Credibility Gap Analysis

Measuring disagreement between human judges and GPT-4 judge across task categories.

1. The Crisis: AI Judge Bias.
2. The Solution: Human-Validated Baseline
3. Conclusion: Human + SOTA Recommendation

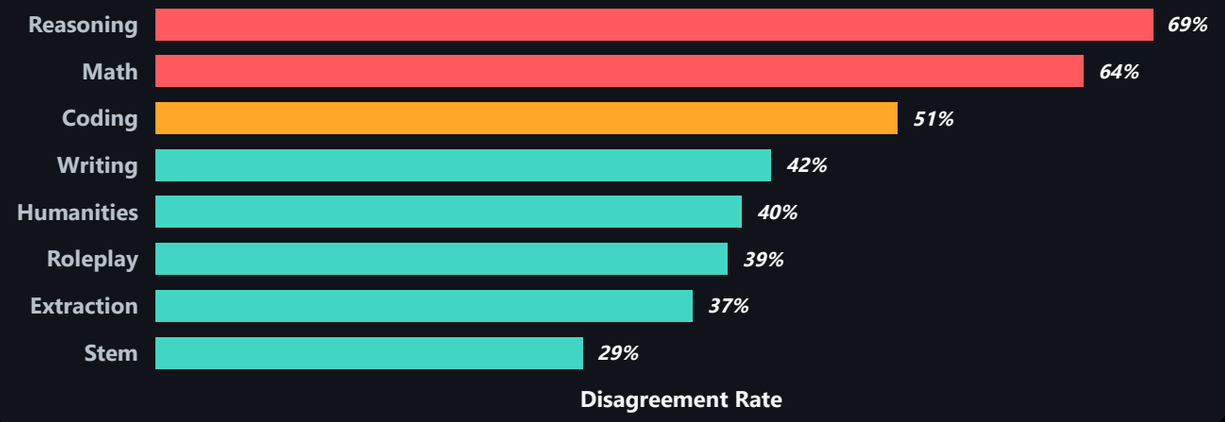
All

53.6%
Overall Agreement

46.4%
Overall Disagreement

3,502
Total Comparisons

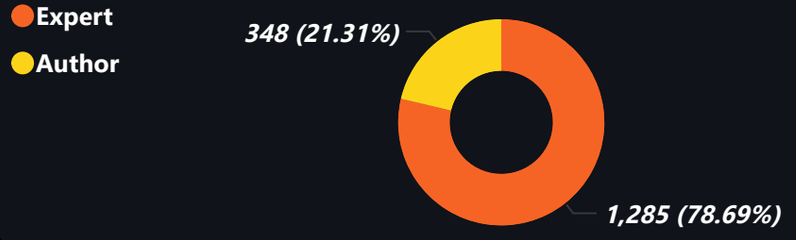
Disagreement by Task Category



Outcome Bias Across Models

Model	Win Rate : GPT	Win Rate : Human	Win Rate Difference
Gpt-4	85.8%	73.0%	12.79%
Claude-V1	70.6%	65.1%	5.56%
Vicuna-13B-V1.2	45.8%	48.0%	-2.25%
Llama-13B	12.7%	15.6%	-2.94%
Gpt-3.5-Turbo	60.9%	65.2%	-4.37%
Alpaca-13B	24.3%	28.7%	-4.42%

How Often Judges Disagree with GPT



🔍 Why Evaluation Credibility Is at Risk

⚠️ **GPT-4 judge disagrees with human judges in ~46% of comparisons.**

- **Experts account for ~79% of all disagreements with GPT.**
- **Reasoning (≈69%) and Math (≈64%) show the highest disagreement rates.**
- **Several models appear stronger only when judged by GPT → model-to-model bias.**

Recommendation: Human validation is essential for credible evaluation.

Human-Validated Model Matrix

The most trustworthy representation of LLM strengths across tasks, derived only from real human judgements

1. The Crisis: AI Judge Bias.
2. The Solution: Human-Validated Baseline
3. Conclusion: Human + SOTA Recommendation

All

53.6%
Overall Agreement

46.4%
Overall Disagreement

3,502
Total Comparisons

Win Rate and Total Appearances by Task and Model

Human-Preferred Performance by Task

Model	Coding	Extraction	Humanities	Math	Reasoning	Roleplay	Stem	Writing
Alpaca-13B	32%	29%	22%	22%	39%	30%	23%	32%
Claude-V1	69%	68%	66%	54%	63%	63%	72%	66%
Gpt-3.5-Turbo	71%	71%	65%	75%	54%	66%	57%	63%
Gpt-4	67%	75%	81%	76%	68%	69%	82%	65%
Llama-13B	11%	14%	10%	24%	27%	13%	7%	19%
Vicuna-13B-V1.2	51%	35%	56%	42%	48%	45%	55%	52%

🔍 Why This Baseline Matters

- GPT-3.5-turbo is the top coding model (best value & functional correctness).
- Claude-v1 excels at writing, closest to human-like style and empathy.
- GPT-4 dominates in difficult reasoning and STEM tasks.

These insights provide the only unbiased baseline for fair benchmarking.

Task-Based Model Guidance

Practical recommendations for each task, aligned with humans and supported by State-Of-The-Art (SOTA) results

All

1. The Crisis: AI Judge Bias.
2. The Solution: Human-Validated Baseline
3. Conclusion: Human + SOTA Recommendation

53.6%
Overall Agreement

46.4%
Overall Disagreement

3,502
Total Comparisons

Human-Validated + SOTA-Aligned Model Recommendations

Task Category	Human-Preferred Model	2025 SOTA Recommendation	Reason & Validated Evidence
Coding/Agentic	GPT-3.5-turbo (Best Value)	Gemini 3 Pro / Claude 3.7 Sonnet	Gemini 3 Pro leads LiveCodeBench Elo (2,439). Claude 3.7 Sonnet has superior real-world performance on SWE-Bench.
Extraction & STEM	GPT-4 (Highest Win-Rate)	GPT-5.1 / Gemini 3 Pro	GPT-5.1 shows strong general performance (GPQA 88.1%). Gemini 3 Pro excels due to native multimodality (diagrams/charts).
Reasoning/Math	GPT-4 (Highest Win-Rate)	Gemini 3 Pro	Gemini 3 Pro leads in novel reasoning (ARC-AGI-2) and symbolic math (AIME), making it the technical SOTA.
Writing/Creative	Claude-v1 (Best Style)	Claude 3.7 Sonnet / GPT-5.1	Anthropic models still excel at coherent, safe, and long-form human-like dialogue, maintaining their lead in stylistic preference.

🧠 Reasoning & Math

Baseline: GPT-4
SOTA: Gemini 3 Pro
Why: Best novel reasoning & symbolic math accuracy (ARC-AGI-2, AIME)

👩‍💻 Coding & Agents

Baseline: GPT-3.5-turbo
SOTA: Gemini 3 Pro / Claude 3.7 Sonnet
Why: Leading SWE-Bench & LiveCodeBench reliability

✍️ Writing & Creativity

Baseline: Claude-v1
SOTA: Claude 3.7 Sonnet / GPT-5.1
Why: Strongest natural prose + style + safety alignment

🔬 STEM & Extraction

Baseline: GPT-4
SOTA: GPT-5.1 / Gemini 3 Pro
Why: Best fact-based performance and accuracy scores

*Conclusion: AI evaluation alone misleads; a hybrid approach using **human-validated baselines** + **verified SOTA improvements** delivers the most trustworthy model recommendations.*