting its family above the neutral panel by about fourteen points, and it does so without recognizing the face as its own. The bias is implicit, stylistic, and family-wide, which makes same-vendor leaderboards and self-grading agents structurally generous to themselves. Two blunter biases, position and length, sit beneath it. Objectivity in machine evaluation is not a property of a better judge; it is a property of a panel, assembled across families, with the crude knobs held still.

---

Code, data, the pre-registered design, and the cached judgments: https://github.com/hankimis/self-preference. MIT licensed.

## References


[1] L. Zheng and others, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." 2023.

[2] A. Panickssery, S. R. Bowman, and S. Feng, "LLM Evaluators Recognize and Favor Their Own Generations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[3] K. Wataoka, T. Takahashi, and R. Ri, "Self-Preference Bias in LLM-as-a-Judge." 2024.

[4] C. A. E. Goodhart, "Problems of Monetary Management: The U.K. Experience," *Monetary Theory and Practice*, 1984.