The Cultural Lens of LLM Judges: Cultural Specificity and Judge-Dependent Rankings in Chinese LLM Evaluation

Zengmingyu He
CS321M: AI Measurement Science
Stanford University
zachx@stanford.edu

Abstract

LLM-as-judge is widely used for open-ended language model evaluation, but judge models are themselves trained systems with their own language distributions, alignment preferences, and cultural priors. This project studies whether LLM-as-judge evaluation is stable for Chinese open-ended tasks, and whether judge-dependent variance increases as tasks become more culturally specific. We constructed 27 author-designed Chinese prompts across three levels of cultural specificity and evaluated responses from U.S.-origin and China-origin assistants[1] using blind and labeled LLM judging. Within this author-designed prompt set, 16 especially difficult objective Chinese internet-culture subtasks were additionally paired with author-verified answer keys. The main outcome is not which model is "best," but whether scores and rankings are robust to judge choice. We treat LLM judges as noisy measurement instruments and analyze judge agreement, ranking sensitivity, same-origin score differences, blind-vs-labeled effects, ground-truth fidelity, and variance components.

1 Overview and Motivation

Open-ended LLM evaluation increasingly relies on other LLMs as judges, especially for tasks without a single exact answer, such as writing, dialogue, cultural explanation, and instruction following. However, a judge model is not a neutral oracle. Its score may reflect answer quality, but it may also reflect judge-specific preferences, training distribution, rubric interpretation, language context, or interaction with the answer model.

Chinese open-ended evaluation is a useful stress test for this problem because task quality often depends on native expression, platform norms, and internet discourse. Some relevant signals live in semi-closed, platform-specific ecosystems such as WeChat, Xiaohongshu, Bilibili, and Zhihu, rather than in a single globally indexed open web. This creates a measurement concern: different judges may assign different scores not because the answer quality changed, but because their exposure to Chinese internet context differs. The goal is not to assume that China-origin assistants understand Chinese contexts better, but to test whether evaluation results depend on judge choice.

[1]Throughout this project, we use origin as a convenient shorthand for a coarse metadata proxy. It does not refer to corporate registration jurisdiction, legal domicile, nationality, model language, or intrinsic model capability. Instead, it summarizes public information about a model's developer/product ecosystem, primary deployment context, target user base, language environment, and alignment setting. The variable is useful only as it helps organize possible judge-dependent measurement variation; it should not be interpreted causally.

Central question. Is LLM-as-judge evaluation culture-neutral for Chinese open-ended tasks, or does cultural specificity amplify judge-dependent variance?

2 Related Work and Gap

LLM-as-judge methods are widely used to evaluate open-ended model outputs because they are cheaper and more scalable than human evaluation. MT-Bench and Chatbot Arena are representative examples of earlier work using strong LLMs and human preferences to evaluate conversational models [Zheng et al., 2023, Chiang et al., 2024]. We cite them as methodological background, not as benchmarks that contain or validate the later model versions used in this project. Recent work also shows that LLM judges can suffer from reliability issues, judge bias, position effects, verbosity preference, and other artifacts [Thakur et al., 2024, Gu et al., 2024]. This motivates treating LLM judges as noisy measurement instruments rather than ground-truth annotators.

Chinese LLM evaluation has also grown rapidly. C-Eval evaluates Chinese foundation models across academic subjects [Huang et al., 2023], CMMLU provides a Chinese multitask language understanding benchmark [Li et al., 2024], and AlignBench evaluates Chinese alignment using multi-dimensional LLM-as-judge protocols [Liu et al., 2024]. These works establish useful background for Chinese-language evaluation, but this project does not treat them as historical leaderboards for the specific 2026-era API models studied here. Less is known about whether Chinese cultural specificity amplifies judge-dependent ranking instability across models from different broad origins. This project addresses that gap by varying task cultural specificity and comparing blind versus labeled judging.

3 Methods

Task set. We constructed 27 author-designed Chinese prompts across three levels of cultural specificity (10 prompts for Level 0, 10 prompts for Level 1, and 7 prompts for Level 2). The substantive prompt ideas were designed and checked by the author; AI assistance was used only for formatting, organization, and wording polish. Within the broader author-designed task set, the objective internet-culture subset required additional manual curation: the author collected candidate items through public internet search and informal, non-identifying online consultation, then wrote and checked explicit answer keys. While we originally planned to construct exactly 10 prompts per level, Level 2 contains 7 prompts because several Level 2 prompts are designed as multi-item evaluation tasks containing multiple sub-questions or test cases (e.g., evaluating up to 10 internet slang items within a single prompt, as shown in Appendix E). This allowed us to cover a wider breadth of Chinese Internet Context items while maintaining a balanced total of 27 prompt blocks. To make the classification intuitive, we highlight representative tasks for each level (with full details provided in Appendix E):

- **Level 0 (Culture-Neutral)**: Explanations, reasoning, information extraction, planning, or coding tasks that do not depend on region-specific or platform-specific knowledge. A representative task asks the model to compare retrieval-augmented generation (RAG) and instruction fine-tuning for reducing hallucination. Another requires implementing a thread-safe sliding-window API rate limiter under memory constraints.

- **Level 1 (Chinese Expression)**: Language-polishing and communication tasks where the core construct is natural written expression in standard modern Chinese. For example, one prompt presents a translationese-heavy sentence ("我被给了一个极好的机会去展示我的技能在昨天的会议上"[2]) and requests a natural, idiomatic rewriting. Another task asks the model to draft a polite WeChat message requesting coffee and career advice from a senior colleague.

[2]"我被给了一个极好的机会去展示我的技能在昨天的会议上", *wