Rubric as Reward: Decomposing Verification Signals for Logical Reasoning in GRPO

Anonymous authors
Paper under double-blind review

Abstract

Reinforcement learning from verifiable rewards (RLVR) has improved LLM reasoning, yet reward functions remain monolithic: a model producing a correct answer via flawed reasoning receives the same signal as one reasoning validly but extracting the wrong answer. We propose rubric-grounded rewards, a framework that decomposes reward into independently weighted criteria spanning a verifiable-to-soft spectrum. Applied to logical reasoning, our five-criterion rubric separates answer correctness, Z3-checked step validity, and format compliance (all machine-verifiable) from premise utilization and reasoning completeness (requiring judgment). We train Qwen2.5-3B-Instruct via GRPO under five reward conditions and evaluate on 166 hard FOLIO and ProntoQA examples. Three findings emerge: (1) rubric-structured verifiable rewards achieve the highest accuracy (51.8%, +6.6pp over baseline) with the most balanced True/False/Unknown performance; (2) rubric profiling reveals that conditions with near-identical accuracy exhibit substantially different quality profiles, exposing an "optimization tax" where RL training improves verifiable criteria while degrading soft ones; and (3) reward structure matters independently of reward content, as decomposing the same verification signals into explicit criteria outperforms their monolithic composite.

1 Introduction

Reinforcement learning from verifiable rewards has emerged as a powerful paradigm for improving LLM reasoning, with symbolic solvers providing dense training signals in mathematics (Shao et al., 2024), code generation (DeepSeek-AI, 2025), and logical inference (Xu et al., 2025; Chen et al., 2025). The core appeal is that verification bypasses the noise and cost of human preference labels. However, current verification-based rewards are monolithic: they conflate multiple quality dimensions into a single scalar. A model that produces a correct final answer through invalid intermediate reasoning receives the same reward as one that reasons flawlessly but misidentifies the conclusion. This conflation limits both the training signal (the policy cannot distinguish which quality dimension to improve) and evaluation interpretability (aggregate accuracy obscures systematic reasoning failures).

In educational assessment, this problem is well understood. Rubric-based evaluation decomposes performance into independently scored criteria with explicit weights, enabling both targeted feedback and diagnostic comparison (Lee & Hooker, 2024). Recent work has applied rubric-structured rewards in non-verifiable domains such as helpfulness scoring (Wang et al., 2024b) and open-ended dialogue evaluation (Kim et al., 2024), where all criteria require LLM judges. We observe, however, that logical reasoning occupies a semi-verifiable middle ground: some quality dimensions (correctness, step validity, format) can be machine-checked, while others (premise utilization, argument completeness) require judgment. This makes logical reasoning an ideal testbed for combining formal verification with rubric structure.

We introduce rubric-grounded rewards, a framework that decomposes the reward signal into five independently weighted criteria (Figure 1), spanning a spectrum from fully verifiable (Z3

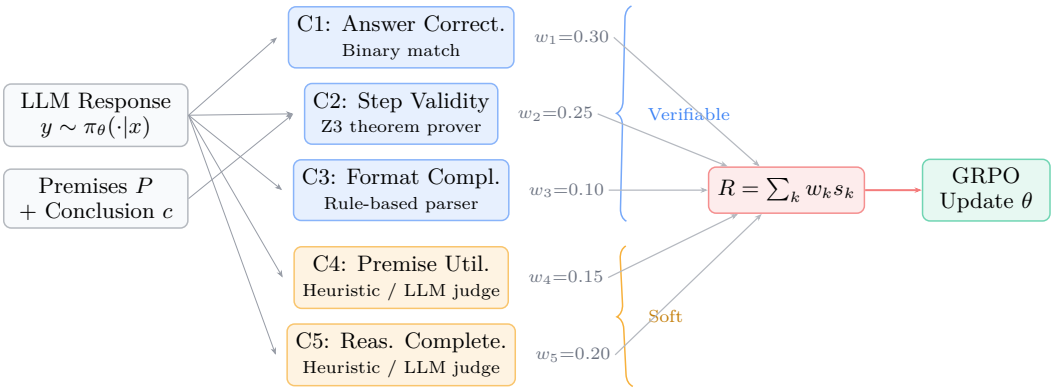

Figure 1: Rubric-grounded reward for logical reasoning. Given an LLM response and task context, five criteria independently score quality dimensions spanning a verifiable-to-soft spectrum. Scores are aggregated via explicit weights into a composite reward for GRPO policy updates. Criteria C1–C3 use deterministic verification (binary match, Z3 prover, regex parser); C4–C5 use lightweight heuristics during training and LLM judges during evaluation.

theorem proving) to soft (heuristic and LLM-judged). Our approach is the first to combine formally checkable and judgment-based criteria within a single rubric for RL training.

