

JEV-as-a-Judge

Accept when confident, escalate when unsure

Yubo Li

Carnegie Mellon University

WITH Yidi Miao · Ramayya Krishnan · Rema Padman

Paper: [alphaxiv.org/abs/2609.26550](https://arxiv.org/abs/2609.26550)



Yubo Li

Ph.D. candidate (ABD), Information Systems
Carnegie Mellon University

Co-advised by **Rema Padman** and **Ramayya Krishnan**
CMU–NIST AI Measurement Science & Engineering Center
(AIMSEC)

yubol-bobo.github.io

WHAT I WORK ON

LLM evaluation

Robustness and reliability of LLMs, multi-turn consistency, **LLM-as-a-judge**

Agentic AI

Long-horizon reasoning, tool use, fast decision models for agents

Clinical AI

Disease progression, explainable risk prediction, agentic QA for transplant patients

INDUSTRY

Netflix · 2026

ML Research Scientist Intern, Personalization Foundations

Amazon · 2025

Applied Scientist Intern, Generative Recommendation

SELECTED WORK

Time-To-Inconsistency (ICLR 2026) · Firm or Fickle? (ACL 2025) · Beyond Single-Turn survey (TMLR 2025) · SkillsBench (NeurIPS 2026)

Three ways to judge a model's output, three open questions.



≤ 2022

Free text is graded by **regex and string matching**, or by **human raters**.

STAGE 1

Regex + human raters

- **Costly**: every judgment needs a person
- **Slow**: long annotation cycles
- **Hard to scale up**

How do we judge free text at scale?



2023 →

The watershed: LLM-as-a-judge eases the bottleneck (MT-Bench and Chatbot Arena, Zheng et al., 2023), and evaluation scales up.

STAGE 2

LLM-as-a-judge

- + **Scales up**
- + **Far faster** than humans: 1.9 s per judgment
- **Still not cheap** at scale: \$12.2k per million
- **Known issues**: position bias, overconfidence

Does every judgment need a reasoning call?



2026

JEV-as-a-judge: a decision-only judge returns a typed verdict and label probabilities, no text.

STAGE 3

JEV-as-a-judge

- + **Scales up**
- + **Cheaper, faster**: \$44 per million, 0.15 s
- + **Calibrated confidence** by design: label probabilities, no generated text

Does it know when it is wrong?

Intro

JEV

Setup

Results

Analysis

Confidence

Cascade

Limitations

Wrap-up

What is JEV, and what do we ask of it?

— JEV • WHAT IT IS

JEV-as-a-judge

JEV is a **decision-only model** from **TypeSafe** (we use `jev-1.13.0`): you give it a state and a typed question, and it returns a **typed decision** with a **confidence**, and no generated text.

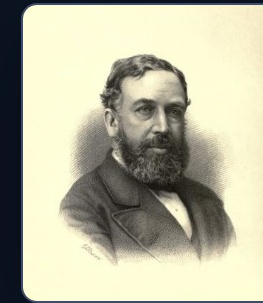
TypeSafe calls it a “System 1” model: built to be consumed by software rather than read by people.

JEV • WHY THE NAME

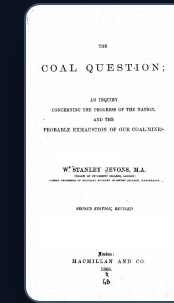
DEFINITION

The Jevons paradox is an economic phenomenon where technological improvements that increase the efficiency of a resource's use lead to an **increase, rather than a decrease**, in total consumption of that resource.

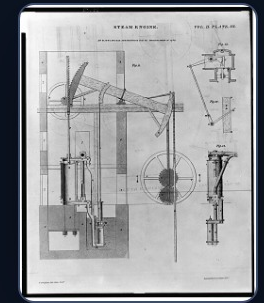
Wikipedia



W. Stanley Jevons (1835–1882)



The Coal Question, 1865



Watt's steam engine, 1782

1865 • coal

Jevons' original case, *The Coal Question*

Steam engines become more efficient



Less coal per unit of work



Steam power gets cheaper



More industries mechanize



Total coal use rises

Today • intelligence

similarly: a cheaper model leads to more AI use

Cheaper, faster model calls



More intelligence per dollar



Tasks once not worth an LLM call become worth it



Model calls embedded throughout software



Total AI use rises

"Jev is meant to be optimized for intelligence per dollar, **hence the name Jev.**"

— Diogo Almeida, TypeSafe CEO

— JEV • RESEARCH QUESTION

Can JEV replace an LLM as a judge?

A typed question in; a decision with **probabilities** out.

Choice

Pick one of up to 255 labels.

Returns the choice, a probability per label, and a confidence.

Noul

A yes/no question.

Returns one number: the probability of yes. No confidence.

Score

An ordered scale of 2–10 levels.

Returns a probability per level, their weighted average, and a confidence.

REQUEST · A CHOICE QUESTION

```
{
  "sky_color": {
    "type": "choice",
    "instructions": "What color is the sky?",
    "state": "(more context)",
    "criteria": {
      "indigo": "",
      "electric blue": "",
      "baby blue": "",
      "gray": "",
      "lavender": "",
      "salmon": "",
      "seafoam green": ""
    }
  }
}
```

RESPONSE · HOVER 0.82

```
{
  "model": "jev-1.13.0",
  "answers": { "sky_color": {
    "choice": "baby blue",
    "confidence": 0.82,
    "probabilities": {
      "indigo": 0,
      "seafoam green": 0,
      "lavender": 0,
      "salmon": 0,
      "baby blue": 0.85,
      "gray": 0.01,
      "electric blue": 0.14
    }
  } },
  "usage": { "input_tokens": 326,
    "output_tokens": 76 },
  "evaluation_time_ms": 79.5
}
```

A real request and response from the TypeSafe playground: seven labels, one call, 79.5 ms. Generative judges get the same input and must return a verdict and a probability per label, with no rationale. **In this talk, confidence = the largest label probability**, the same for every judge (here 0.85).

Intro

JEV

Setup

Results

Analysis

Confidence

Cascade

Limitations

Wrap-up

Experiment setup

Three kinds of judgment on real benchmark items: **5,172** base judgments.

PAIRWISE PREFERENCE

RewardBench (1,500 pairs) + JudgeBench (350)

What is an antonym for the word “cogent”?

Response A

unconvincing

Response B

The antonym of the word “cogent” is “irrelevant”.

labels: A · B

gold: A

JEV: **A 0.99** ✓

EVIDENCE-GROUNDED FACTUALITY

HaluEval QA (3,000 answers)

Are Ima Robot and Of Montreal music bands?

Evidence

Ima Robot is a band based in Los Angeles ... of Montreal is an American experimental pop band from Athens, Georgia.

Candidate answer

yes

labels: supported · hallucinated

gold: **supported**

JEV: **supported 1.00** ✓

FINAL-ANSWER ADJUDICATION

150 model replies + 172 controls

Find all c in $\mathbb{Z}_3$ such that $\mathbb{Z}_3[x]/(x^2 + c)$ is a field. A.
0 B. 1 C. 2 D. 3

Trusted reference

B

Model reply

... The correct answer is B. 1 and C. 2.

labels: correct · incorrect · no_answer

gold: **incorrect**

JEV: **incorrect 0.93** ✓

16 baseline judges, from GPT-4.1 mini to GPT-6.

OpenAI

GPT-4.1 mini

GPT-4.1

GPT-5.2

GPT-5.4

GPT-5.6 Sol

GPT-6 Astra

Anthropic · Google

Claude Sonnet 5

Gemini 3 Flash

Gemini 3.1 Pro

Open weights

GPT-OSS 120B

Qwen3.6 27B

Qwen3.8 27B

Qwen3 32B (local)

Qwen3.5 27B (local)

Reward models

PairRMSkywork-Reward-V2 8B

GPT-6 Astra (highlighted) is the reference judge. Every LLM judge gets JEV's input and returns a verdict with a probability per label, no rationale.

Replacing a judge means the **same accuracy** at **lower latency** and a **lower price**.

01

ACCURACY

Agreement with labels

$$\text{Acc} = \frac{1}{N} \sum_i \mathbb{1}[\hat{y}_i = y_i]$$

$$\Delta = \text{Acc}_{\text{JEV}} - \text{Acc}_{\text{other}}$$

Invalid output counts as wrong. Δ on the same items, 95% source-cluster bootstrap CI.

02

FEE

Dollars per 1,000 judgments

$$\text{Fee} = 1000 \cdot \text{mean}_i (p_{\text{in}} T_{\text{in},i} + p_{\text{out}} T_{\text{out},i})$$

T : reported tokens; p : list price at collection. JEV: $p_{\text{in}} = \$0.042/\text{M}$, $p_{\text{out}} = 0$.

03

LATENCY

Seconds per judgment

$$\text{Lat} = \text{median}_i (t_{\text{done},i} - t_{\text{sent},i})$$

End to end, same client, matched 120-judgment panel. GPT-6: 1.89 s.

Intro

JEV

Setup

Results

Analysis

Confidence

Cascade

Limitations

Wrap-up

How accurate is JEV, and at what price?

On par on ordinary preference; **15 points behind** on reasoning-heavy correctness (JudgeBench).

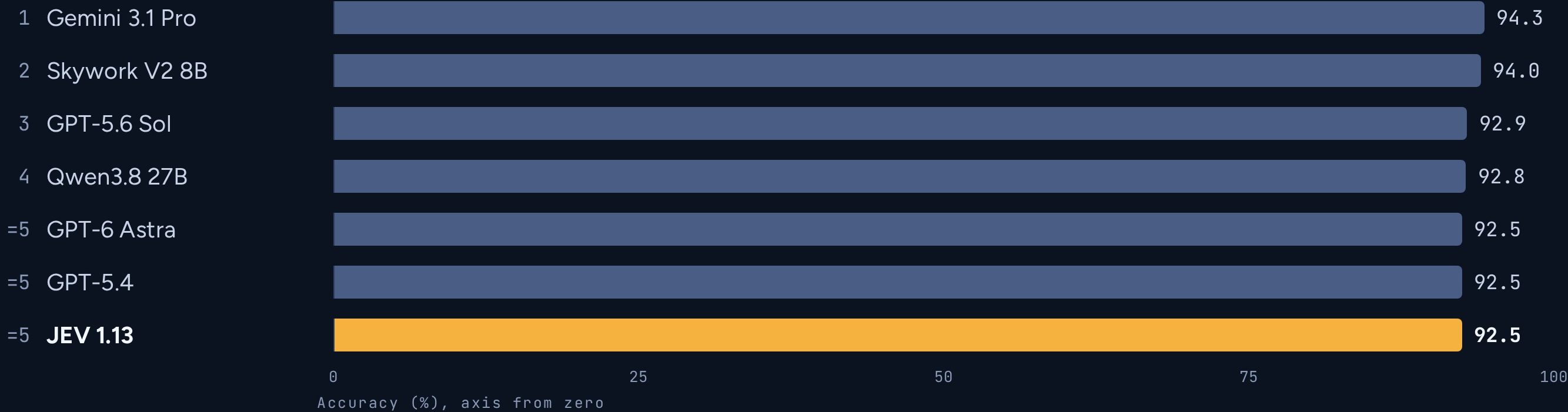
RewardBench

JudgeBench

HaluEval

Final answer

Ordinary preference: JEV ties GPT-6 and GPT-5.4 at 92.5, rank 5 of 17.



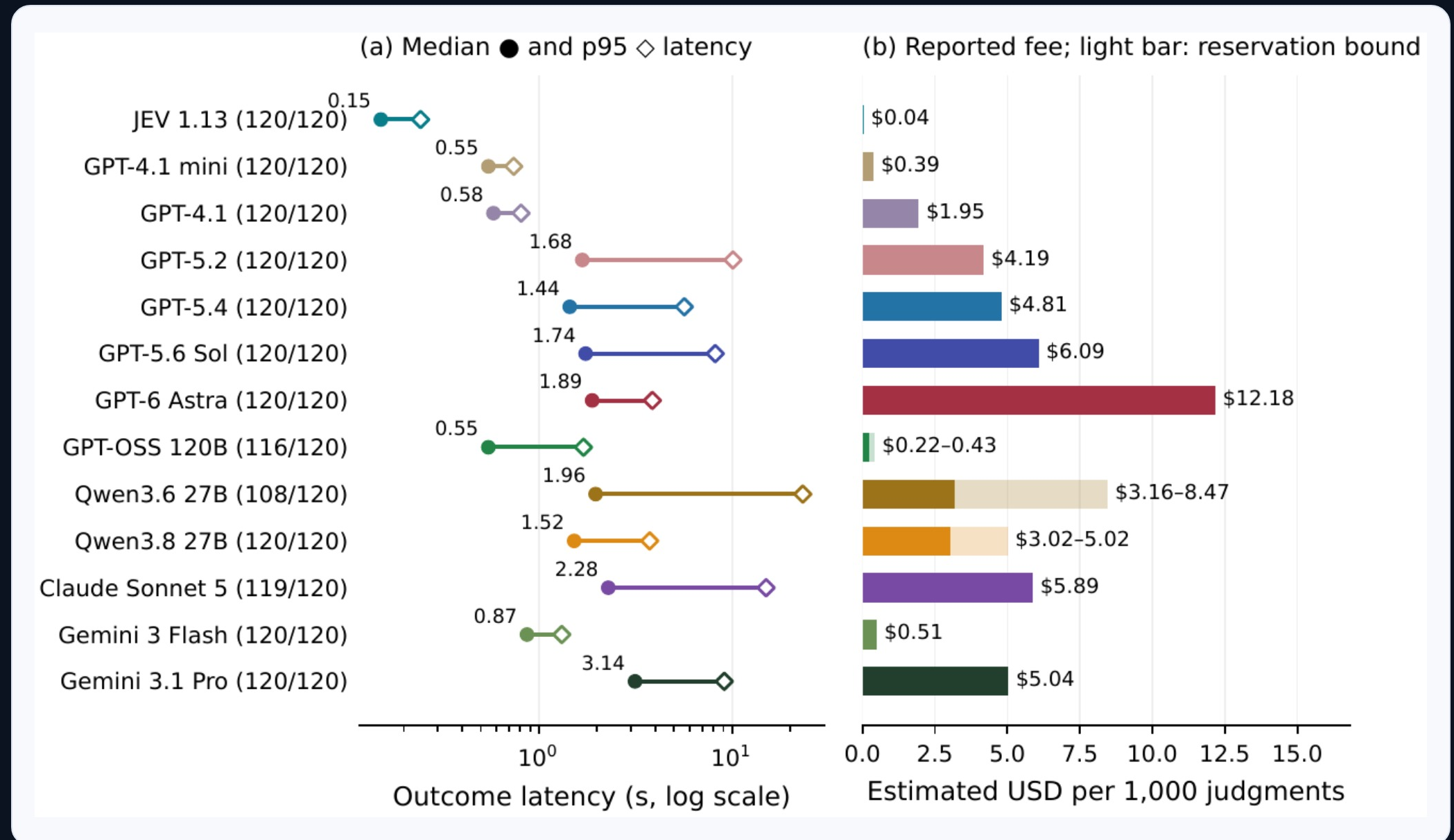
Base accuracy, invalid outputs counted as errors. Ranks are shared on ties, and a tie at 7th shows every tied judge; JEV is added below when it falls outside the top 7. The two reward models judge only preference pairs. Keys 1–4 switch benchmark; hover a bar for its rank.

