When Accuracy Hides Path Dependence: Criterion Fidelity in LLM Judges

Ali Uyar
Independent Researcher

Abstract

Large language model judges are commonly validated by ordinary task accuracy or judge–reference agreement, implicitly assuming that a judge that looks accurate is following the provided criterion. We argue that this is incomplete: in criterion-conditioned evaluation, a valid judge should change when the criterion meaning changes and remain stable when wording changes but meaning does not. We formalize this property as *criterion fidelity* and measure it with *criterion families*—matched groups of base, paraphrase, and counterfactual variants evaluated with mirrored-order continuation scoring. Across two controlled task families—QA-Key, a high-margin keyed-answer setting, and BioRubric, a low-margin rubric-conditioned pairwise setting where both candidates are factually true—ordinary quality and criterion fidelity dissociate. Under standard prompting, QA-Key gaps are small but real, while BioRubric gaps are large: for Qwen2.5-14B-Instruct, BASEACC $= 0.783$ but GCF $= 0.498$; for Llama-3.1-8B-Instruct, BASEACC $= 0.727$ but GCF $= 0.217$. Strict criterion-emphasis prompting is not a reliable remedy: it removes the gap on QA-Key/Qwen2.5-14B-Instruct, narrows it on BioRubric/Qwen2.5-14B-Instruct, and worsens both quality and order instability on BioRubric/Llama-3.1-8B-Instruct. Failure analysis shows that underreaction to criterion changes dominates paraphrase instability, but the largest rubric-based failures are often mediated by presentation-sensitive path dependence rather than by a clean majority of direct semantic criterion refusal. On the most pathological slice, Llama-3.1-8B-Instruct evaluated with the strict prompt becomes nearly a pure first-position chooser on BioRubric while remaining almost balanced over underlying candidate identity. These results position criterion fidelity as a distinct validity axis for LLM judges and show that criterion-conditioned instability, not just answer-level inaccuracy, should be part of judge evaluation.

1 Introduction

Large language models are increasingly used as judges: to score candidate outputs, compare systems, filter synthetic data, and stand in for more expensive human evaluation (Chiang and Lee, 2023; Zheng et al., 2023). In practice, these uses are often validated with ordinary accuracy, agreement, or correlation. That standard is useful, but it quietly assumes something stronger: if a judge gets the base case right, then it must be implementing the provided criterion.

This assumption is testable and incomplete. A judge can be *right for the wrong reason*. In criterion-conditioned evaluation, the input, candidate outputs, and rubric or reference are all part of the object being judged. A model that selects the right winner on the base item but fails to change when the criterion changes, or flips when the criterion wording changes but its meaning does not, is not faithfully implementing the criterion even if its base accuracy looks respectable.

We study this property as *criterion fidelity*. Rather than flattening evaluation into independent rows, we group matched items into *criterion families*: a base variant, two meaning-preserving paraphrases, and a counterfactual criterion that should flip the winner. This gives a direct test of whether a judge's decision is actually controlled by the criterion. We evaluate two controlled families. QA-Key provides a high-margin keyed-answer setting close to reference-based question answering. BioRubric provides a harder rubric-conditioned pairwise setting in which both candidates are factually true and only the rubric weights determine the winner.

The resulting picture is sharper—and more interesting—than the narrower "judges choose beliefs over criteria" story. We find that ordinary judge quality and criterion fidelity dissociate on both task families, with especially large gaps on BioRubric. We also find that these failures are

not dominated by a clean majority of direct semantic criterion refusal. Instead, the largest rubric-based failures are often mediated by presentation-sensitive path dependence: the decision depends on candidate serialization even after mirrored-order aggregation, and stronger criterion-emphasis prompting does not reliably fix this pathology. On the most extreme slice, Llama-3.1-8B-Instruct under the strict prompt is almost a pure first-position chooser on BioRubric.