Contributions.

1. A five-criterion rubric reward for semi-verifiable logical reasoning, combining Z3-based verification with soft quality criteria (§2).
2. Controlled experiments across five reward conditions showing that rubric-structured rewards achieve the highest accuracy and most balanced answer-type performance (§3).
3. Evidence that reward structure matters beyond reward content: decomposing identical verification signals into explicit criteria outperforms their monolithic composite, and rubric profiling reveals an "optimization tax" invisible to accuracy alone.

2 Method

2.1 Task Formulation

We study first-order logical reasoning: given a set of premises $P = \{p_1, \ldots, p_n\}$ and a conclusion c, the model must determine whether c is True (follows from P), False (contradicted by P), or Unknown (neither provable nor refutable), outputting structured reasoning in <reasoning>…</reasoning> <answer>…</answer> format.

2.2 Rubric Design

We define a five-criterion rubric $\mathcal{C} = \{C_1, \ldots, C_5\}$ with associated weights $\mathbf{w} = (w_1, \ldots, w_5)$ and per-criterion scorers $s_k : \mathcal{Y} \times \mathcal{X} \to [0, 1]$. The composite reward is $R(y, x) = \sum_{k=1}^{5} w_k \cdot s_k(y, x)$, scaled to $[0, 3]$ for GRPO compatibility. Table 1 summarizes the rubric.

C1: Answer Correctness (weight 0.30, verifiable). Binary: 1.0 if the extracted answer matches the ground truth label, 0.0 otherwise.

C2: Step Formal Validity (weight 0.25, verifiable). Each reasoning step is parsed and verified via the Z3 SMT solver (de Moura & Bjørner, 2008). Steps citing valid inference rules (modus ponens, universal instantiation, etc.) are checked for logical entailment from the premises. Each step receives graded credit $v(s_i) \in \{0, 0.3, 0.5, 1.0\}$ depending on whether the derived proposition is provably entailed (1.0), satisfiable but not entailed (0.5), or syntactically valid but unverifiable (0.3). The criterion score is the mean: $s_2 = \frac{1}{|S|} \sum_i v(s_i)$.

Table 1: Five-criterion rubric for logical reasoning. Criteria are ordered from fully verifiable (deterministic, no model calls) to soft (requiring judgment). Training and evaluation use different scorers for C4–C5 to balance cost and fairness.

	Criterion	Weight	Type	Scorer
C1	Answer Correctness	0.30	Verifiable	Binary match vs. ground truth
C2	Step Formal Validity	0.25	Verifiable	Z3 graded: $\bar{v} = \frac{1}{\lvert S \rvert} \sum_i v(s_i)$
C3	Format Compliance	0.10	Verifiable	Incremental tag/step parser
C4	Premise Utilization	0.15	Soft	Train: text heuristic; Eval: GPT-4o-mini
C5	Reasoning Completeness	0.20	Soft	Train: structural heuristic; Eval: GPT-4o-mini

C3: Format Compliance (weight 0.10, verifiable). Incremental credit for structural elements: 0.2 each for <reasoning>, </reasoning>, <answer>, </answer> tags, and 0.2 for $\geq$2 numbered steps.

C4: Premise Utilization (weight 0.15, soft). Measures whether reasoning engages the provided premises. During training, a text-overlap heuristic scores the fraction of premises whose key terms appear in the response. During evaluation, a GPT-4o-mini judge scores on a 0–1 rubric.

C5: Reasoning Completeness (weight 0.20, soft). Measures whether the argument chain is sufficient to justify the conclusion. During training, a structural heuristic based on step count and reasoning length provides a fast proxy. During evaluation, GPT-4o-mini judges completeness on a 0–1 rubric.

The train-time heuristic and eval-time judge asymmetry for C4–C5 avoids costly API calls during GRPO's K=8 generations per prompt (1,328 calls per step) while ensuring fair cross-condition comparison at evaluation.

2.3 GRPO Training

We use Group Relative Policy Optimization (DeepSeek-AI, 2025), which eliminates the critic network by normalizing rewards within each prompt's generation group. For prompt x with K sampled completions $\{y_1, \ldots, y_K\}$, the advantage is:

$$\hat{A}_i = \frac{R(y_i, x) - \mu_R}{\sigma_R + \epsilon}, \qquad \mu_R = \frac{1}{K} \sum_{j=1}^{K} R(y_j, x), \quad \sigma_R = \text{std}(\{R(y_j, x)\}_{j=1}^{K}) \tag{1}$$

where $R(y, x)$ is the rubric composite reward. The policy is updated via clipped importance-weighted gradients with a KL penalty against the reference policy π_{ref}. Crucially, because our rubric decomposes R into independently weighted criteria, the advantage signal reflects which quality dimensions differ across completions, providing richer gradient information than a monolithic reward.