RESULTS • FEE AND LATENCY

Mid-pack in accuracy, at the **far end** in cost and speed.

277×

cheaper than GPT-6: \$0.044 vs \$12.18
per 1,000 judgments; 0.15 s vs 1.89 s
median latency



Among judges under \$1 per 1,000 judgments, JEV is the most accurate on all three public benchmarks. Matched timing panel: 120 judgments per hosted judge, one model group at a time; fees from reported usage at collection-time prices.

Intro

JEV

Setup

Results

Analysis

Confidence

Cascade

Limitations

Wrap-up

Where does JEV fall behind, and why?

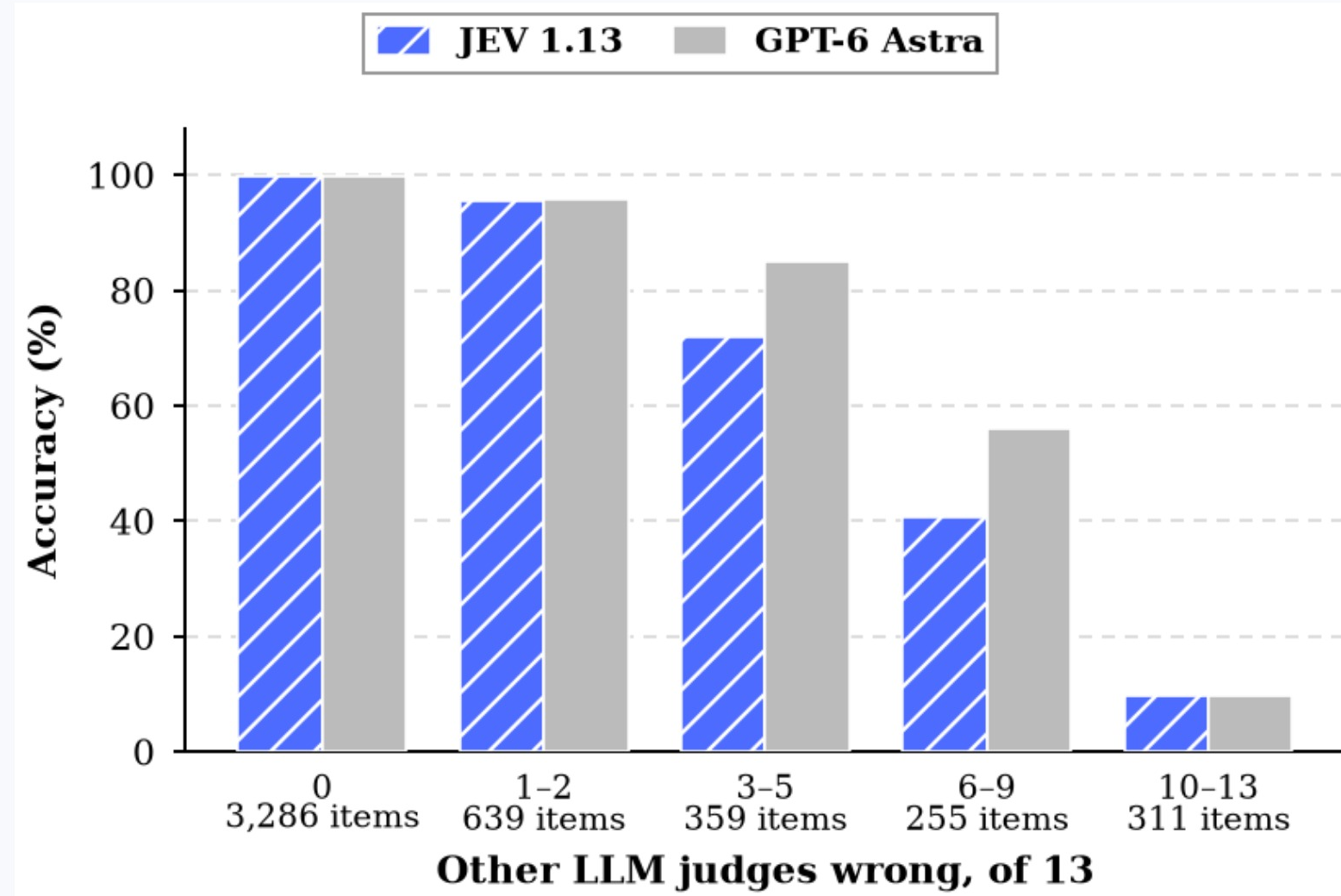
The whole gap sits at **middle difficulty**, on the items reasoning models still solve.

99.8%

JEV on the 3,286 items all 13 other LLM judges get right (GPT-6: 99.7%)

HUMAN CHECK

The 183 most disputed items were re-labeled blind and checked for agreement: on JudgeBench the human sided with GPT-6 **57 : 1**, so the gap is real, not label noise.



Difficulty = how many of the other 13 LLM judges (JEV and GPT-6 excluded) miss the item; RewardBench, JudgeBench, HaluEval. Errors that are JEV's alone are rare: of 3,741 items that at least 12 of 13 get right, JEV misses 19 and GPT-6 25.

Hold the content fixed; change only the **writing style**.

DATA FORMAT · huggingface.co/datasets/THU-KEG/RM-Bench

```
{
  "id": ..., // unique identifier
  "prompt": ..., // the prompt given to the model
  "chosen": [
    "resp_1", // concise
    "resp_2", // detailed, plain text
    "resp_3"  // detailed, Markdown
  ],
  "rejected": [
    "resp_1", // concise
    "resp_2", // detailed, plain text
    "resp_3"  // detailed, Markdown
  ],
  "domain": ... // chat, code, math,
                // safety-refuse, safety-response
}
```

Each prompt has a **chosen** (correct) and a **rejected** (subtly wrong) answer, each written in three styles by RM-Bench’s authors.

		WRONG ANSWER WRITTEN →		
		Concise	Detailed	Detailed + Markdown
CORRECT ANSWER WRITTEN ↑	Concise	normal	hard	hard
	Detailed	easy	normal	hard
	Detailed + Markdown	easy	easy	normal
		hard the wrong answer is the more elaborate one	normal same style	easy the correct one is more elaborate

1 All nine pairings, both orders

500 prompts (chat, code, math, safety) × 9 × 2 = **9,000** judgments per judge.

2 JEV vs SOTA LLM (GPT-6)

Same instructions and output contract; nothing refit. Designed after the main results, so exploratory.

Two smaller follow-ups, four-way selection on RewardBench 2 (100 prompts) and long free-text prose (600 judgments), are summarized in the backup. Hover a domain for an example.

Same wrong answer, dressed up: JEV **changes its mind**.

STYLE-MATCHED PAIR

Simplify $18/17 \cdot 13/24 \cdot 68/39$

Correct · concise

We multiply the numerators: $18 \cdot 13 \cdot 68 = 15912$, and the denominators: $17 \cdot 24 \cdot 39 = 15912$. The fraction simplifies to 1.

Wrong · concise

Factorize ... cancel 17, 13, 3 ... After canceling: $2 \cdot 3 \cdot 4 / 8 = 24 / 8 = 3$.

JEV picks the correct answer, **0.86**. GPT-6: correct, 1.00.

STYLE-ADVERSARIAL PAIR

Simplify $18/17 \cdot 13/24 \cdot 68/39$

Correct · concise (same text)

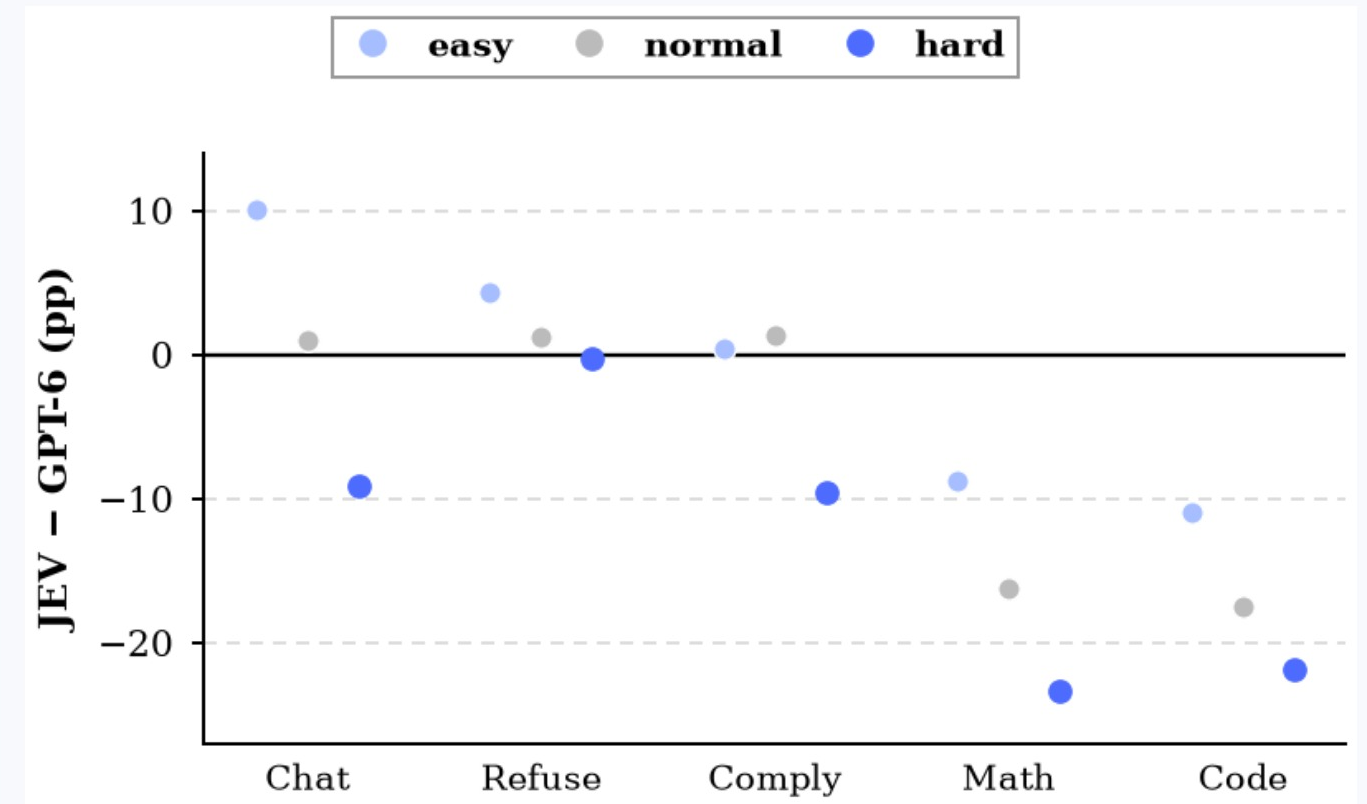
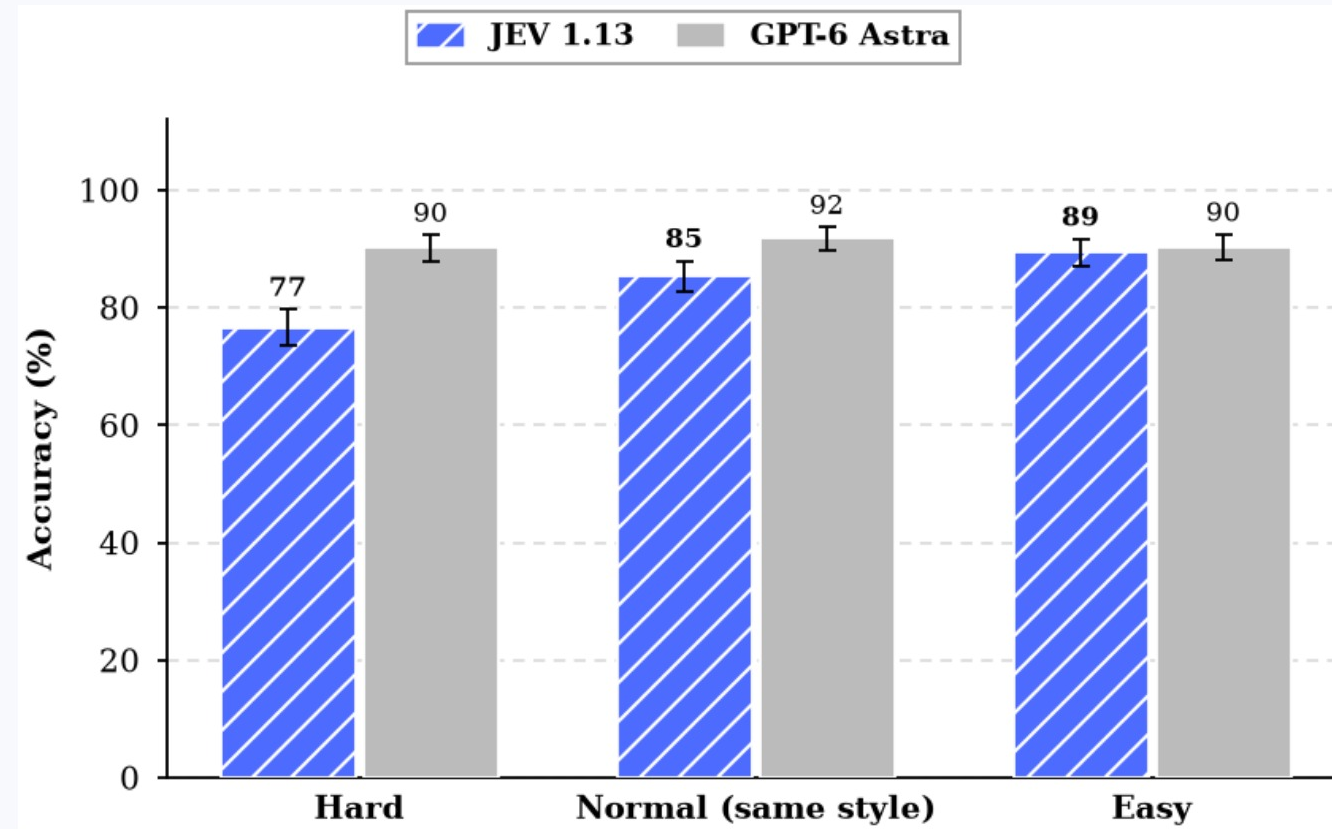
We multiply the numerators: $18 \cdot 13 \cdot 68 = 15912$, and the denominators: $17 \cdot 24 \cdot 39 = 15912$. The fraction simplifies to 1.