Our contribution is therefore a validity paper, not a new judge benchmark. We make four claims supported by the completed evidence. First, ordinary judge quality is not the same as criterion fidelity. Second, the gap is not only a knowledge-conflict artifact: it appears in a non-QA family where both candidates are true. Third, failures decompose into underreaction and overreaction, with underreaction dominant on the meaningful slices. Fourth, strict criterion-emphasis prompting is an unreliable baseline rather than a solution. A narrower mechanism claim—that the largest failures are mostly direct semantic refusal to follow the criterion—is only partially supported and is treated as such.

2 Criterion Fidelity as a Family-Level Validity Property

A criterion family f contains a fixed task input x_f, two underlying candidates $c_{f,1}$ and $c_{f,2}$, and four logical variants $v \in \{b, p_1, p_2, c\}$: one base criterion, two paraphrases that preserve criterion semantics, and one counterfactual criterion that flips the gold winner. The family is the evaluation unit.

For each logical variant we evaluate both candidate orders, $o \in \{AB, BA\}$. Let $\pi_{f,v,o}$ be the rendered chat prefix that ends immediately before the answer label, and let $\lambda_o(c) \in \{A, B\}$ be the displayed label of underlying candidate c under order o. We score literal continuations "A" and "B" by continuation log-probability from the identical rendered prefix and then remap both orders back to underlying candidate IDs:

$$S_{f,v}(c) = \frac{1}{2} \sum_{o \in \{AB,BA\}} \log p_\theta(\lambda_o(c) \mid \pi_{f,v,o}) \,. \tag{1}$$

We predict the higher-scoring underlying candidate unless $|S_{f,v}(c_{f,1}) - S_{f,v}(c_{f,2})| \le \varepsilon$ with $\varepsilon = 10^{-6}$, in which case we record a tie. Ties are logged and counted incorrect for the binary headline metrics.

This setup separates three distinct questions. Ordinary quality asks whether the base item is correct. Criterion sensitivity asks whether the judge responds correctly when the criterion semantics change. Criterion invariance asks whether the judge remains correct when the criterion wording changes without semantic change. We report four family-level headline metrics:

$$\textsc{BaseAcc} = \frac{1}{|\mathcal{F}|} \sum_f \mathbf{1}[\hat{y}_{f,b} = y_{f,b}], \tag{2}$$

$$\textsc{GCF} = \frac{1}{|\mathcal{F}|} \sum_f \mathbf{1}[\forall v \in \{b, p_1, p_2, c\} : \hat{y}_{f,v} = y_{f,v}], \tag{3}$$

$$\textsc{Sens@BC} = \frac{\sum_f \mathbf{1}[\hat{y}_{f,b} = y_{f,b} \wedge \hat{y}_{f,c} = y_{f,c}]}{\sum_f \mathbf{1}[\hat{y}_{f,b} = y_{f,b}]}, \tag{4}$$

$$\textsc{Inv@BC} = \frac{\sum_f \mathbf{1}[\hat{y}_{f,b} = y_{f,b} \wedge \hat{y}_{f,p_1} = y_{f,p_1}]}{\sum_f \mathbf{1}[\hat{y}_{f,b} = y_{f,b}]} \tag{5}$$

$$\times \mathbf{1}[\hat{y}_{f,p_2} = y_{f,p_2}]. \tag{6}$$

We refer to $1 - \textsc{Sens@BC}$ as *underreaction* and $1 - \textsc{Inv@BC}$ as *overreaction*. As supporting diagnostics we also report OrderDisagreement, the fraction of logical variants whose AB and BA order-specific decisions disagree before aggregation, and the tie rate after mirrored-order aggregation.

Two aspects of the framework matter for interpretation. First, the family-level metrics are intentionally harsher than row-wise accuracy: a family fails GCF if *any* of its four logical variants fails. Second, mirrored-order evaluation does not create path dependence; it reveals it. If a judge behaves differently across candidate serializations, criterion-conditioned evaluation is already unstable.

3 Controlled Task Families

QA-Key and BioRubric are designed to isolate criterion dependence under different margins and different failure temptations.

QA-Key. QA-Key is a controlled keyed-answer family built from six Wikidata-backed relations: place of birth, place of death, date of birth, date of death, capital, and currency. Each family fixes a question and two answer candidates. The base criterion is an official answer key; the counterfactual criterion swaps in a donor answer from the same relation and coarse answer type; the paraphrases preserve answer-key semantics. This family gives a clean high-margin test close to reference-conditioned QA, and therefore remains the most natural place to observe direct semantic criterion failures of the kind highlighted by Lee et al. (2026).

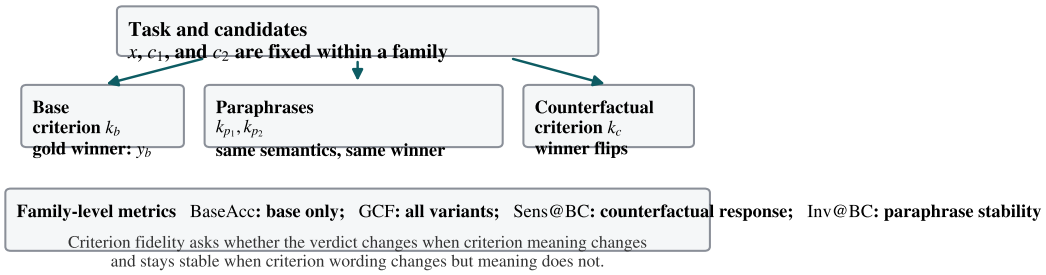

Figure 1: Criterion families keep the task and candidates fixed while varying only the criterion. A judge with high criterion fidelity should change when the counterfactual criterion changes the gold winner and remain stable when paraphrases preserve criterion meaning.

Table 1: Controlled task families. Both families use frozen splits of 2000/400/600 criterion families for train/dev/test.

Task family	Controlled object	Built families	Key caveat
QA-Key	Keyed-answer preference under controlled answer-key swaps over six Wikidata relations spanning place, date, capital, and currency.	6,168	High-margin setting: one candidate matches the key and the other does not. Useful anchor, but still close to reference-conditioned QA.
BioRubric	Two-sentence biographies built from Wikidata truthy statements. Both candidates are factual, share one fact, and differ on one rubric-weighted fact from the same slot.	4,385	Low-margin comparative setting: the criterion must determine which true candidate is preferred. This is the decisive non-QA family.

BioRubric. BioRubric removes knowledge-conflict as the dominant explanation. Each family describes a human entity using two deterministic two-sentence biographies built from Wikidata truthy statements. The biographies share one occupation fact and differ on one additional fact drawn from the same slot type (notable work, award, or field). The criterion is an explicit weighted rubric: in the base variant one distinguishing fact receives higher weight; in the counterfactual variant the other distinguishing fact receives higher weight;

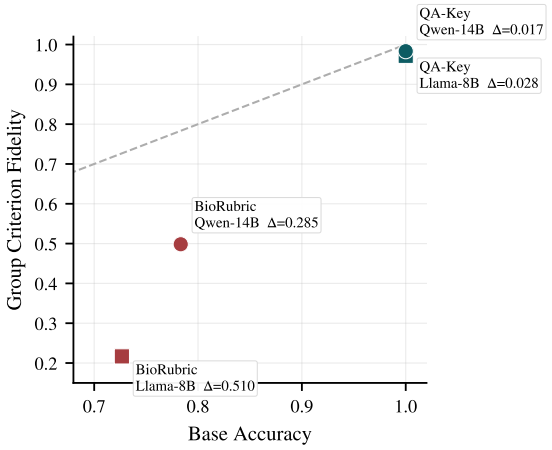

Figure 2: BASEACC versus GCF under standard prompting. The QA slices cluster near the top-right with small but nonzero gaps. The decisive result is BioRubric, where ordinary quality looks much better than family-level criterion fidelity.

paraphrases preserve the rubric semantics. Both candidates are factual. The only thing that should change the winner is the criterion.

The final frozen source contains 6,256 raw QA-Key rows, which become 6,174 canonical rows and 6,168 valid built families before split truncation. For BioRubric, 55,854 truthy fact rows over 5,081 entities become 4,385 valid built families before split truncation. We freeze both families to train/dev/test sizes of 2000/400/600 and keep the processed datasets immutable for the headline results.

4 Ordinary Quality Hides Criterion-Fidelity Failure

Table 2 and Figure 2 establish the paper's central empirical result. Ordinary judge quality and criterion fidelity are not the same object. On QA-Key, the dissociation is modest but statistically real under standard prompting for both models. On

Table 2: Standard-prompt headline results on the frozen test sets. GAP = BASEACC − GCF. All confidence intervals are 95% family bootstrap intervals.

Task	Model	BASEACC	GCF	GAP [95% CI]	SENS@BC	INV@BC	Order dis.	Tie
QA-Key	Qwen-14B	1.000	0.983	0.017 [0.007, 0.027]	0.985	0.998	0.007	0.0004
QA-Key	Llama-8B	1.000	0.972	0.028 [0.015, 0.042]	0.972	1.000	0.180	0.0000
BioRubric	Qwen-14B	0.783	0.498	0.285 [0.250, 0.322]	0.740	0.896	0.292	0.0029
BioRubric	Llama-8B	0.727	0.217	0.510 [0.468, 0.550]	0.429	0.775	0.944	0.0354

BioRubric, the dissociation is large and decisive: Qwen2.5-14B-Instruct loses 0.285 points from BASEACC to GCF, and Llama-3.1-8B-Instruct loses 0.510.

This immediately gives two conclusions. First, criterion fidelity is distinct from ordinary judge quality. Second, the gap is not only a knowledge-conflict artifact. BioRubric is the crucial family here: both candidates are true, and the gold winner changes only because the rubric changes. The BioRubric results therefore show a criterion-conditioned validity failure in a setting where the judge cannot lean on factual correctness alone.

The decomposition into sensitivity and invariance clarifies the structure of the gap. On the large-gap slices, SENS@BC is systematically below INV@BC. For Qwen2.5-14B-Instruct on BioRubric under the standard prompt, SENS@BC is 0.740 while INV@BC is 0.896; for Llama-3.1-8B-Instruct the gap is even larger, 0.429 versus 0.775. The dominant failure is therefore *under-reaction*: the judge fails more often when criterion semantics change than when criterion wording changes but its meaning does not.

This does not mean that paraphrase instability is absent. Rather, it means that the main failure is not ordinary phrasing brittleness. The judges often fail to let the criterion change control the verdict strongly enough. That is already visible in QA-Key, but it becomes unmistakable in BioRubric.

5 Strict Criterion-Emphasis Prompting Is Not a Reliable Remedy

A natural baseline is to instruct the model more forcefully to follow the criterion exactly. Figure 3 shows why this is not the paper's story.

On QA-Key, stricter prompting can remove or nearly remove the gap. For Qwen2.5-14B-Instruct, the standard-prompt gap of 0.017 collapses to 0.000. For Llama-3.1-8B-Instruct, the gap drops from 0.028 to 0.005. These improvements are real, but they occur in the high-margin keyed-answer family where the criterion signal is already strong.

On BioRubric, the picture is qualitatively different. Qwen2.5-14B-Instruct improves, but the problem remains: BASEACC rises from 0.783 to 0.863, GCF from 0.498 to 0.670, and the gap remains a substantial 0.193. Llama-3.1-8B-Instruct is worse. Under the strict prompt, BASEACC falls from 0.727 to 0.577, GCF falls from 0.217 to 0.132, order disagreement rises from 0.944 to 0.998, and the tie rate more than doubles to 0.0879.

The lesson is not that prompting never matters. It clearly does. The lesson is that stronger criterion-emphasis instructions do not reliably make the criterion control the decision. On some slices they help, on others they leave a large residual gap, and on the most pathological slice they amplify the wrong signal.

6 Path Dependence in Rubric-Conditioned Evaluation

The key question after the main results is what kind of failure produces the large BioRubric gaps. The audits and diagnostics point to a clear answer: on the largest-gap slices, criterion-fidelity failure is often mediated by presentation-sensitive path dependence.

Manual audits make the pattern concrete. On BioRubric under standard prompting, `order_effect` accounts for 34/50 audited failures for Qwen2.5-14B-Instruct and 36/50 for Llama-3.1-8B-Instruct. Under strict prompting, the same label still dominates (32/50 for Qwen2.5-14B-Instruct and 36/50 for Llama-3.1-8B-Instruct), and Llama-3.1-8B-Instruct adds 14/50 audited `model_tie` cases. Direct semantic criterion failures remain visible—for example, 15/50 audited Qwen2.5-14B-Instruct standard failures and 17/50 audited Qwen2.5-14B-Instruct strict failures are labeled `genuine_criterion_failure`—but they are not the majority channel on the largest-gap slices.

This does *not* mean the criterion is irrelevant

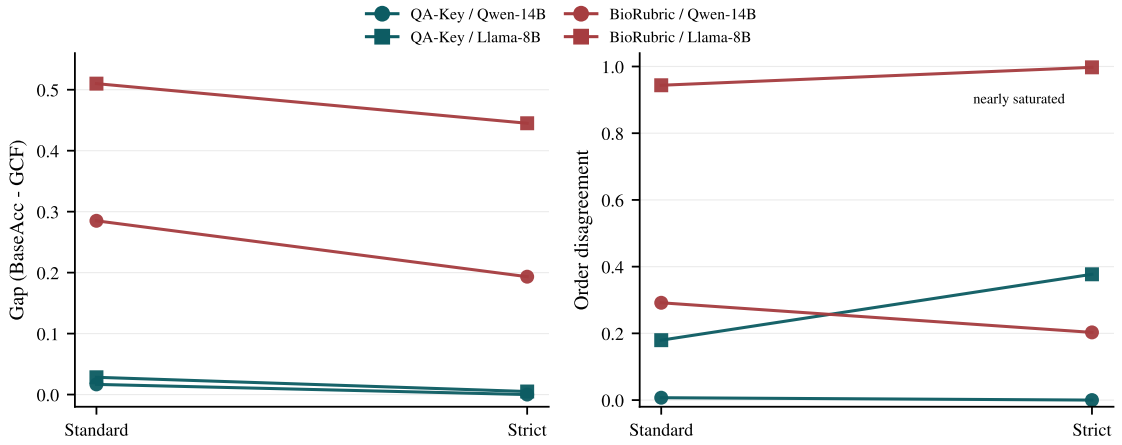

Figure 3: Effect of the strict criterion-emphasis prompt relative to the standard prompt. Left: criterion-fidelity gap. Right: order disagreement. Strict prompting helps some slices, but it is not a reliable remedy and can amplify instability.

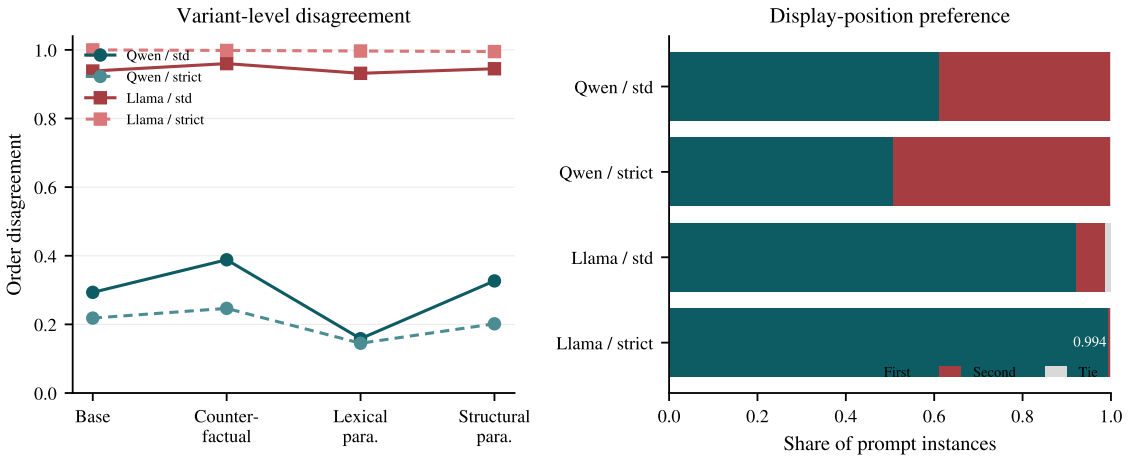

Figure 4: Path-dependence diagnostics on BioRubric. Left: variant-level order disagreement is highest on the counterfactual variants for Qwen2.5-14B-Instruct and is nearly saturated across all variants for Llama-3.1-8B-Instruct. Right: display-position preference reveals the extreme strict-prompt Llama-3.1-8B-Instruct pathology: nearly every prompt instance chooses the first-presented candidate even though the underlying candidate identity remains nearly balanced.

and only order matters. It means that the observable route by which the criterion fails to control the verdict is often order-sensitive instability. The difference matters. A model can fail criterion-conditioned evaluation because it refuses the criterion semantically, because it is brittle to criterion wording, or because its pairwise decision process is path dependent. The completed evidence says that the last route is a major part of the story in close rubric-based settings.

The diagnostics sharpen this claim. First, order disagreement is highest exactly where criterion sensitivity should matter most. For Qwen2.5-14B-Instruct on BioRubric under the standard prompt, the counterfactual variant has the highest disagreement rate (0.388). Under the strict prompt it remains the most unstable variant (0.247). Second, restricting to order-stable families does not eliminate the issue on Qwen2.5-14B-Instruct. The BioRubric gap remains 0.217 on order-stable families under the standard prompt and 0.123 under the strict prompt (Appendix A). Path dependence therefore explains a large part of the failure, but not all of it.

For Llama-3.1-8B-Instruct, however, path dependence is the main event. Under standard prompting, 599 of 600 BioRubric families are order-unstable. Under strict prompting, *every* test family is order-unstable. The right panel of Figure 4 shows what this instability looks like in practice.

On BioRubric with the strict prompt, Llama-3.1-8B-Instruct chooses the first-presented candidate 99.44% of the time, the second-presented candidate only 0.46% of the time, and yet remains almost perfectly balanced over the underlying candidate identities because mirrored-order evaluation presents each candidate first half the time. This is not random noise. It is a serialization pathology that breaks criterion-conditioned pairwise evaluation.

Representative failure (BioRubric / Llama / standard). In the family for *Loretta Graziano Breuning*, both biographies are factual and share the same occupation: one says she worked in *happiness*, the other in *management*. The base rubric rewards *happiness* more heavily; the counterfactual rewards *management*. Llama disagrees across candidate orders on all four logical variants and ends up predicting the wrong winner on both the base and counterfactual variants. This is criterion-conditioned invalidity in a non-QA setting, mediated by path dependence rather than by a simple factual conflict.

This example also clarifies why the paper remains stronger as a validity paper than as a narrow "belief override" paper. Clean direct semantic failures do occur: QA-Key provides them, and some BioRubric families do as well. But the largest rubric-based failures are often failures of *stable criterion implementation*. The judge is not simply choosing its own preferred answer over the criterion; it is failing to realize the same criterion-conditioned comparison consistently across equivalent serializations.

7 Discussion and Limitations

The main lesson is that criterion fidelity is a reusable validity axis for LLM judges. A judge can look accurate on the base item while failing the stronger requirement that it respond correctly to criterion changes and remain stable under criterion-preserving rewordings. In that sense, criterion fidelity is not another accuracy metric. It asks whether the judge is implementing the intended evaluation function.

The paper is also clearest about what it *does not* show. The strongest completed evidence does not support the broad claim that LLM judges mostly ignore criteria in favor of their own beliefs. That narrower mechanism appears in some QA-Key cases and in some audited BioRubric failures, but it does not dominate the large-gap rubric slices. The more honest and stronger statement is that ordinary judge quality can hide criterion-conditioned instability, and that in close rubric-conditioned pairwise settings this instability is often presentation-sensitive.

This reframing makes the BioRubric result more interesting, not less. Position bias and serialization effects in LLM judges have been noted before (Zheng et al., 2023; Shi et al., 2024; Xu et al., 2026). Our result is different in two ways. First, the pathology is measured against a criterion-conditioned validity object rather than against generic agreement alone. Second, the pathology appears in a setting where both candidates are true and the criterion should be the decisive signal. In that setting, path dependence is not a nuisance confound on top of the main result. It is one of the main ways criterion fidelity fails.

The strict-prompt results reinforce this point. Extra criterion emphasis helps only when the base decision problem is already high margin, and it can make the low-margin rubric task worse. On Llama-3.1-8B-Instruct/BioRubric, the strict prompt seems to sharpen the wrong comparator: the model becomes more serialization-sensitive and less criterion-faithful at the same time. "Reminding" the model to follow the criterion is not equivalent to making the criterion causally control the decision.

Our study has several limits. It uses controlled tasks rather than naturally occurring evaluation pipelines; the models are two open instruct models rather than the full judge ecosystem; and the paper stops short of proposing a mainline mitigation. These are deliberate choices. The goal is not a new benchmark suite or a benchmark-sprawl survey, but a clean validity test. The controlled families let us keep the task and candidates fixed while changing only the criterion, and that is exactly what makes the resulting failures interpretable.

8 Conclusion

We introduced criterion fidelity as a family-level validity axis for LLM judges and evaluated it on two controlled task families. Ordinary judge quality and criterion fidelity dissociate even in a high-margin QA setting and diverge sharply in a non-QA rubric-conditioned family. The resulting failures are not captured by ordinary base accuracy and are not reliably repaired by stronger criterion-

emphasis prompts. In the hardest rubric-based slices, criterion-fidelity failure is often mediated by order-sensitive path dependence rather than by a clean majority of direct semantic criterion refusal. For judge evaluation, that is the main takeaway: accuracy alone can hide whether a model is actually implementing the criterion.

References

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Dongryeol Lee, Yerin Hwang, Taegwan Kang, Minwoo Lee, Younhyung Chae, and Kyomin Jung. 2026. Judging against the reference: Uncovering knowledge-driven failures in LLM-judges on QA evaluation. *arXiv preprint arXiv:2601.07506*.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic study of position bias in LLM-as-a-judge. *arXiv preprint arXiv:2406.07791*.

Yuzheng Xu, Tosho Hirasawa, Tadashi Kozuno, and Yoshitaka Ushiku. 2026. Am I more pointwise or pairwise? Revealing position bias in rubric-based LLM-as-a-judge. *arXiv preprint arXiv:2602.02219*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