2.4 Experimental Conditions

We train five conditions with progressively richer reward signals, all using Qwen2.5-3B-Instruct (Qwen Team, 2024) with LoRA (Hu et al., 2022) (r=32, α=64, targeting all attention and MLP projections), 250 GRPO steps, batch size 8, K=8 generations, learning rate 5×10^{-6}, bf16, single H100 GPU, single seed:

1. Baseline: Zero-shot Qwen2.5-3B-Instruct (no fine-tuning).
2. Outcome-only: Correctness + format reward (no step verification).
3. Z3-verified: Correctness + format + Z3 step validity as a monolithic weighted sum.
4. Rubric-verifiable: C1+C2+C3 with explicit per-criterion weights (renormalized). Identical verification signals to Condition 3, different reward structure.
5. Rubric-hybrid: Full C1–C5 rubric with heuristic C4/C5 during training.

Table 2: Results on 166 hard test examples. C1–C5: per-criterion rubric scores (0–1). Total: weighted composite. Best per column in bold. Rubric-verifiable achieves the highest accuracy using the same verification signals as Z3-verified, differing only in reward structure.

Condition	Acc	C1	C2	C3	C4	C5	Total
1. Baseline	.452	.452	.060	.442	.616	.628	.413
2. Outcome-only	.458	.458	.155	.777	.611	.596	.465
3. Z3-verified	.500	.500	.348	.831	.551	.547	.512
4. Rubric-verif.	.518	.518	.295	.771	.575	.547	.502
5. Rubric-hybrid	.464	.464	.314	.787	.563	.547	.490

Conditions 2–3 reuse trained models from prior experiments; 4–5 are newly trained (107 and 156 minutes on H100 respectively). The comparison between Conditions 3 and 4 is particularly informative: both use the same three verification signals (correctness, Z3, format), but Condition 3 combines them as a flat weighted sum while Condition 4 decomposes them into an explicit rubric with per-criterion normalization. This isolates the effect of reward structure from reward content.

3 Results

We evaluate all five conditions on 166 hard test examples (109 FOLIO (Han et al., 2022) + 57 ProntoQA (Saparov & He, 2023), excluding trivially easy synthetic examples) with LLM-judged C4/C5 to ensure fair comparison.

3.1 Main Results

Finding 1: Rubric structure improves accuracy. Rubric-verifiable achieves the highest accuracy (51.8%, +6.6pp over baseline, +1.8pp over Z3-verified) despite using the same three verification signals as Condition 3. The only difference is reward structure: explicit per-criterion weights with renormalization versus a flat composite. This suggests that decomposing reward into independently normalized criteria provides cleaner gradient signal to the policy, even when the underlying verification content is identical.

Finding 2: Rubric profiling reveals hidden quality differences. Conditions 1 and 2 differ by only 0.6pp in accuracy, yet their rubric profiles diverge substantially (Figure 2): outcome-only scores +9.5pp on C2 (step validity) and +33.5pp on C3 (format compliance), demonstrating that GRPO teaches structural reasoning behavior even when aggregate accuracy barely moves. Without rubric decomposition, this improvement would be entirely invisible.

Finding 3: The optimization tax. All RL-trained conditions score lower on C4 (premise utilization) and C5 (reasoning completeness) than the untrained baseline (0.55–0.61 vs. 0.62–0.63). This "optimization tax" suggests that RL training narrows reasoning toward patterns that maximize verifiable criteria at the expense of broader premise engagement and argument completeness. Notably, this degradation is consistent across all four trained conditions regardless of whether soft criteria are included in the reward (Condition 5), suggesting that heuristic proxies for C4/C5 are insufficient to counteract the strong gradient signal from verifiable criteria.

Finding 4: Heuristic soft criteria add limited value. Rubric-hybrid (C1–C5) does not outperform rubric-verifiable (C1–C3), likely because the lightweight heuristic proxies for premise utilization and reasoning completeness are too noisy to provide useful gradient signal during training. The hybrid condition does achieve slightly higher C2 than rubric-verifiable (0.314 vs. 0.295), hinting that soft criteria provide some indirect signal for step quality, but this does not translate to accuracy gains.