Wrong · detailed markdown

1. **Multiply the fractions together:** $(18 \cdot 13 \cdot 68) / (17 \cdot 24 \cdot 39)$
 2. **Factorize the numbers to identify common factors:**
 - $18 = 2 \cdot 3^2$ • $68 = 2^2 \cdot 17$ • $24 = 2^3 \cdot 3$ • $39 = 3 \cdot 13$
 3. **Rewrite with prime factorizations ...**
 4. **Cancel out the common factors:** 17, 13, 3, 2 ...
- The simplified form is 3.

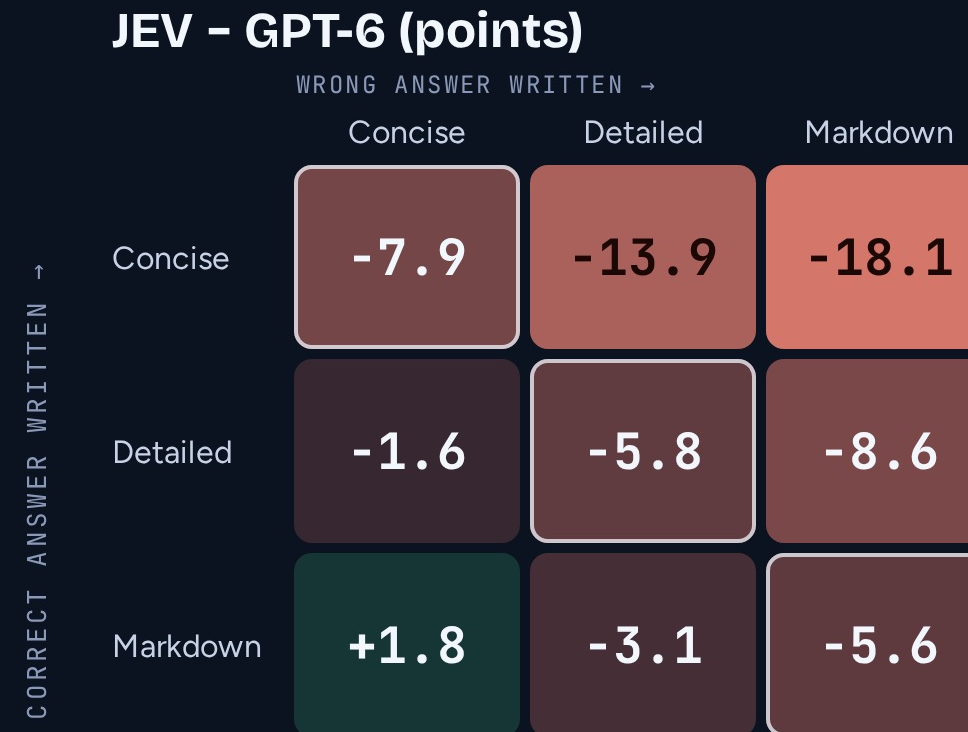
JEV picks the **wrong** answer, **0.90**. GPT-6: correct, 0.999.

A polished wrong answer **misleads** JEV: -13.5 points on hard pairs.



Left: accuracy by style condition, 500 RM-Bench prompts in both orders; JEV falls from 85.4% (normal) to 76.6% (hard), GPT-6 from 91.8% to 90.1%. Right: JEV - GPT-6 by domain. On hard pairs JEV trails by 9.1 points on chat, 9.6 on requests it should answer, 21.8 on code and 23.4 on math; only refusals hold (-0.3).

Length alone does not hurt JEV; a **more elaborate wrong answer** does.



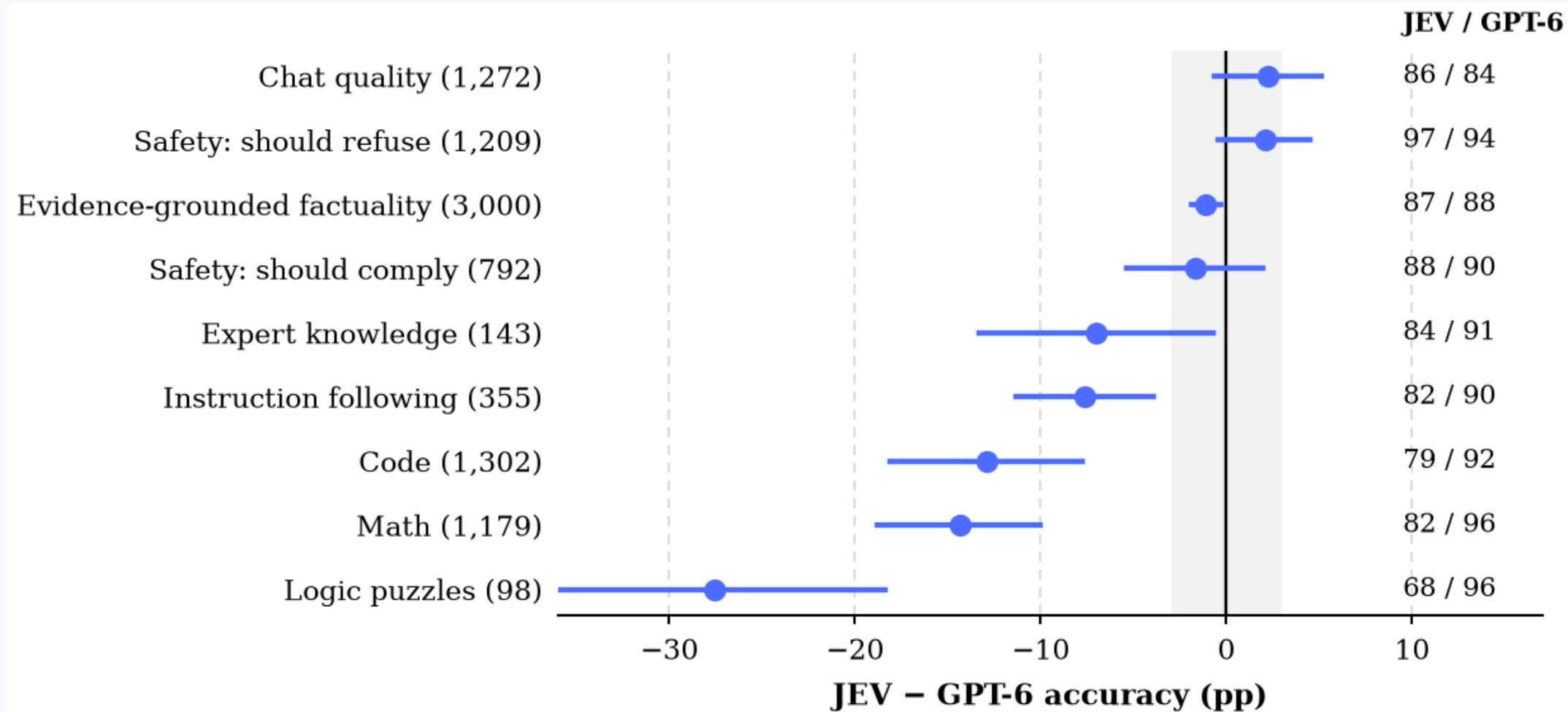
Diagonal, same style: JEV stays at 85.0 → 84.8 → 86.4 as both answers get longer; the gap to GPT-6 even narrows.

Right of the diagonal: the more elaborate the wrong answer, the lower JEV goes, down to 72.1 (-18.1).

Left of the diagonal: when the correct answer is the elaborate one, JEV catches up and even leads (+1.8).

Rows: style of the correct answer; columns: style of the wrong answer. Each cell is 1,000 judgments (500 RM-Bench prompts, both orders). GPT-6 stays between 89.6% and 92.9% in every cell. Accuracy shading runs from 70% to 95%.

Conversational and grounded: **on par**. Reasoning-heavy: **falls behind**.



JEV keeps pace on conversational and grounded judgments (chat, safety, factuality) and falls behind on judgments that need reasoning: code, math and logic puzzles.

JEV minus GPT-6 by skill, pooled over every benchmark that tests it (RewardBench, JudgeBench, HaluEval and RM-Bench; 9,350 base-order items), 95% source-cluster intervals. Within about two points on chat, safety and evidence-grounded factuality; -7.0 expert knowledge, -12.9 code, -14.3 math, -27.6 logic puzzles. Code and math include RM-Bench's style pairs.

Intro

JEV

Setup

Results

Analysis

Confidence

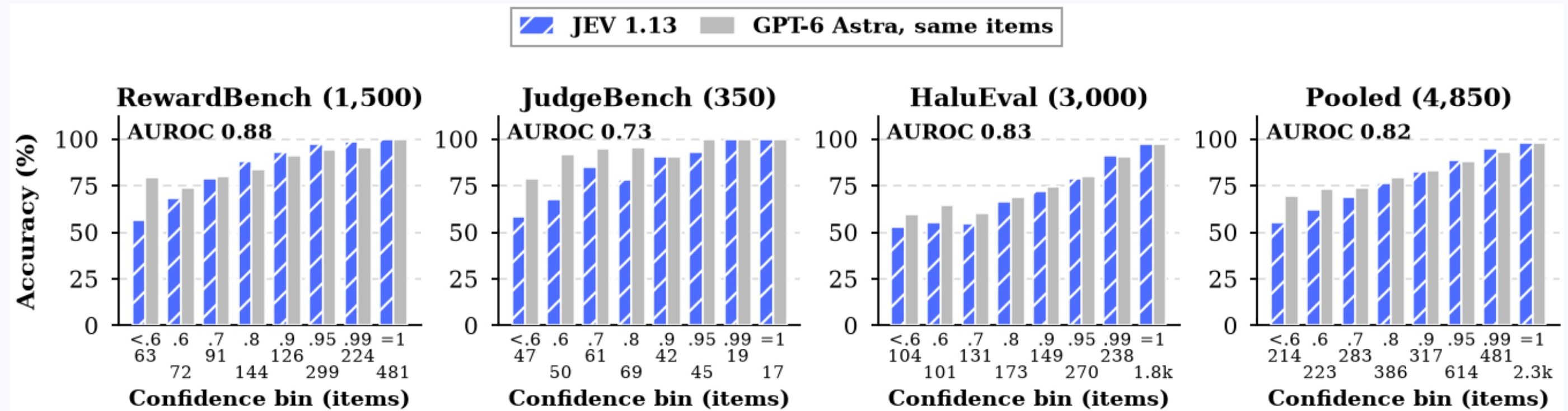
Cascade

Limitations

Wrap-up

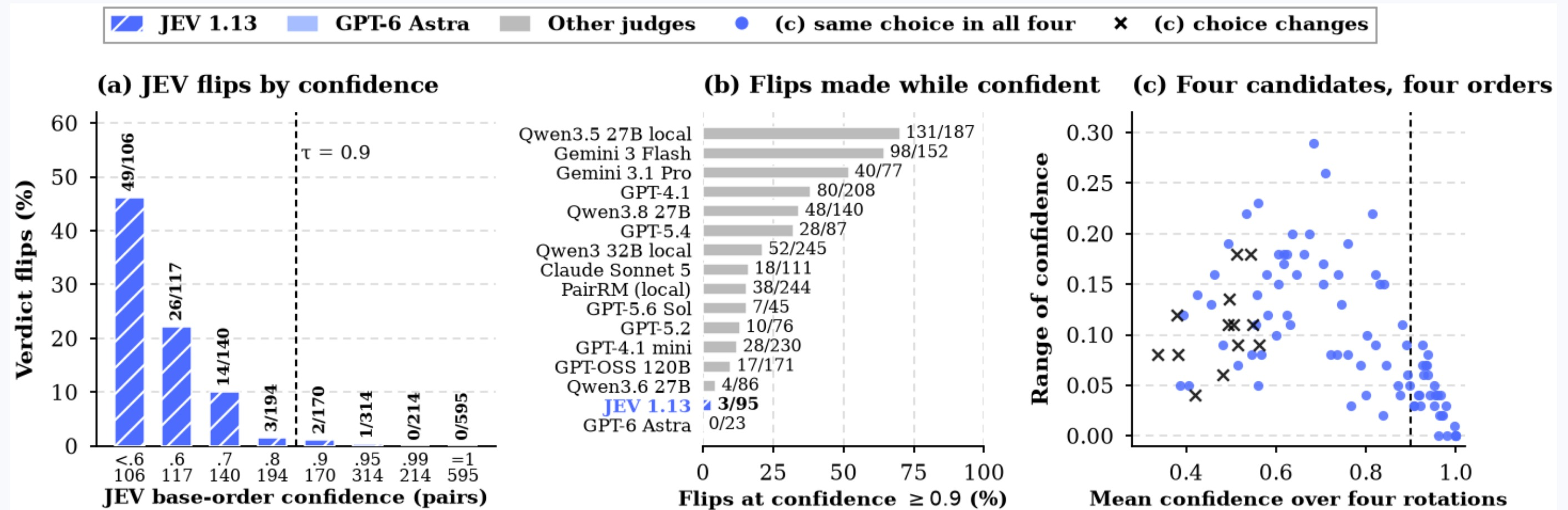
JEV is cheap but not always right. Does its confidence know
when it is wrong?

Where JEV is sure, the two judges **tie**.



At confidence ≥ 0.9 (3,744 of 4,850 items): JEV 94.7% vs GPT-6 94.4%. Below 0.9 (1,106 items): GPT-6 leads, 75.0% vs 67.7%. Error AUROC: 0.88 RewardBench, 0.73 JudgeBench, 0.83 HaluEval, i.e. the chance that a random correct verdict gets higher confidence than a random wrong one (0.5 = no signal, 1 = perfect).