A Additional Metrics and Diagnostics

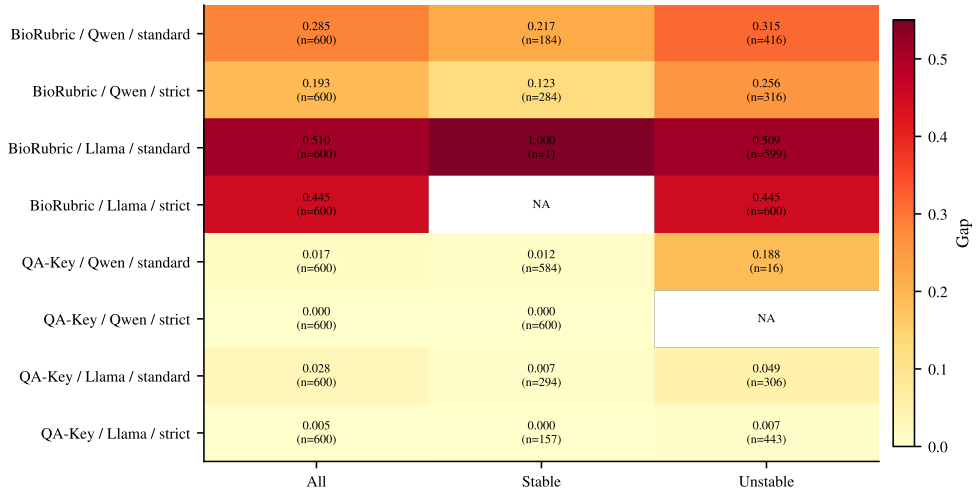

Figure 5: Gap decomposition by order stability. On Qwen2.5-14B-Instruct/BioRubric, a substantial residual gap remains even after restricting to order-stable families. On Llama-3.1-8B-Instruct/BioRubric, almost every family is order-unstable, so path dependence dominates the slice.

Table 3: Strict-prompt deltas relative to the standard prompt. Positive values indicate an increase under the strict prompt.

Task	Model	ΔBase	ΔGCF	ΔGap	ΔSens	ΔInv	ΔOrder
QA-Key	Qwen-14B	0.000	0.017	−0.017	0.015	0.002	−0.007
QA-Key	Llama-8B	−0.020	0.003	−0.023	0.023	0.000	0.197
BioRubric	Qwen-14B	0.080	0.172	−0.092	0.109	0.031	−0.089
BioRubric	Llama-8B	−0.150	−0.085	−0.065	−0.030	−0.099	0.054

Table 4: Failure-audit summary for the principal slices. The audit taxonomy gives `order_effect` precedence over `genuine_criterion_failure`; percentages therefore describe the observed failure channel rather than an exhaustive partition of all conceptual criterion failures.

Slice	n_{aud}	Genuine criterion failure	Order effect	Model tie	Other
QA-Key / Qwen / standard	10	70%	20%	10%	0%
QA-Key / Llama / standard	17	24%	76%	0%	0%
BioRubric / Qwen / standard	50	30%	68%	2%	0%
BioRubric / Llama / standard	50	6%	72%	22%	0%
BioRubric / Qwen / strict	50	34%	64%	2%	0%
BioRubric / Llama / strict	50	0%	72%	28%	0%

Table 5: Variant-level order disagreement and position preference on the BioRubric slices. The strict-prompt Llama slice is nearly a pure first-position chooser while remaining almost balanced over the underlying candidate identities.

Slice	Base	CF	Para 1	Para 2	First-pos	Second-pos	Tie	Choose c_1	Choose c_2
BioRubric / Qwen / standard	0.293	0.388	0.158	0.327	0.612	0.388	0.0002	0.632	0.368
BioRubric / Llama / standard	0.945	0.958	0.932	0.945	0.922	0.066	0.0117	0.515	0.473
BioRubric / Qwen / strict	0.218	0.247	0.145	0.202	0.506	0.493	0.0006	0.672	0.327
BioRubric / Llama / strict	1.000	0.998	0.997	0.995	0.994	0.0046	0.0010	0.500	0.499

B Representative Examples

Clean direct criterion failure (QA-Key / Qwen / standard). For the question "What currency is used in Sao Tome and Principe?", the candidates are "The answer is Sao Tome and Principe dobra" and "The answer is Eastern Caribbean dollar." Under the base criterion, the gold winner is the dobra candidate; under the counterfactual criterion, the gold winner flips to the Eastern Caribbean dollar. Qwen predicts the dobra candidate on all four variants with no order disagreement or ties. This is a clean direct semantic criterion failure.

Strict-prompt serialization pathology (BioRubric / Llama / strict). For the Claude Sautet family, both biographies are factual and differ only in the cited work. Under the strict prompt, all four logical variants are order-disagreeing, the base variant collapses to a tie, and the slice-level statistics show a first-position selection rate of 0.994 while the underlying candidate identity remains nearly balanced. This is the clearest failure case for the claim that stronger criterion emphasis is not equivalent to stronger criterion control.