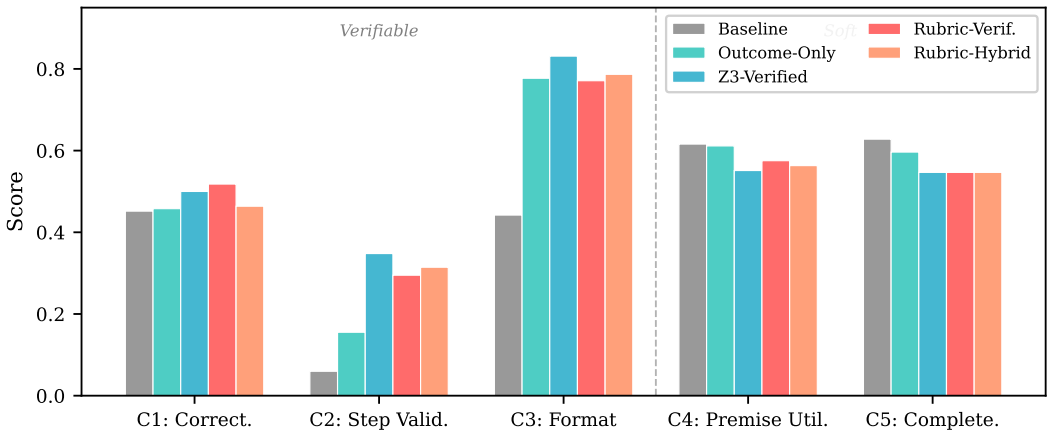

Figure 2: Per-criterion rubric scores across all five conditions. The dashed vertical line separates verifiable criteria (C1–C3, left) from soft criteria (C4–C5, right). RL training consistently improves verifiable criteria while degrading soft ones, an "optimization tax" that is invisible to accuracy alone.

Table 3: Accuracy by answer type and data source. Rubric-verifiable is the most balanced condition, achieving the highest accuracy on both False and Unknown categories. Conditions 3 and 5 show pronounced True-prediction bias.

Condition	True	False	Unkn.	FOLIO	PQA
1. Baseline	.480	.400	.472	.532	.298
2. Outcome-only	.453	.455	.472	.523	.333
3. Z3-verified	.600	.382	.472	.477	.544
4. Rubric-verif.	.520	.491	.556	.541	.474
5. Rubric-hybrid	.587	.345	.389	.468	.456

3.2 Answer-Type Balance

Table 3 reveals systematic biases that aggregate accuracy obscures. Conditions 3 (Z3-verified) and 5 (rubric-hybrid) achieve high True accuracy ($\geq$0.587) but low False accuracy ($\leq$0.382), indicating True-prediction bias. Rubric-verifiable is the most balanced condition: it achieves the highest accuracy on both False (0.491) and Unknown (0.556) while maintaining competitive True accuracy (0.520). This balance is particularly important for logical reasoning, where correctly rejecting invalid arguments and recognizing underdetermined conclusions is as important as confirming valid ones.

On data sources, rubric-verifiable leads on FOLIO (0.541), which tests complex natural language premises, while Z3-verified leads on ProntoQA (0.544), which uses synthetic chains more amenable to formal verification. This suggests that rubric structure particularly benefits tasks requiring flexible integration of multiple reasoning quality signals.

4 Related Work

Formal verification for LLM reasoning. Several recent works use symbolic solvers to provide reward or supervision signals for logical reasoning. LogicReward (Xu et al., 2025) uses Z3 for step-wise verification during RL training. FOVER (Kamoi et al., 2025) generates formally verified training data for process reward models. VeriCoT (Feng et al., 2025) validates chain-of-thought consistency via logical checks. ProSFI (Chen et al., 2025) generates formal intermediaries to bridge natural and symbolic reasoning. All of these approaches use monolithic reward composites that aggregate verification signals into a single scalar, without the criterion-level decomposition that enables the diagnostic insights we report.

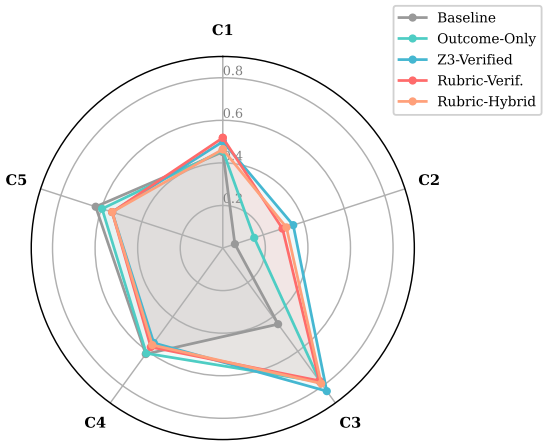

Figure 3: Rubric quality profiles (radar chart). Each axis represents one criterion (C1–C5). The baseline (gray) has high C4/C5 but low C2/C3; trained conditions show the opposite pattern. Rubric-verifiable (red) achieves the largest C1 while maintaining moderate soft scores.