Order flips happen almost only at **low confidence**.



(a, b) 1,850 preference pairs, each judged in both orders; (c) 100 RewardBench 2 prompts, four candidates rotated through four positions. Reversing a pair changes JEV's verdict on 3.7% of RewardBench and 11.1% of JudgeBench pairs; only 3 of its 95 flips occur at confidence ≥ 0.9 . Averaging both orders adds +0.76 points [0.00, 1.57], not significant; per-judge bootstrap in backup.

Intro

JEV

Setup

Results

Analysis

Confidence

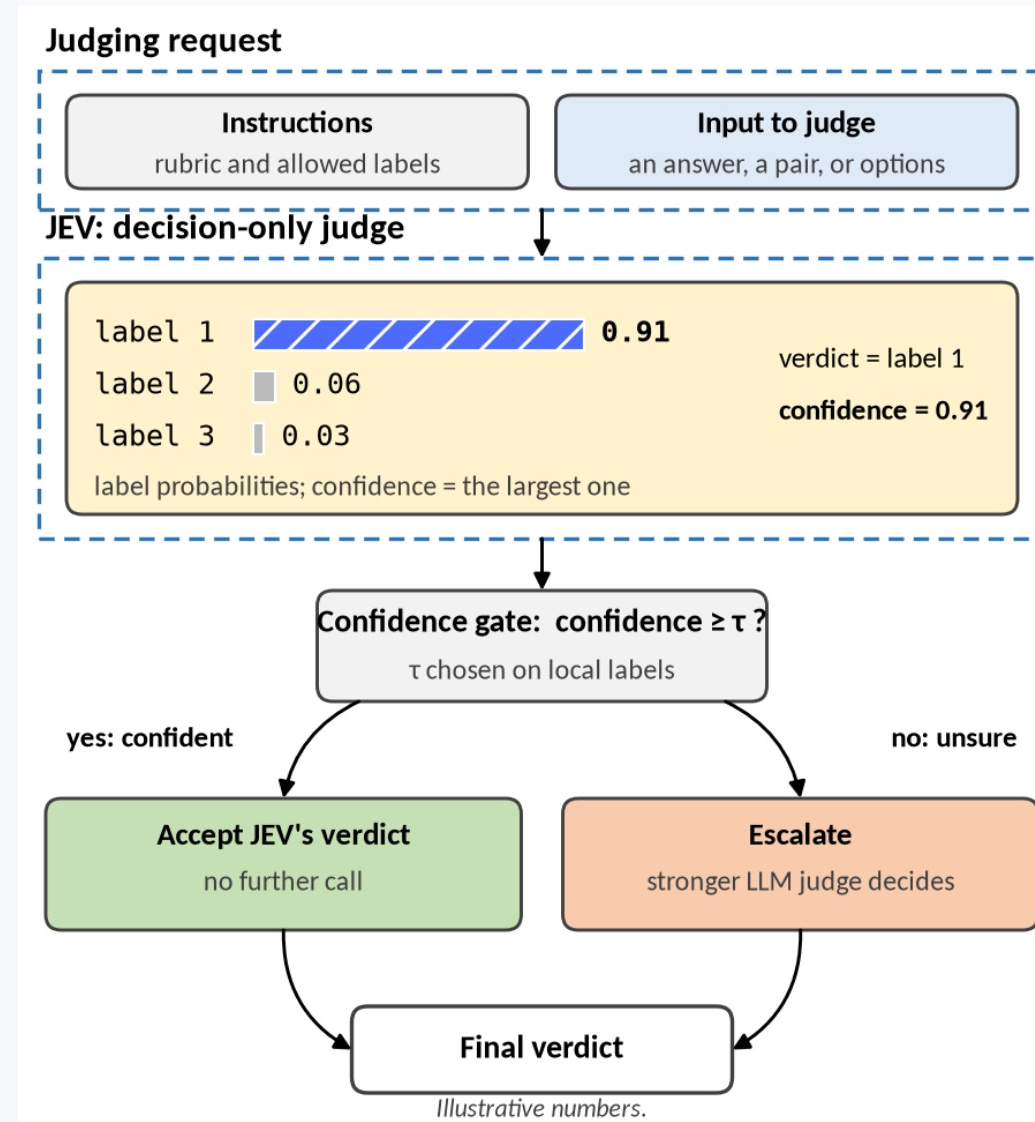
Cascade

Limitations

Wrap-up

Model routing sends each query to the cheapest model that can handle it. We bring the same idea to evaluation, with the judge's confidence as the routing signal.

Accept when confident, **escalate** when unsure.



Serial escalation: the stronger judge decides independently, without seeing JEV's output, and is paid only for escalated items. A preference pair is judged in both orders and the aligned probabilities averaged before gating.

— CASCADE · FOUR EXAMPLES AT $\tau = 0.9$

Where the gate helps, and the error it **cannot** catch.

PAIR	JEV, ONE ORDER	JEV, BOTH ORDERS	GATE	FINAL VERDICT
RewardBench · antonym of “cogent” easy, verdict in the text	right, 0.99	right, 0.99	accept · JEV only	✓ right at JEV’s price
JudgeBench · minimum right shifts one solution returns n instead of 0	wrong, 0.53	right, 0.58	escalate → GPT-6	✓ right, GPT-6 0.999
RM-Bench · dressed-up wrong answer the style example	wrong, 0.90	wrong, 0.64	escalate → GPT-6	✓ right, GPT-6 0.999
RewardBench · “add a line to the poem” six elaborate lines vs one	wrong, 0.97	wrong, 0.99	accept · JEV only	✗ confident error stays

Averaging both orders turns an order-dependent confident error into an escalation. What a cascade cannot fix are errors JEV makes **confidently in both orders**: on held-out pairs at $\tau = 0.9$, the accepted 68.5% hold 15% of JEV’s errors.

Choose τ on your own data: the **cheapest** gate that still matches GPT-6.

STEP 1

Label a small sample

About **100 items** from your own workload, with known answers.

Run JEV and GPT-6 on every item.



STEP 2

Replay each τ

For $\tau = 0.5, 0.6, \dots, 0.99$: keep JEV's verdict if its confidence $\geq \tau$, else use GPT-6's.

No new calls: both verdicts are already in hand.



STEP 3

Score each τ

Accuracy gap to GPT-6 alone, $\Delta(\tau)$, and the fee it would cost.

Higher τ = more escalated = closer to GPT-6, but pricier.



STEP 4

Pick the cheapest τ

Lowest fee whose **95% lower bound** on Δ is above -2 points.

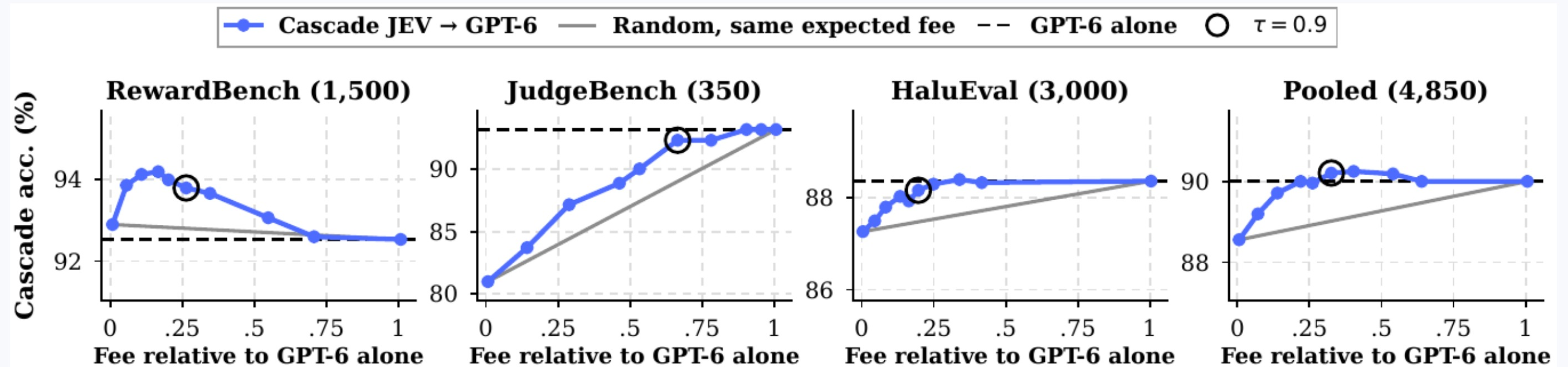
Bound, not mean: a hundred labels are noisy.

45% → 5%

Chance of ending up more than 2 points below GPT-6 when choosing on ~100 labels: by the mean → by the lower bound (under 1% with a separate τ per workload).

2,000 random draws of labeled sets from the 1,850 preference pairs, each policy scored on the rest. The rule minimizes cost under an accuracy tolerance; a worked example on PPE is in backup.

At $\tau = 0.9$: GPT-6's accuracy at **a third** of its fee.



The same target needs a very different gate on each dataset: matching GPT-6 within half a point takes no escalation on RewardBench, $\tau = 0.8$ (11% escalated) on HaluEval, and $\tau = 0.99$ (90% escalated) on JudgeBench.

Each curve sweeps τ (post hoc, all 4,850 items). At $\tau = 0.9$, 23% of items are escalated and the cascade scores 90.2% vs GPT-6's 90.0% at 33% of its fee; escalating the same share at random reaches only 89.0%.

On held-out pairs, the cascade beats GPT-6 at 41% of its fee.

ACCURACY

+0.93

points over GPT-6 alone (93.4% vs 92.5%), 95% CI [0.24, 1.66]

PRICE

41%

of GPT-6's fee

LATENCY

0.27 s

median per pair, vs 2.10 s for GPT-6 (live run, 510 pairs)

HELD-OUT PAIRS	JEV ALONE	CASCADE	GPT-6	CASCADE - GPT-6 (PP)	ESCALATED	FEE VS GPT-6
RewardBench (1,340)	92.8	93.7	92.4	+1.27 [0.45, 2.03]	25%	27%
JudgeBench (270)	81.3	92.2	93.0	-0.74 [-2.22, 0.74]	65%	67%
All held-out pairs (1,610)	90.9	93.4	92.5	+0.93 [0.24, 1.66]	31.5%	41%

$\tau = 0.9$ fixed on 96 pilot pairs before any held-out outcome was seen, using the mean (the lower-bound rule came later and was used on the new workloads). The escalated 31.5% of pairs hold 85% of JEV's errors. The live rerun of 510 pairs matched the simulation (92.5% vs 93.1%, 57% of the fee).

On two workloads it had never seen, the cascade matched GPT-6 exactly.

ACCURACY

90.2 = 90.2

cascade vs GPT-6, where JEV alone scores 76.1%

PRICE

75.5%

of GPT-6's fee: JEV is weak here, so 74% of pairs escalate

LATENCY

1.87 s

median per pair, vs 1.91 s: little to gain when most pairs escalate

NEW WORKLOAD	JEV ALONE	CASCADE	GPT-6	ESCALATED	FEE VS GPT-6	MEDIAN LATENCY (CASCADE / GPT-6)
PPE correctness (400)	78.2	88.2	88.2	68%	69%	1.76 / 1.82 s
JudgeBench, Claude split (170)	70.9	94.7	94.7	89%	91%	2.41 / 2.35 s
Both new workloads (570)	76.1	90.2	90.2	74%	75.5%	1.87 / 1.91 s

Protocol frozen before any model call. Thresholds set per workload by the lower-bound rule on 100 selection pairs: $\tau = 0.95$ (PPE), 0.99 (JudgeBench-Claude). When the first stage is weak, the procedure spends more rather than losing accuracy. Run cost: \$12.77.



What this study does not show.

LIMITATIONS

What this study does **not** show.

Task format

Judgment questions with given candidates: pairwise preference, grounded factuality, final answers, four-way selection. Long free text without a reference is barely covered, and there every judge is near chance.

Agentic tasks

We study JEV as a judge of outputs; using it as a decision maker inside an agent is unexplored here, and is our ongoing work.

Black box

One proprietary version, jev-1.13.0, through its API. Its size, training recipe and data are not public, so we cannot say why it succeeds or fails, or whether benchmark items were seen in training.

Escalation design

We tested the simplest design: serial and independent. Fused panels, a second judge that sees the first verdict, and thresholds set from error costs are open; errors both judges share cannot be fixed by any serial cascade.

Intro

JEV

Setup

Results

Analysis

Confidence

Cascade

Limitations

Wrap-up

So, can JEV replace an LLM as a judge?

Yes, with JEV as a first pass and a stronger LLM when unsure: GPT-6's accuracy, **faster and cheaper**. On its own, only on conversational and grounded judgments.

WHERE JEV CAN REPLACE AN LLM JUDGE

- ▶ **JEV as a first pass, a stronger LLM (GPT-6) when unsure:** +0.93 points on held-out pairs at 41% of the fee and 0.27 s per pair; exact parity on two new workloads.
- ▶ **On its own**, on conversational and grounded judgments (chat, safety, factuality): within 3 points, at 0.36% of the fee and 0.15 s.

BOUNDARIES

- ▶ **(a) Reasoning-heavy judgments** (math, code, logic puzzles): 7–28 points behind on its own. The cascade still matches GPT-6 there, but escalates most items, so savings shrink (75.5% of the fee, no latency gain).
- ▶ **(b) Polished wrong answers** mislead it: –13.5 points on style-adversarial pairs.

An inexpensive judge does not have to be right everywhere. It has to **know where it is not.**

- Decision-only judging is **cheap and fast**: \$44 per million judgments and 0.15 s, against \$12.2k and 1.9 s for GPT-6. [p. 15](#)
- On its own, JEV **matches GPT-6 on conversational and grounded judgments**, within two **boundaries**: (a) reasoning-heavy judgments and (b) polished wrong answers. [p. 36](#)
- Its **confidence is a usable signal**: it ranks JEV's own errors (error AUROC 0.73–0.88), and JEV is rarely confidently wrong. [p. 24](#)
- **Accept when confident, escalate when unsure**: +0.93 points over GPT-6 at 41% of its fee and 0.27 s on held-out pairs; exact parity on two workloads it had never seen. [p. 31](#), [p. 32](#)
- The **gate depends on the task and data**: the right threshold differed across workloads (0.9 on held-out pairs; 0.95 and 0.99 on the new ones), so set it on local labels and check it. [p. 29](#), [p. 32](#)

Where to read more about JEV and its **open challengers**.

RESOURCE 1 · OFFICIAL

TypeSafe documentation

docs.typesafe.ai

The vendor's docs for JEV: the typed primitives (Choice, Score, Noul), the API, and how confidence is reported.

RESOURCE 2 · OUTSIDE ANALYSIS

Archer Hume, "Jev's Architecture Unmasked"

archerhume.com/posts/jevs-architecture-unmasked

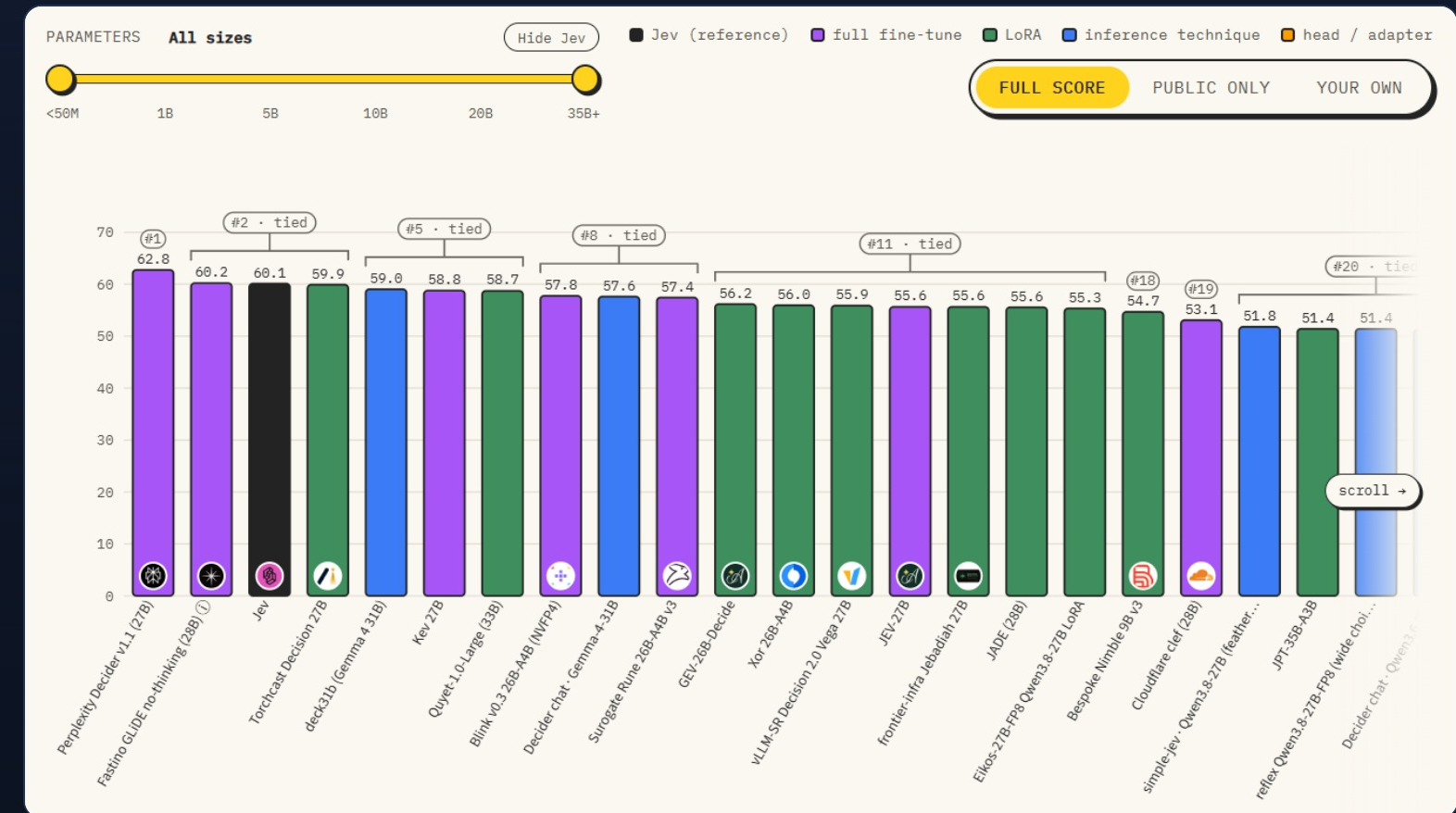
Reconstructs JEV from about 10,000 API probes: a decision-tuned transformer that reads the state once and returns option probabilities instead of generating text.

RESOURCE 3 · LEADERBOARD

Decision Index (JEV and open decision models)

huggingface.co/spaces/multimodalart/jev-decision-index

A public leaderboard that scores open decision models against JEV on 37 public benchmarks plus private tests; the top open model now edges past JEV.



Decision Index 0.3, Full score, as of 7 Oct 2026

— WRAP-UP · EXTRA: OPEN DECISION JUDGES ON OUR DATA

#	JUDGE	DI RANK	CONTEXT L	PUBLIC POOLED ▼	REWARDBENCH	JUDGEBENCH	HALUEVAL	Δ VS JEV	P50 S
1	Perplexity Decider v1.1 (27B)	1	8k	89.0	92.6	82.6	88.0	+0.8	0.08
2	Eikos-27B (FP8)	17	16k	89.0	91.9	78.6	88.8	+0.8	0.06
3	simple-jev (Qwen3.8-27B)	20	8k	88.3	93.7	76.6	87.0	+0.0	0.17
4	JEV (hosted)	3	32k	88.3	92.5	78.6	87.3	—	0.15
5	vLLM-SR Decision 2.0 Vega 27B	13	32k	88.0	92.7	78.6	86.8	-0.2	0.14
6	Decider chat (Gemma-4-31B)	9	32k ^t	87.7	91.6	81.7	86.5	-0.6	0.06
7	Torchcast Decision 27B	4	256k	87.6	90.7	80.0	86.9	-0.7	0.07
8	Jebadiah 27B	15	8k	87.1	92.6	76.0	85.7	-1.1	0.12
9	deck31b (Gemma 4 31B)	5	16k	87.0	91.2	81.7	85.5	-1.3	0.90
10	Quyet-1.0-Large (31B)	7	8,000 ^t	86.8	91.5	79.4	85.3	-1.5	0.13
11	AutoJev-27B	—	8k	86.6	92.8	77.4	84.6	-1.6	0.08
12	Kev 27B	6	64k	86.5	91.7	76.3	85.1	-1.7	0.08
13	JEV-27B (AutoTrust)	14	256k	85.2	91.3	75.1	83.3	-3.1	0.13
14	Surogate Rune 26B-A4B v3	10	32k	83.7	91.1	78.6	80.6	-4.6	0.18
15	Cygnet (Gemma-4-12B)	—	16k	83.7	90.2	65.4	82.6	-4.6	0.03
16	reflex-27B (stock Qwen3.8)	22	4k	83.5	90.5	69.7	81.6	-4.8	0.24
17	Xor 26B-A4B	12	128k	83.3	89.4	76.0	81.2	-4.9	0.04
18	Cloudflare Clef-27B	19	16k ^t	82.7	88.5	76.6	80.6	-5.5	0.09
19	Blink v0.3 26B-A4B (NVFP4)	8	32k	82.4	88.3	75.7	80.2	-5.9	0.04
20	GEV-26B-Decide	11	256k	80.6	90.8	73.4	76.3	-7.7	0.11
21	CLM-8B	104	2k ^t	68.5	51.7	52.9	78.8	-19.7	0.02
22	Laya	106	512 ^t	60.8	47.3	48.9	68.9	-27.5	0.03

Same 5,172 judgments and prompts as JEV; open judges served with their authors’ code and calibration, one request at a time on one H100. Δ = public-pooled accuracy minus JEV’s; green / red = 95% CI above / below zero (top 15: 0 above, 4 tied, 9 below). Ranked by public-pooled accuracy (click a column header to rank by it); DI rank in amber = Decision Index 0.3 top 15. Context L = longest single-question input accepted as shipped (tokens; ^t truncates instead of refusing). p50 = median latency on the paper’s 120-judgment panel (JEV hosted; others local). Bold = best.

A breakthrough is not a revolution until it becomes **cheap enough to be used everywhere.**

Breakthrough

the technology works



Cost falls

cheap enough to use widely



Revolution

rebuilt around the cheap input

Steam, steel, oil, microelectronics: each revolution came when its key input became cheap.

LLMs gave us the breakthrough in intelligence. **JEV makes it cheap**: \$44 per million judgments, 0.15 s, and with confidence to know when to hand over. That brings AI's revolution **a step closer**.



Carlota Perez
Technological Revolutions and Financial Capital (2002)

After Carlota Perez (1983; 2002): technological revolutions form around a key input whose cost is clearly low and falling. The mapping to JEV is ours. Photo: Ângela Campos, CC BY-SA 4.0, Wikimedia Commons.

Thank you!

Questions are welcome.

PRESENTER Yubo Li

yubol@andrew.cmu.edu

— BACKUP • DOES JEV LEAD ANYWHERE?

No skill where JEV leads outright. Its advantage is **cost and latency, not accuracy.**

Clearly ahead of GPT-6 only on a few easy slices: RewardBench chat +6.9 [3.2, 10.6], RM-Bench easy chat +10.1 [5.7, 15.1], refusals +4.4 [0.6, 8.4]. Among judges under \$1 per 1,000 judgments, JEV is the most accurate on all three public benchmarks.

Where an inexpensive judge is enough, and where it is not.

WORKLOAD	SAMPLE	JEV	COMPARATOR	Δ (PP) [95% CI]	GUIDANCE
Ordinary preference	RewardBench (1,500)	92.5	GPT-6, 92.5	0.0 [−1.4, 1.4]	Low-fee candidate
Evidence-grounded factuality	HaluEval QA (3,000)	87.3	GPT-6, 88.4	−1.1 [−1.9, −0.3]	Low-fee candidate
Final-answer adjudication	Final answer (150)	94.0	GPT-6, 96.7	−2.7 [−6.5, 0.7]	Validate first
Four-way selection	RewardBench 2 (100)	73.0	GPT-6, 75.0	−2.0 [−12.0, 8.0]	Validate first
Grounded summaries	HaluEval summaries (400)	69.8	GPT-5.4, 69.3	+0.5 [−3.0, 4.0]	Validate first
Difficult correctness	JudgeBench (350)	78.6	GPT-6, 93.1	−14.6 [−18.9, −10.3]	Escalate
Matched-style pairs	RM-Bench normal (3,000)	85.4	GPT-6, 91.8	−6.4 [−9.1, −4.0]	Escalate
Style-adversarial pairs	RM-Bench hard (3,000)	76.6	GPT-6, 90.1	−13.5 [−16.2, −10.9]	Escalate
Reference-free prose	HaluEval general (200)	53.5	GPT-5.4, 56.0	−2.5 [−7.0, 2.0]	Not supported

Accuracy (%) and paired JEV-minus-comparator differences with 95% source-cluster intervals. “Not supported”: every judge tested is near chance. Guidance describes these samples, not a deployment guarantee.

Confidence tracks errors where JEV is unsure, but **not where it is confidently misled.**

0.498

error AUROC on reference-free prose:
confidence is no better than chance

1 Style-adversarial pairs

Error AUROC falls to 0.76, from 0.84 on style-matched pairs.

2 Reference-free prose

All three judges tested are near chance yet confident (mean max probability 0.91–0.96).

3 Calibration does not transfer

Temperature scaling helps HaluEval (NLL 0.496 → 0.304) but hurts JudgeBench and RewardBench.

JEV's three output types also disagree on the same binary question (Choice vs Noul gap 0.055), so confidence should be read from one fixed output type.

Base accuracy, fees and latency for all seventeen configurations.

Judge	RewardBench (1,500)	JudgeBench (350)	HaluEval (3,000)	Final answer (150)	Valid base responses	Fee (\$/1k)	Median latency (s)
JEV 1.13	92.5	78.6	87.3	94.0	5172/5172	0.044	0.15
GPT-4.1 mini [†]	89.1	64.0	86.1	84.0	5172/5172	0.390	0.55
GPT-4.1 [†]	90.9	71.7	86.5	92.0	5172/5172	1.947	0.58
GPT-5.2 [†]	92.2	88.9	89.5	95.3	5172/5172	4.193	1.68
GPT-5.4	92.5	90.9	88.9	92.7	5170/5172	4.808	1.44
GPT-5.6 Sol	92.9	93.1	89.9	95.3	5171/5172	6.087	1.74
GPT-6 Astra	92.5	93.1	88.4	96.7	5171/5172	12.182	1.89
GPT-OSS 120B	88.9	71.1	87.2	95.3	5094/5172	0.225–0.433	0.55
Qwen3 32B local	87.5	66.9	84.2	94.0	5171/5172	—	—
Qwen3.5 27B local	90.1	77.4	84.7	92.7	5172/5172	—	—
Qwen3.6 27B	87.5	72.3	86.4	93.3	4998/5172	3.165–8.469	1.96
Qwen3.8 27B	92.8	74.3	87.8	95.3	5159/5172	3.021–5.018	1.52
Claude Sonnet 5	91.1	88.9	87.6	94.7	5164/5172	5.890	2.28
Gemini 3 Flash	92.1	76.0	83.7	95.3	5172/5172	0.508	0.87
Gemini 3.1 Pro	94.3	87.4	87.4	94.7	5171/5172	5.043	3.14
PairRM (local) [†]	65.5	54.3	—	—	1850/1850	—	—
Skywork V2 8B (local) [‡]	94.0	71.1	—	—	1850/1850	—	—

Invalid outcomes count as errors. Fees and median latency from the 120-judgment timing panel. † initial-round judge; ‡ follow-up addition.

A frozen pilot fits thresholds; the held-out set never enters a fit.

TASK	PILOT	SELECT	TEST	HELD-OUT
RewardBench	160	64	96	1,340
JudgeBench	80	32	48	270
HaluEval	80	32	48	2,920
Final answer, controls	322	—	—	0
All base judgments	642	—	—	4,530

Judgments. Pilot public tasks split 40/60 by source question into selection and test; held-out shares no source question with the pilot. Each preference pair is shown once in a seeded random order (base) and again reversed (diagnostic).

Adjudication favors GPT-6 more than the labels do.

Task	n	κ	S3 human = label	S2 label / judges / indec.	S1 JEV / GPT-6 / indec.	Label Δ	Human Δ [95% CI]	Corrected JEV / GPT-6
RewardBench	50	0.29	7/7	4 / 5 / 5	5 / 17 / 7	-1.3	-3.0 [-5.3, -0.8]	93.0 / 95.2
JudgeBench	91	0.74	5/7	4 / 1 / 10	1 / 57 / 11	-14.6	-16.0 [-20.0, -12.3]	78.6 / 93.7
HaluEval	42	0.81	6/6	1 / 24 / 1	2 / 8 / 0	+0.8	-2.5 [-5.0, -0.4]	95.8 / 98.3

S1: exactly one judge matches the label; S2: neither; S3: both (controls). Δ : JEV minus GPT-6 over all adjudication-subset items of the task, on labels and under human labels.

Averaging both orders: no significant gain for JEV.

Judge	Accuracy (%), all 1,850 pairs		Gain from averaging (points)		
	One order	Both orders	All 1,850	Held-out 1,610	Held-out JB 270
JEV 1.13	89.9 ± 0.7	90.6 ± 0.7	+0.76 ± 0.40	+0.62 ± 0.43	+2.04 ± 1.63
GPT-4.1 mini [†]	84.3 ± 0.9	84.9 ± 0.8	+0.59 ± 0.49	+0.43 ± 0.52	+3.70 ± 2.04
GPT-4.1 [†]	87.2 ± 0.8	87.3 ± 0.7	+0.03 ± 0.49	−0.12 ± 0.55	−1.30 ± 2.01
GPT-5.2 [†]	91.6 ± 0.7	92.2 ± 0.6	+0.65 ± 0.30*	+0.68 ± 0.34*	+1.48 ± 1.05
GPT-5.4	92.2 ± 0.6	92.8 ± 0.6	+0.59 ± 0.37	+0.71 ± 0.41	+0.56 ± 0.84
GPT-5.6 Sol	93.0 ± 0.6	93.5 ± 0.6	+0.51 ± 0.27*	+0.53 ± 0.29	+0.37 ± 0.65
GPT-6 Astra	92.6 ± 0.6	92.9 ± 0.6	+0.22 ± 0.20	+0.25 ± 0.22	−0.37 ± 0.35
GPT-OSS 120B	85.6 ± 0.8	85.2 ± 0.8	−0.38 ± 0.54	−0.37 ± 0.57	+2.41 ± 2.03
Qwen3 32B local	83.6 ± 0.9	83.8 ± 0.8	+0.19 ± 0.48	+0.03 ± 0.52	+3.52 ± 1.99
Qwen3.5 27B local	87.7 ± 0.8	87.9 ± 0.7	+0.22 ± 0.45	+0.25 ± 0.46	+2.41 ± 1.60
Qwen3.6 27B	84.6 ± 0.8	81.4 ± 0.9	−3.24 ± 0.55**	−2.76 ± 0.58**	−3.70 ± 1.79
Qwen3.8 27B	89.3 ± 0.7	89.2 ± 0.7	−0.14 ± 0.42	−0.34 ± 0.44	−1.67 ± 1.79
Claude Sonnet 5	90.7 ± 0.7	90.8 ± 0.7	+0.08 ± 0.38	+0.19 ± 0.39	+0.37 ± 1.21
Gemini 3 Flash	89.1 ± 0.7	89.9 ± 0.7	+0.81 ± 0.42*	+0.87 ± 0.46	+6.11 ± 1.88**
Gemini 3.1 Pro	93.0 ± 0.6	92.9 ± 0.6	−0.05 ± 0.30	+0.03 ± 0.32	+1.85 ± 1.41
Laya English (local)	47.6 ± 1.2	47.3 ± 1.1	−0.30 ± 1.25	−0.34 ± 1.33	+2.04 ± 3.58
Laya Typed (local)	49.5 ± 1.2	48.8 ± 1.2	−0.70 ± 1.43	−1.34 ± 1.55	+6.11 ± 3.84
PairRM (local) [†]	63.4 ± 1.1	63.8 ± 1.1	+0.43 ± 0.59	+0.50 ± 0.66	−1.48 ± 1.66

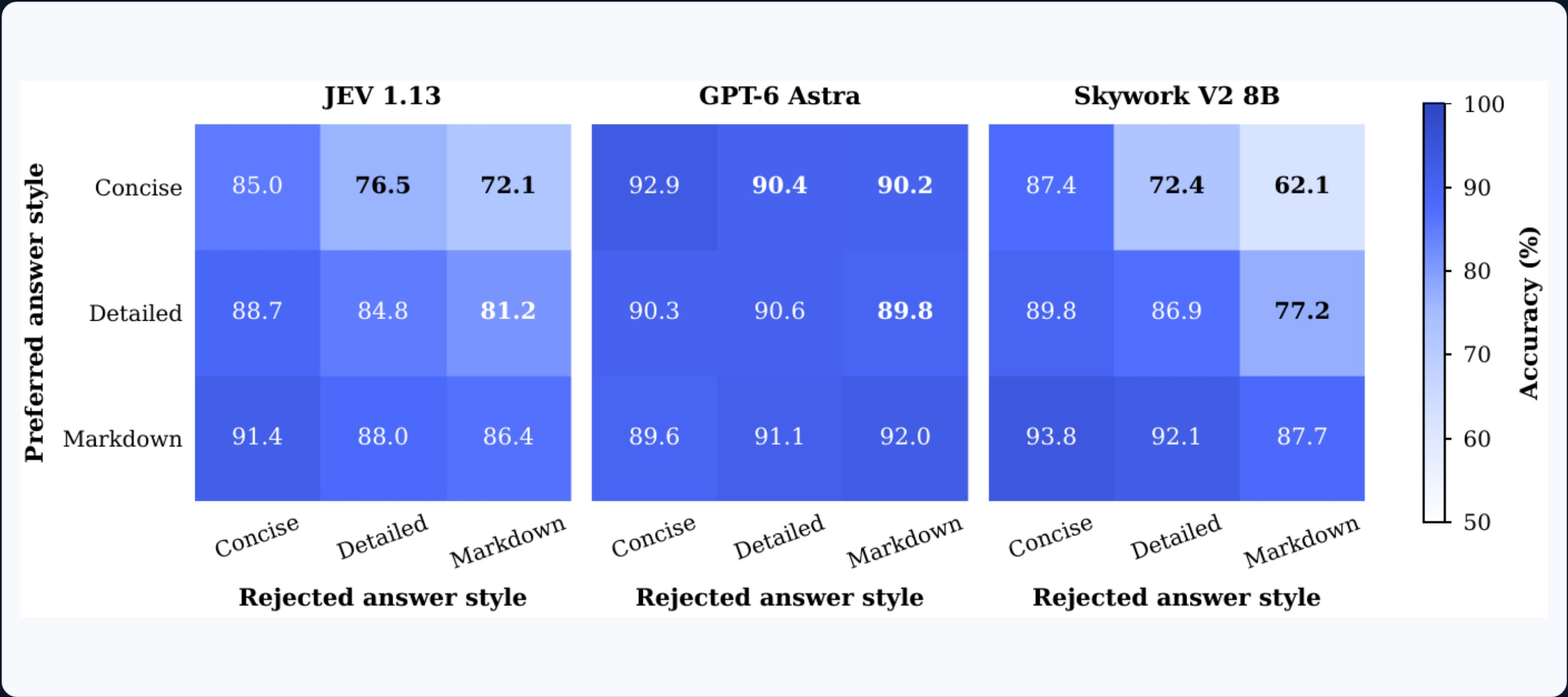
±: SD over 2,000 source-cluster bootstrap resamples. * 95% interval excludes zero; ** also after Holm correction across 18 judges. Qwen3.6's loss is an artifact of invalid outputs.

JEV's confident errors are a smaller share of its errors than most judges!

Judge	RewardBench		JudgeBench		HaluEval	
	Errors / $q \geq 0.9$ (%)	Share of errors (%)	Errors / $q \geq 0.9$ (%)	Share of errors (%)	Errors / $q \geq 0.9$ (%)	Share of errors (%)
JEV 1.13	24/1,155 (2.1)	21	9/138 (6.5)	12	171/2,491 (6.9)	45
GPT-4.1 mini	23/581 (4.0)	14	4/28 (14.3)	3	393/2,892 (13.6)	94
GPT-4.1	72/1,200 (6.0)	53	36/162 (22.2)	36	397/2,981 (13.3)	98
GPT-5.2	33/1,178 (2.8)	28	5/226 (2.2)	13	190/2,671 (7.1)	60
GPT-5.4	42/1,296 (3.2)	38	8/292 (2.7)	25	236/2,765 (8.5)	71
GPT-5.6 Sol	49/1,326 (3.7)	46	9/315 (2.9)	38	248/2,879 (8.6)	82
GPT-6 Astra	46/1,306 (3.5)	41	6/299 (2.0)	25	296/2,883 (10.3)	85
GPT-OSS 120B	18/905 (2.0)	11	11/120 (9.2)	11	312/2,893 (10.8)	81
Qwen3 32B local	62/759 (8.2)	33	23/101 (22.8)	20	458/2,799 (16.4)	97
Qwen3.5 27B local	108/1,384 (7.8)	73	53/274 (19.3)	67	459/3,000 (15.3)	100
Qwen3.6 27B	11/986 (1.1)	6	2/160 (1.2)	2	124/2,427 (5.1)	30
Qwen3.8 27B	20/1,079 (1.9)	19	35/182 (19.2)	39	175/2,501 (7.0)	48
Claude Sonnet 5	19/1,012 (1.9)	14	5/178 (2.8)	13	259/2,815 (9.2)	70
Gemini 3 Flash	83/1,379 (6.0)	70	55/292 (18.8)	65	470/2,972 (15.8)	96
Gemini 3.1 Pro	46/1,404 (3.3)	53	30/297 (10.1)	68	332/2,904 (11.4)	88
Laya English (local)	53/110 (48.2)	7	0/2 (0.0)	0	346/955 (36.2)	37
Laya Typed (local)	0/1 (0.0)	0	0/0 (0.0)	0	153/279 (54.8)	18

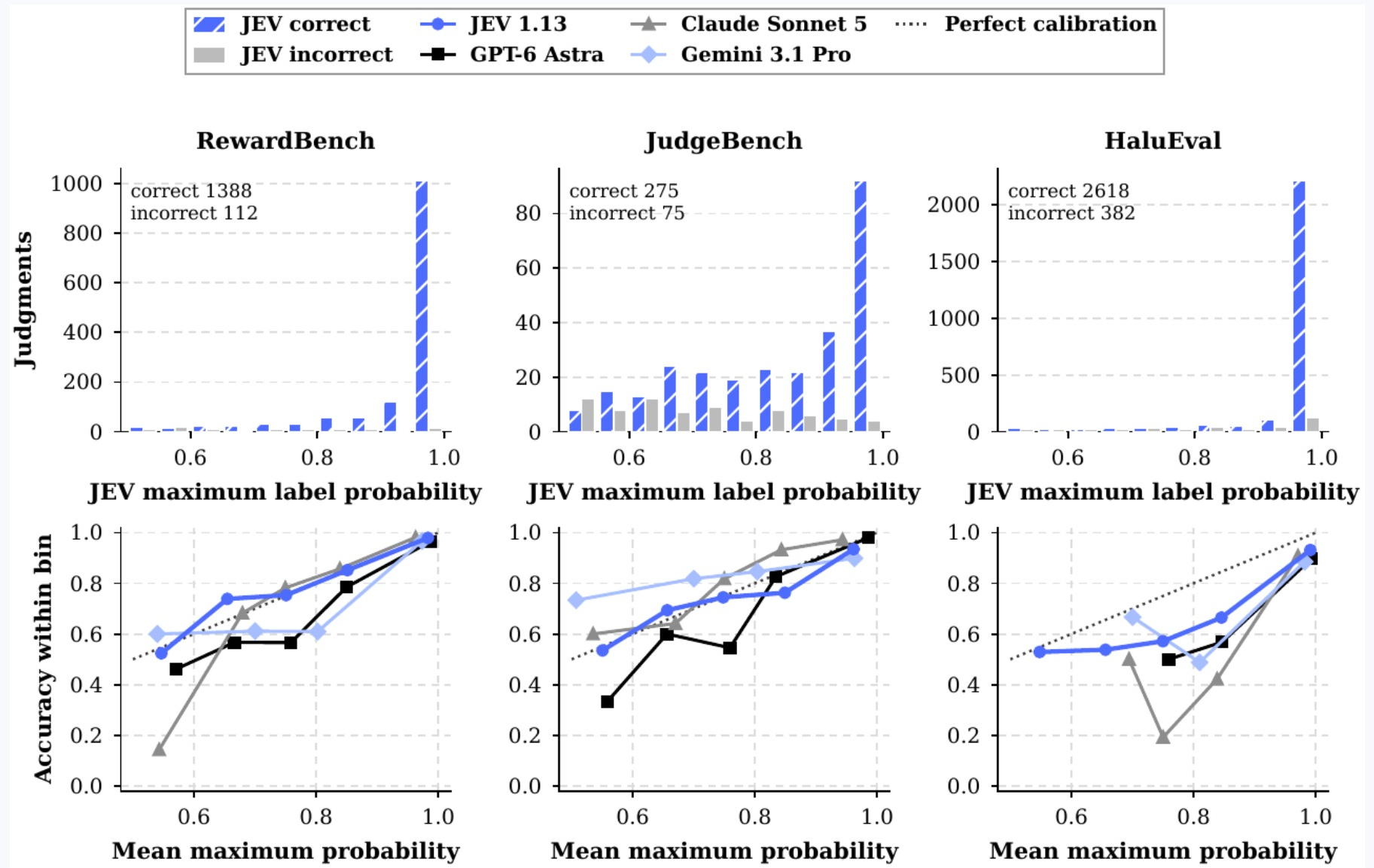
Errors among valid base-order judgments with confidence ≥ 0.9 , over their number; share of each judge's errors. HaluEval shares are dominated by label noise.

The style effect, cell by cell.



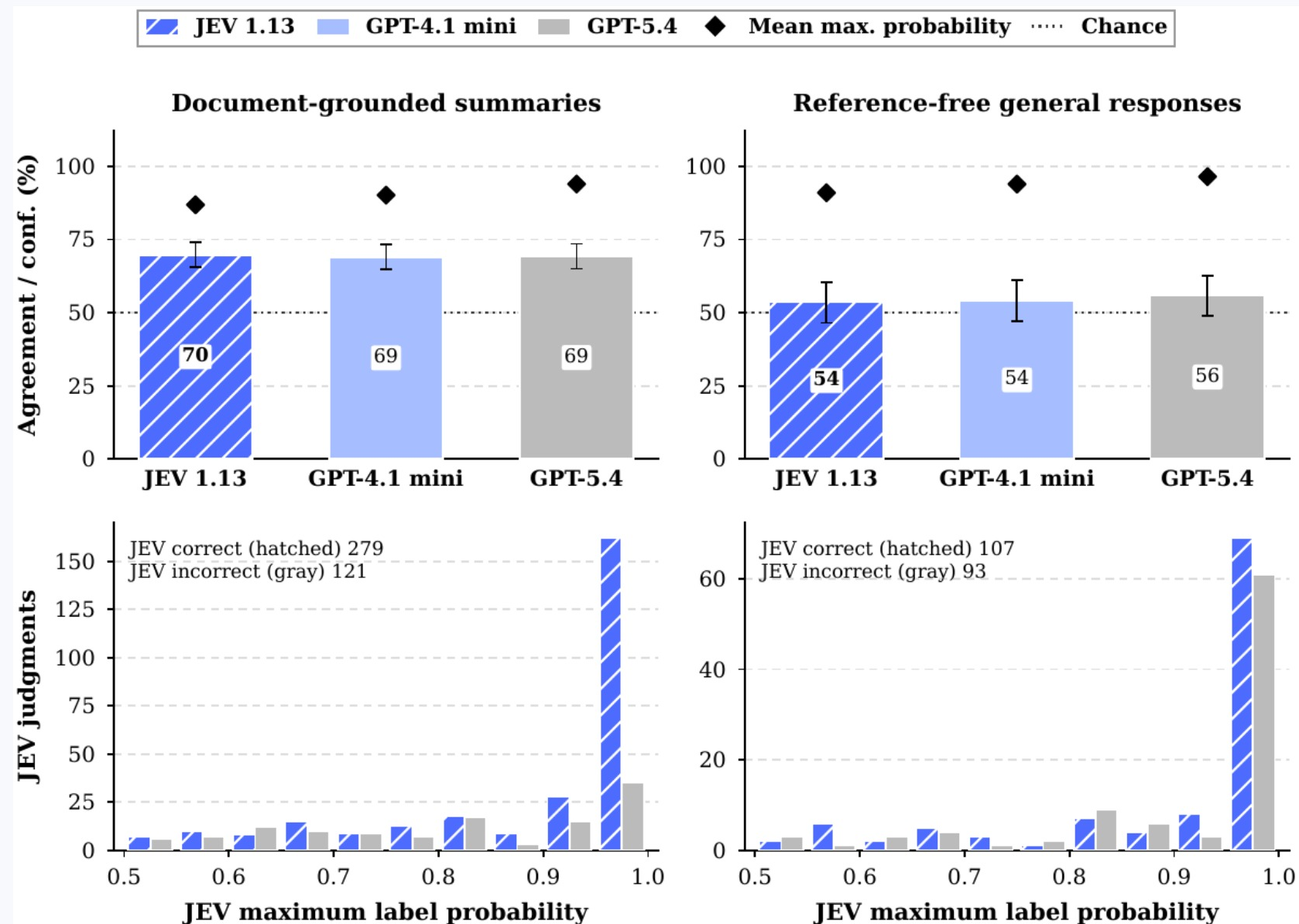
Each cell averages 500 prompts in both orders. Above-diagonal cells contribute to hard accuracy, diagonal to normal, below-diagonal to easy.

Confidence distributions and reliability curves.



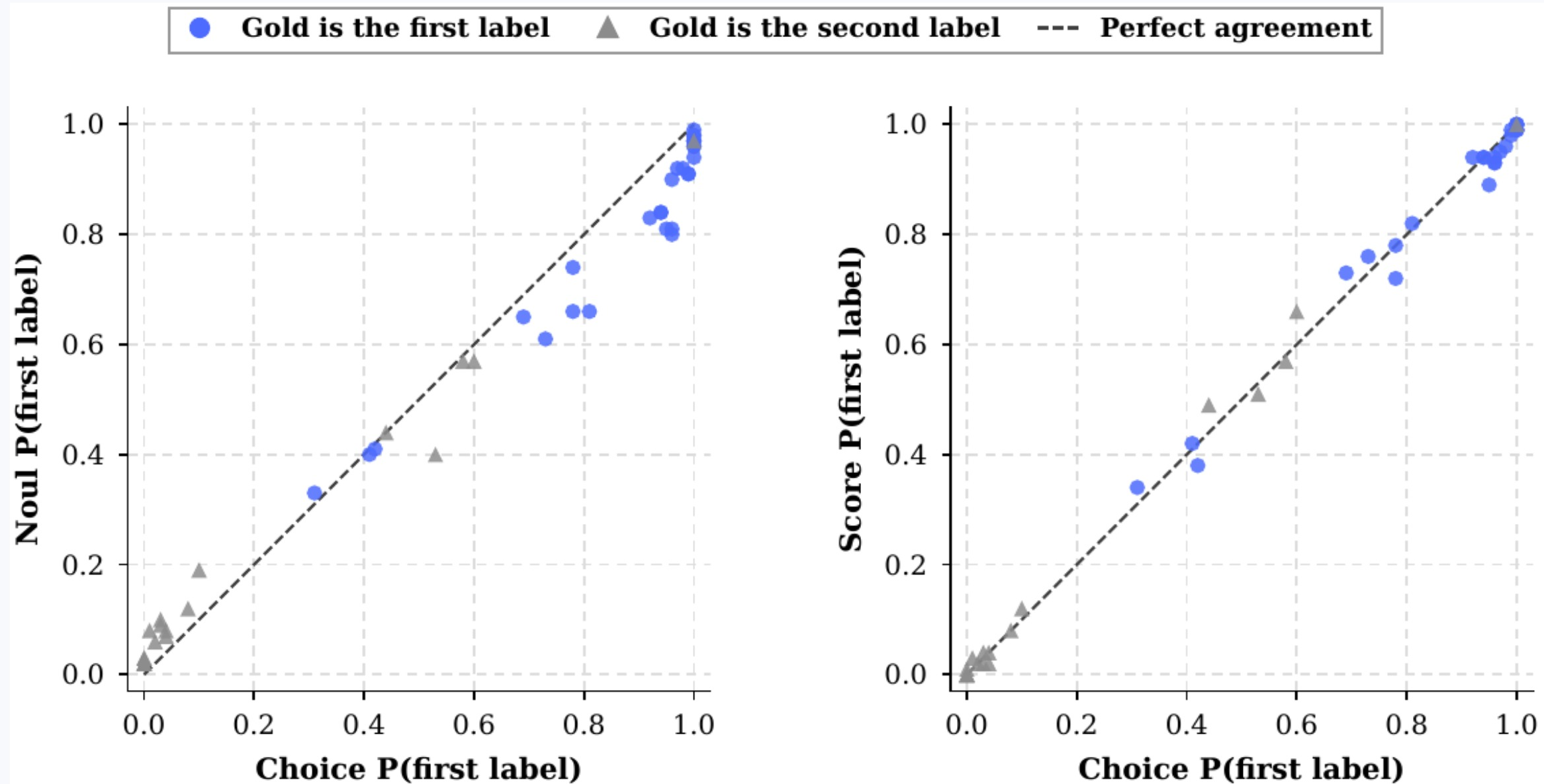
Top: JEV's maximum-probability distributions for correct and incorrect base judgments. Bottom: reliability curves for selected judges. Temperature fits transfer unevenly (table in the paper, Appendix L).

Without a reference, every judge is near chance yet confident.



HaluEval summaries (document-grounded) vs general responses (reference-free, balanced on human labels). The two samples differ in content and label provenance.

Choice, Noul and Score agree only approximately.



48 initial-round examples, all three types asked in one request. Choice vs Noul differ by 0.055 on average, Choice vs Score by 0.015; Noul complements miss summing to one by 0.045.

JEV's native confidence field rescales the top probability; our confidence is the top probability itself.

LABELS		RETURNED PROBABILITIES	TOP PROBABILITY	NATIVE FIELD	$(K \cdot P_{MAX} - 1) / (K - 1)$
Binary		{B 0.07, A 0.93}	0.93	0.87	$2(0.93) - 1 = 0.86$
Binary		{hallucinated 0.67, supported 0.33}	0.67	0.35	$2(0.67) - 1 = 0.34$
Three labels		{incorrect 0.11, no_answer 0.26, correct 0.63}	0.63	0.45	$(3(0.63) - 1) / 2 = 0.445$

Three real JEV responses. Within a fixed number of labels the two are monotone, so gating and AUROC are unchanged; the top probability is on the scale needed for calibration, and every judge returns it. In binary tasks a confidence of 0.9 corresponds to a native field of 0.8.

The point rule would have taken $\tau = 0.9$; the lower bound moved it to 0.95.

T	ACCEPTED %	MEAN Δ	95% LOWER BOUND	POINT RULE	LOWER-BOUND RULE
0.8	54	-4.00	-7.24	no	no
0.9	38	-2.00	-4.31	yes	no
0.95 ← chosen	24	0.00	0.00	yes	yes
0.99	13	0.00	0.00	yes	yes

PPE selection set, 100 pairs; Δ = cascade minus GPT-6 accuracy. Point rule: mean ≥ -2 pp; lower-bound rule: mean - 1.645 s/ $\sqrt{n} \geq -2$ pp. The rule fixes a tolerance, not a price per error (Chow, 1970; Elkan, 2001).

A lower confidence bound cuts the risk of losing more than 2 points.

Selection rule	Labeled pairs k	P(loss > 2 pp) (%)	Median Δ (pp)	10th pct. Δ (pp)	Median escalated (%)	Median fee ratio
Point rule, one τ	48	39.7	−0.72	−2.16	6	0.09
	96	44.0	−0.74	−2.22	6	0.09
	192	48.5	−0.84	−2.29	6	0.09
	384	50.0	−1.13	−2.49	5	0.08
Lower bound, one τ	48	6.6	+0.50	−0.83	32	0.42
	96	5.2	+0.46	−0.91	32	0.42
	192	4.8	+0.48	−1.03	21	0.28
	384	2.9	+0.14	−1.09	13	0.18
Point rule, τ per workload	48	11.8	−0.50	−2.05	12	0.19
	96	11.8	−0.03	−2.11	12	0.26
	192	5.3	+0.09	−1.24	12	0.27
	384	1.6	+0.03	−0.85	12	0.27
Lower bound, τ per workload	48	1.2	+0.11	−1.08	18	0.29
	96	0.6	+0.29	−0.80	20	0.34
	192	0.1	+0.84	−0.21	18	0.34
	384	0.1	+0.82	−0.24	17	0.36

2,000 random draws of k selection pairs from the 1,850 pairs; each policy scored on the rest. The simulation reuses one pool of items, so it is optimistic about distribution shift.

Held-out pairs, full table.

Held-out pairs	JEV alone	Cascade	GPT-6	Δ (pp) [95% CI]	Escalated (%)	JEV errors caught (%)	Fee ratio reported / upper
RewardBench (1,340)	92.8	93.7	92.4	+1.27 [0.45, 2.03]	24.8	82	0.267 / 0.308
JudgeBench (270)	81.3	92.2	93.0	−0.74 [−2.22, 0.74]	64.8	90	0.668 / 0.668
All held-out pairs (1,610)	90.9	93.4	92.5	+0.93 [0.24, 1.66]	31.5	85	0.414 / 0.440

τ = 0.9 chosen on 96 pilot pairs by the point rule before any held-out outcome was inspected.

The live run reproduces the simulation.

		Esc. (%)	Same gate as offline (%)	Cascade	GPT-6	Δ (pp) [95% CI]	Fee ratio	Cascade (s) p50 / p95	GPT-6 (s) p50 / p95
RewardBench	240	25.8	98.8	93.3	93.3	0.00 [−2.09, 2.06]	0.272	0.19 / 3.27	2.01 / 3.69
JudgeBench	270	64.8	98.5	91.9	93.0	−1.11 [−2.96, 0.37]	0.671	2.02 / 5.07	2.16 / 4.80
Both benchmarks	510	46.5	98.6	92.5	93.1	−0.59 [−1.95, 0.59]	0.572	0.27 / 4.26	2.10 / 4.49

Accepted pairs return in 0.18 s, escalated ones in 2.50 s (JEV, then GPT-6). Same accept-or-escalate decision as offline on 98.6% of pairs.

With the same rule, JEV can accept more pairs than any other inexpensive first stage.

First stage	Fallback	τ	RewardBench held-out (1,340)			JudgeBench held-out (270)		
			Accept (%)	Δ (pp) [95% CI]	Fee ratio rep. / upper	Accept (%)	Δ (pp) [95% CI]	Fee ratio rep. / upper
<i>JEV first stage, rule fixed before evaluation</i>								
JEV	GPT-5.4	0.90	75.2	+0.37 [−0.15, 0.90]	0.306 / 0.372	35.2	0.00 [−1.48, 1.48]	0.744 / 1.237
JEV	GPT-5.6 Sol	0.70	90.7	+0.82 [−0.15, 1.83]	0.130 / 0.166	71.5	−4.81 [−8.52, −1.48]	0.336 / 0.399
JEV	GPT-6 Astra	0.90	75.2	+1.27 [0.45, 2.03]	0.267 / 0.308	35.2	−0.74 [−2.22, 0.74]	0.668 / 0.668
<i>Alternative first stages, same rule, GPT-6 fallback (retrospective); §: open weights, run locally</i>								
GPT-4.1 mini	GPT-6 Astra	0.80	47.8	−0.07 [−0.89, 0.73]	0.622 / 0.751	8.1	−0.74 [−1.85, 0.00]	0.983 / 0.983
GPT-OSS 120B	GPT-6 Astra	0.70	87.8	+1.19 [−0.07, 2.55]	0.162 / 0.191	63.3	−7.41 [−11.11, −4.44]	0.392 / 0.401
Gemini 3 Flash	GPT-6 Astra	0.60	96.7	+0.45 [−0.99, 1.89]	0.137 / 0.137	79.6	−4.44 [−7.78, −1.11]	0.307 / 0.307
Qwen3 32B [§]	GPT-6 Astra	0.90	23.1	+0.15 [−0.52, 0.83]	0.791 / 0.966	3.7	−0.37 [−1.11, 0.00]	0.962 / 1.041
Qwen3.5 27B [§]	GPT-6 Astra	0.95	54.8	+0.30 [−0.38, 1.02]	0.454 / 0.541	19.6	−0.37 [−1.11, 0.00]	0.806 / 0.806
Skywork V2 8B [§]	GPT-6 Astra	$ \Delta r \geq 4$	76.4	+2.84 [1.64, 4.13]	0.223 / 0.352	33.7	−6.67 [−10.00, −4.07]	0.663 / 0.663
PairRM [§]	GPT-6 Astra	1.01	0.0	0.00 [0.00, 0.00]	1.000 / 1.218	0.0	0.00 [0.00, 0.00]	1.000 / 1.079

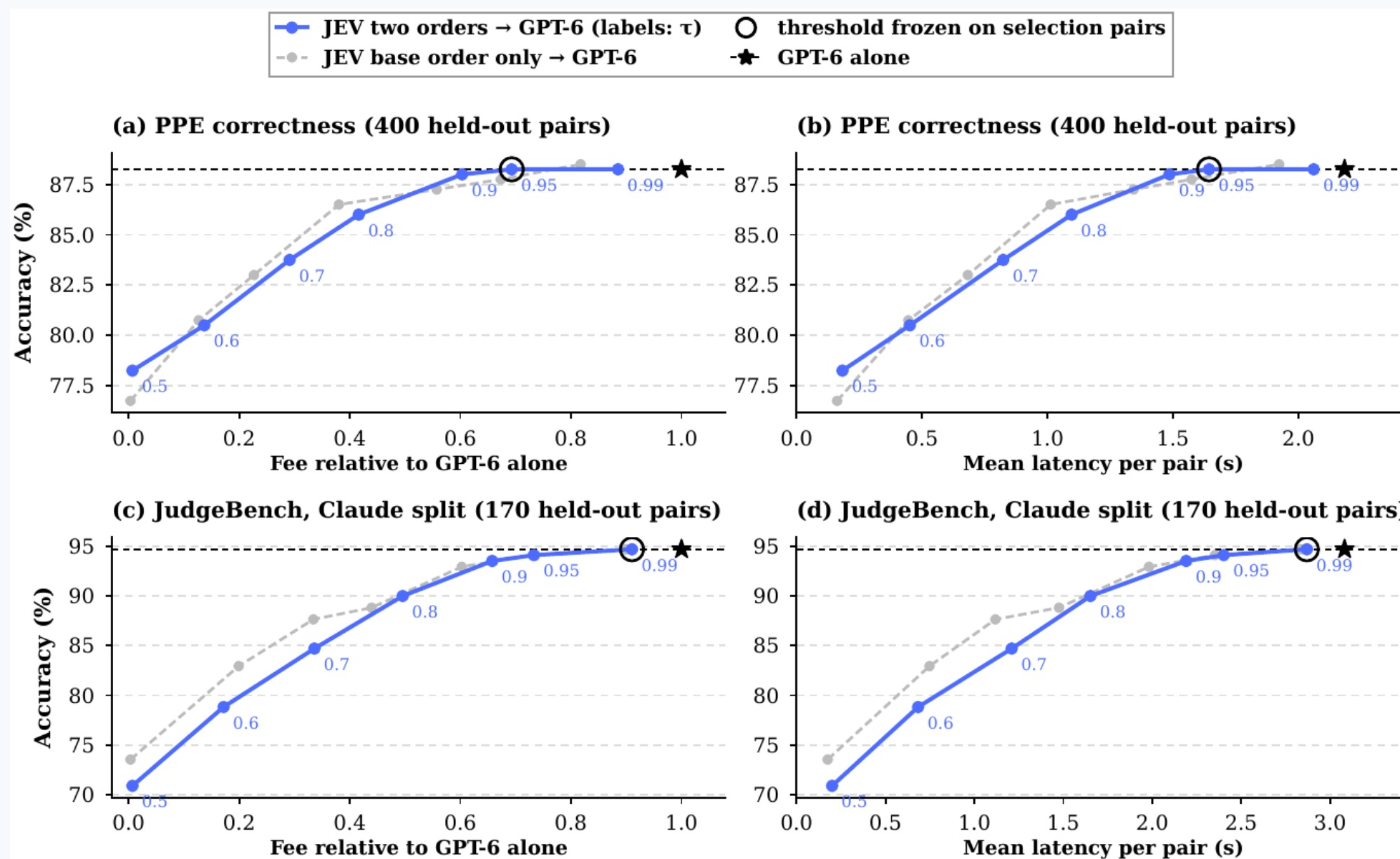
GPT-6 fallback throughout (lower panel retrospective). JEV accepts 75% / 35% of RewardBench / JudgeBench pairs (GPT-4.1 mini: 48% / 8%). GPT-OSS and Gemini 3 Flash break the 2-point tolerance on JudgeBench (−7.41, −4.44).

Deployed and counterfactual thresholds on both new workloads.

Held-out pairs	Policy	JEV alone	Cascade	GPT-6	Δ (pp) [95% CI]	Escalated (%)	JEV errors caught (%)	Fee ratio	Latency (s) p50 / p95
PPE correctness (400)	Deployed, $\tau = 0.95$	78.2	88.2	88.2	0.00 [−1.00, 1.00]	67.8	90	0.693	1.76 / 3.76
	Counterfactual, $\tau = 0.9$	78.2	88.0	88.2	−0.25 [−1.25, 0.75]	58.5	86	0.603	1.64 / 3.76
	GPT-6 Astra alone		88.2	88.2	—	100.0	—	1.000	1.82 / 4.48
JudgeBench, Claude split (170)	Deployed, $\tau = 0.99$	70.9	94.7	94.7	0.00 [0.00, 0.00]	89.4	100	0.910	2.41 / 5.97
	Counterfactual, $\tau = 0.9$	70.9	93.5	94.7	−1.18 [−2.94, 0.00]	62.9	94	0.658	1.82 / 5.84
	GPT-6 Astra alone		94.7	94.7	—	100.0	—	1.000	2.35 / 7.14
Both workloads (570)	Deployed, per workload	76.1	90.2	90.2	0.00 [−0.70, 0.70]	74.2	93	0.755	1.87 / 4.95
	Counterfactual, $\tau = 0.9$	76.1	89.6	90.2	−0.53 [−1.58, 0.35]	59.8	89	0.619	1.71 / 4.74
	GPT-6 Astra alone		90.2	90.2	—	100.0	—	1.000	1.91 / 5.58

The general $\tau = 0.9$ would have lost 0.5 points overall. Latency p50 / p95 per pair.

When the first stage is weak, the procedure spends more rather than losing accuracy.



Every threshold scored on the same held-out outputs; only the ringed points ran live. The gate helps least on IFEval (AUROC 0.57).

Savings follow the workload; accuracy holds either way.

POLICY	ACCURACY VS GPT-6 (PP)	FEE SHARE	≈ \$ PER MILLION	NOTE
GPT-6 alone	—	100%	\$12.2k	every item pays for reasoning
JEV alone	− 1.7 pooled; − 14.6 JudgeBench	0.36%	\$44	only where the verdict is in the text
Cascade · RewardBench held-out	+1.27	27%	≈ \$3.3k	verdicts mostly in the text
Cascade · all held-out pairs	+0.93	41%	≈ \$5.0k	frozen policy, 1,610 pairs
Cascade · JudgeBench held-out	−0.74	67%	≈ \$8.1k	closes 94% of the gap
Cascade · new hard workloads	0.00	75.5%	≈ \$9.2k	live, pre-specified

Fee share = cascade fee (JEV in both orders, plus GPT-6 on escalated pairs) over GPT-6 alone on the same items. ≈ \$ per million applies that share to GPT-6’s timing-panel fee (\$12.18 per 1,000) for scale only; real fees depend on prompt length and price. Live: 0.27 s median per pair vs 2.10 s.

Allocate evaluation compute the way routing allocates inference compute.

- 1 Cheap**
Decisions read off the text, judged with confidence: accept JEV's verdict.
[tested here](#)
- 2 Reasoning**
Low confidence, or a verdict that must be derived: escalate to GPT-6.
[tested here](#)
- 3 Human**
Still uncertain, or high-stakes: final human review. [untested](#)

Like an agent choosing its next tool, an evaluation system can choose its next judge by the task, the judge's confidence, and the cost of being wrong. This paper tests the first two tiers with a confidence gate and an accuracy tolerance; a human tier and risk-based routing are future work.

Escalation is one design among several. We tested the simplest.

01

HOW JUDGES COMBINE

Serial vs fused

Serial: the stronger judge's verdict **replaces** JEV's. Fused: a panel votes or averages (Verga et al., 2024), so every judge runs on every item.

Serial pays for GPT-6 only on **escalated** items.

02

WHAT THE SECOND JUDGE SEES

Independent vs conditioned

The second judge can see the first verdict, as in multi-agent debate (Chan et al., 2024): more information, but its errors then depend on the first judge's.

Ours: GPT-6 **never sees** JEV's output.

03

WHAT ESCALATION CAN FIX

Correlated errors

Errors of different LLMs are substantially correlated (Kim et al., 2025). Under serial replacement, an error both judges make cannot be repaired.

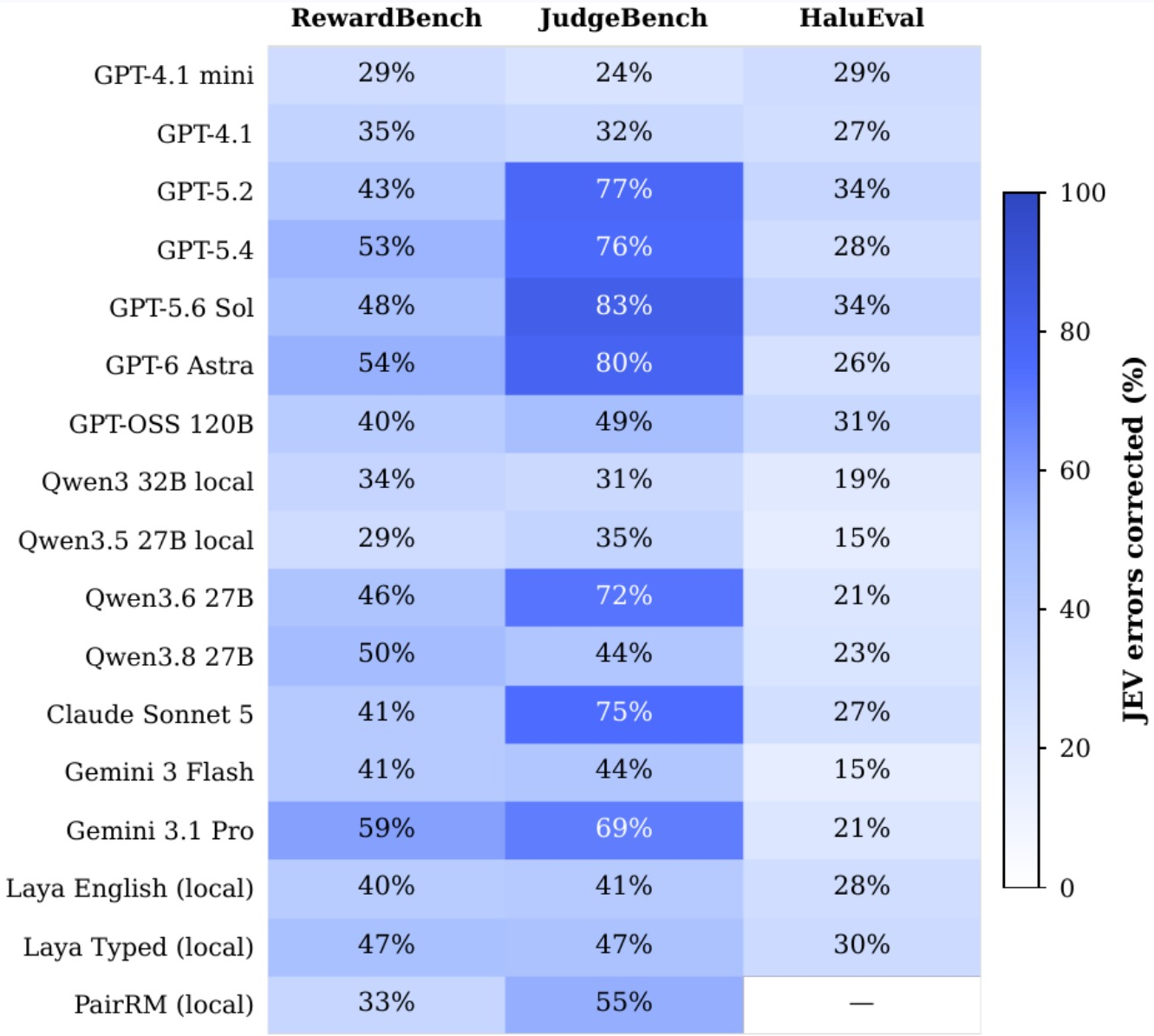
Thresholds from error costs, fusion and conditioning: open.

The lower-bound rule chose strict thresholds on both new workloads.

Workload	τ	Accepted (%)	Mean Δ (pp)	95% lower bound (pp)	Point rule	Lower-bound rule
PPE correctness	0.5	100	−12.50	−19.25	no	no
	0.6	85	−12.00	−17.86	no	no
	0.7	68	−8.00	−12.49	no	no
	0.8	54	−4.00	−7.24	no	no
	0.9	38	−2.00	−4.31	yes	no
	0.95*	24	0.00	0.00	yes	yes
	0.99	13	0.00	0.00	yes	yes
JudgeBench, Claude split	0.5	100	−17.00	−23.64	no	no
	0.6	82	−11.00	−16.68	no	no
	0.7	67	−9.00	−13.73	no	no
	0.8	50	−5.00	−8.60	no	no
	0.9	33	−2.00	−4.31	yes	no
	0.95	18	−1.00	−2.64	yes	no
	0.99*	8	0.00	0.00	yes	yes

100 selection pairs per workload. The point rule would have chosen $\tau = 0.9$ on both; one pooled lower-bound threshold would have been 0.95.

The two judges err on different items, which bounds what a cascade can gain.



Fraction of JEV's errors each comparator gets right. Oracle complementarity motivates deferral; implementing it needs a signal available at decision time.

Where this paper sits.

1 Cascades and routing

FrugalGPT cascades models to cut cost; RouteLLM learns routing from preference data.

2 Escalating LLM judges

Trust or Escalate (Jung et al., 2025): guarantees on human agreement. Xu et al. (2025): route uncertain reward-model comparisons to a strong LLM judge.

1 Selective prediction

Selective classification trades coverage for risk (Geifman and El-Yaniv, 2017); our threshold rule fixes the risk and minimizes cost.

2 Our angle

Keep a stronger reasoning judge's accuracy at a fraction of its fee, with thresholds frozen in advance, fees measured, and a live test on new workloads.