Diagnosing the Reliability of LLM-as-a-Judge via Item Response Theory

Junhyuk Choi[1*], Sohhyung Park[2*], Chanhee Cho[1], Hyeonchu Park[1], Bugeun Kim[1]

[1]*Department of Artificial Intelligence, Chung-Ang University, Seoul, Republic of Korea*

[2]*Department of Industrial Engineering, Seoul National University, Seoul, Republic of Korea*

ICML

International Conference
On Machine Learning

Background & Motivation

LLM-as-a-Judge

- Rapidly adopted across NLP, vision-language, and RLHF reward modeling.
- Scalable and cost-effective, requiring no human annotation

Limitations of Existing Validation

- **Existing practices operate at the level of observed outputs**:
  - **Intrinsic consistency**: inter-rater agreement, internal consistency …
  - **Human alignment**: correlation, Cohen's κ, Krippendorff's α .

Research Gap in Current Validation

- Validation targets **the output score.**
  - inter-rater agreement, human correlation, uncertainty …
- The judge itself is never tested: **a stable instrument?**
  - Consistent across equivalent prompts?
  - Variance: signal vs. measurement error?

Background & Motivation

How can we trust the judgement of LLM-as-a-Judge?

Background & Motivation

How can we trust the judgement of LLM-as-a-Judge?

Separate the instrument from the judgement

Why Item Response Theory?

- **Item Response Theory (IRT):** psychometrics' tool for measuring **latent ability in human testing**.
  - grades exams: recover a student's latent ability from noisy responses
  - The **Graded Response Model (GRM)** of IRT handles **ordinal ratings** — directly applicable to LLM-as-a-Judge.
  - Most judges return **ordered scores** (Likert-style), so we adopt **IRT's GRM**.

$$P\left(Y_{pj} \geq k \mid \theta_j\right) = \frac{1}{1 + \exp\left(-\alpha_p\left(\theta_j - \beta_{pk}\right)\right)}$$

θ_j	Latent quality of evaluation sample j
α_p	Discrimination of prompt variant p
β_{pk}	Ordered thresholds for score transitions of prompt p at category k

Measuring quality: from ratings to θ

An observed rating is the visible shadow of a continuous latent quality θ.

The GRM recovers it in 3 steps:

① **Ordered thresholds:** A rating reaches "k or higher" once θ clears threshold β_{pk}.

② **Logistic link:** Cumulative probability is logistic with slope α_p:

$$P\left(Y_{pj} \geq k \mid \theta_j\right) = \frac{1}{1 + \exp\left(-\alpha_p\left(\theta_j - \beta_{pk}\right)\right)}$$

③ **Fit by posterior:** NUTS gives every sample a posterior mean $\theta_{\hat{j}}$ and uncertainty σ_j^2.

Fit under **three semantics-preserving variants**: typo, newline, paraphrase. Same θ, own (α, β).

Why θ is the right unit to compare

Comparing judges needs one shared quantity. Instrument parameters don't qualify:

- Judges use the **rating range differently** (one {1-5}, one {3-5}).
- So (α, β) fit on **different thresholds.**

Latent quality θ is a property of the **sample**, not the instrument:

- By Eq. (1), every rating shares the **same θ**.
- **θ distribution** is comparable across judges, prompts, models.

θ is the common scale. Every metric builds on it

What the Framework Measures

Phase 1 Intrinsic Consistency

Is the judge a stable instrument?

- **Prompt consistency:** $C_V = \sigma_{\bar{V}} / \mu_{\bar{V}}$
  - How much do scores consistence under minimal variances (typo / newline / paraphrase)?
  - Threshold: $C_V < 0.10$: stable.
- **Marginal reliability:** $\rho = \mathrm{Var}(\hat{\theta}) / \left[\mathrm{Var}(\hat{\theta}) + E[\sigma^2]\right]$
  - How much of θ is real signal, not estimation noise?
  - Threshold: $\rho \geq 0.70$: reliable

What the Framework Measures

Phase 2 Human Alignment

Does its scale match humans?

- **Discrimination breadth:** $\boldsymbol{\theta_{ratio}} = \mathrm{IQR}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{LLM}})/\mathrm{IQR}(\widehat{\boldsymbol{\theta}}_{\mathrm{Hum}})$
  - How widely does it spread quality, compared to humans?
  - $\theta_{ratio} \approx 1$ matched, $\theta_{ratio} \gg 1$ over-spreads.
- **Distributional alignment:** $\boldsymbol{D_W} = \boldsymbol{W_1}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{LLM}}, \widehat{\boldsymbol{\theta}}_{\mathrm{Hum}})$
  - How far is the full θ distribution from humans?
  - smaller D_w means closer distribution

Experimental Setup

Benchmarks: 4 judge methods, NLP + vision tasks

- NLP Task: SummEval, TopicalChat and HelpSteer-2
- Vision Task: VIEScore (CIG/TIE/MIE)

Judge models: 7 frontier LLMs (VLMs)

- Gemini-2.5-Flash, GPT-4o, GPT-4o-mini, Qwen3-30B, Qwen3-235B, Llama-4-Maverick and Llama-4-Scout
- Vision tasks use the VL variants (Qwen3-VL).

Phase 1 : Most Judges Fail Somewhere

Benchmark / Metric		Gemini-2.5	GPT-4o	GPT-4o-mini	Qwen3-30b	Qwen3-235b	Llama-4-M	Llama-4-S
[NLP] SummEval Consistency	C_V	0.07	0.05	0.92	0.15	0.08	0.25	0.07
	ρ	0.55	0.91	0.88	0.60	0.66	0.63	0.78
TopicalChat Understand.	C_V	0.18	0.16	0.20	0.03	0.27	0.21	0.48
	ρ	0.39	0.48	0.49	0.53	0.34	0.46	0.43
HelpSteer-2 Complexity	C_V	1.08	0.16	0.30	0.39	0.32	0.25	0.31
	ρ	0.87	0.89	0.81	0.88	0.89	0.77	0.85
[Vision] VIEScore CIG-PQ	C_V	1.11	0.30	0.82	0.62	0.47	0.17	0.46
	ρ	0.94	0.96	0.93	0.94	0.94	0.92	0.83

NLP task stable; Vision tasks far more prompt-sensitive

- NLP tasks : $C_V < 0.30$; Vision tasks 0.16–1.32 (Gemini-2.5 > 1.0 on every vision task).
- ρ uneven even on NLP; few judges pass both thresholds.

Phase 2 : Same Signal, Different Causes

Benchmark / Metric (θ_{ratio})	Gemini-2.5	GPT-4o	GPT-4o-mini	Qwen3-30b	Qwen3-235b	Llama-4-M	Llama-4-S
[NLP] SummEval Coherence	1.24	1.50	1.66	1.53	1.55	0.98	1.30
TopicalChat Engaging	2.21	3.07	2.16	2.08	2.21	2.05	2.75
HelpSteer-2 Helpfulness	1.80	1.83	1.76	1.61	1.58	1.86	1.65
[Vision] VIEScore CIG-PQ	3.35	3.03	3.11	3.09	2.61	2.76	2.03
TIE-PQ	3.86	4.40	4.14	3.75	3.92	3.47	3.91
MIE-SC	1.87	2.42	1.96	2.15	1.79	2.00	1.60

Same signal ($\theta_{ratio} > 1$), two causes.

- **NLP tasks** : calibration mismatch (r = 0.2-0.5). Fix by rescaling.
- **Vision tasks** : validity gap (r ≈ 0). Rescaling won't help.

Why and How?

Phases 1-2 show where judges fail

Phase 1: few judges pass both C_V and ρ.

Phase 2: calibration gap persists.

Ablation: what fixes their reliability?

Toggle rubric detail, CoT, rating scale.

Re-measure C_V and ρ (mean of 7 judges).

Design Choices That Improve Reliability

Ablation results for TopicalChat. Mean C_V and ρ across all 7 judges. CoT lowers C_V; richer scales raise ρ.

Condition	Gemini-2.5		GPT-4o		GPT-4o-mini		Qwen3-30B		Qwen3-235B		Llama-4-M		Llama-4-S	
	C_V	ρ	C_V	ρ	C_V	ρ	C_V	ρ	C_V	ρ	C_V	ρ	C_V	ρ
Original Prompt	0.21	0.71	0.23	0.66	0.16	0.72	0.15	0.77	0.23	0.73	0.25	0.68	0.29	0.72
Detail rubric	0.12	0.69	0.12	0.68	0.15	0.69	0.12	0.78	0.19	0.71	0.16	0.74	0.28	0.64
CoT	0.14	0.74	0.11	0.69	0.33	0.71	0.16	0.76	0.16	0.72	0.06	0.76	0.24	0.70
Likert 5	0.18	0.92	0.14	0.92	0.18	0.93	0.13	0.93	0.34	0.94	0.19	0.93	0.22	0.92
Likert 7	0.30	0.92	0.21	0.95	0.26	0.93	0.26	0.94	0.26	0.95	0.26	0.90	0.21	0.92

Reliability is an engineering choice.

→ High C_V? Add a detailed rubric + chain-of-thought → **more stable.**

→ Low ρ? Adjust the rating scale → **more reliable.**

Practical Recommendations & Takeaways

If your judge fails Phase 1 (Intrinsic consistency)

- High $C_V \rightarrow$ add detailed rubric + CoT.
- Low $\rho \rightarrow$ adjust rating scale (Likert 5/7).

If your judge fails Phase 2 (Human alignment)

- NLP task, calibration mismatch: rescale outputs.
- Vision task, validity gap: rescaling won't help; different construct.

Our IRT-based framework turns 'can we trust this judge?' into an actionable diagnostic.

Revealing where judges fail and what to do about it.

Contributors





Junhyuk Choi* Sohhyung Park* Chanhee Cho Hyeonchu Park Bugeun Kim

* Equal Contribution

Question: chlwnsgur129@cau.ac.kr & sohhyung@bdai.snu.ac.kr

Paper link: https://arxiv.org/abs/2602.00521

Code: https://github.com/elu-lab/ICML2026-Diagnosing-the-Reliability-of-LLM-as-a-Judge-via-Item-Response-Theory