Rubric-based reward modeling. Decomposed reward modeling has been explored in non-verifiable settings. Lee & Hooker (2024) propose rubric-structured rewards for open-ended generation. HelpSteer2 (Wang et al., 2024b) defines multi-attribute scoring for helpfulness. LLM-as-judge frameworks (Zheng et al., 2024; Kim et al., 2024) provide rubric-based evaluation using model judges. These approaches apply exclusively to soft, non-verifiable domains. Our work is the first to combine formally checkable and judgment-based criteria within a single rubric for RL reward computation.

RLVR and process rewards. DeepSeek-R1 (DeepSeek-AI, 2025) and DeepSeekMath (Shao et al., 2024) demonstrate that verifiable outcome rewards improve reasoning at scale. Step-level process reward models (Lightman et al., 2024; Wang et al., 2024a) provide denser supervision by scoring intermediate steps. Our rubric approach can be viewed as a principled framework for combining outcome-level and process-level signals with explicit weights and independent normalization, bridging the gap between these paradigms.

5 Discussion and Conclusion

We introduced rubric-grounded rewards, a framework for decomposing verification signals into independently weighted criteria for RL training of LLM reasoning. Applied to logical reasoning with GRPO, three findings stand out. First, reward structure matters independently of reward content: the same Z3-based verification signals achieve higher accuracy when decomposed into explicit rubric criteria than when combined as a flat composite. Second, rubric profiling provides diagnostic power that aggregate accuracy lacks, revealing an "optimization tax" where RL training consistently improves machine-verifiable quality dimensions while degrading softer ones. Third, heuristic proxies for judgment-based criteria are insufficient to counteract this tax during training, motivating future work on efficient soft reward approximation.

Limitations. This study uses a single model (3B parameters), a single random seed, and a relatively small test set (166 examples). The heuristic C4/C5 proxies during training are coarse approximations of the LLM judge used at evaluation. Better approximations, such as distilled judge models or learned reward functions for soft criteria, could unlock the hybrid rubric's potential and are a natural next step.

Future directions. The rubric-as-reward framework is domain-agnostic: any task with a mix of verifiable and subjective quality dimensions can benefit from structured decomposition. Natural extensions include distilling lightweight judge models for efficient training-time soft evaluation, extending the rubric to mathematical reasoning (where step verification is well-

developed but premise utilization is underexplored), and investigating whether adaptive criterion weights during training can mitigate the optimization tax.

References

Luoxin Chen, Yichi Zhou, and Huishuai Zhang. Learning to generate formally verifiable step-by-step logic reasoning via structured formal intermediaries. In Submitted to International Conference on Learning Representations (ICLR), 2025.

Leonardo de Moura and Nikolaj Bjørner. Z3: An efficient SMT solver. In Tools and Algorithms for the Construction and Analysis of Systems (TACAS), volume 4963 of Lecture Notes in Computer Science, pp. 337–340. Springer, 2008.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

Yu Feng, Nathaniel Weir, Kaj Bostrom, Sam Bayless, Darion Cassel, Sapana Chaudhary, Benjamin Kiesl-Reiter, and Huzefa Rangwala. VeriCoT: Neuro-symbolic chain-of-thought validation via logical consistency checks. arXiv preprint arXiv:2511.04662, 2025.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Ansong Ni, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic. arXiv preprint arXiv:2209.00840, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations (ICLR), 2022.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Sarkar Snigdha Sarathi Das, and Rui Zhang. Generalizable process reward models via formally verified training data. arXiv preprint arXiv:2505.15960, 2025.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. arXiv preprint arXiv:2405.01535, 2024.

Ryan Lee and Sara Hooker. Decomposed reward modeling with scoring rubrics. arXiv preprint arXiv:2410.15035, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. arXiv preprint arXiv:2305.20050, 2024.

Qwen Team. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In International Conference on Learning Representations (ICLR), 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

Peiyi Wang, Lei Li, Zhihong Shao, R.X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. arXiv preprint arXiv:2312.08935, 2024a.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer2: Open-source dataset for training top-performing reward models. arXiv preprint arXiv:2406.08673, 2024b.

Jundong Xu, Hao Fei, Huichi Zhou, Xin Quan, Qijun Huang, Shengqiong Wu, William Yang Wang, Mong-Li Lee, and Wynne Hsu. LogicReward: Incentivizing LLM reasoning via step-wise logical supervision. arXiv preprint arXiv:2512.18196, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 36, 2024.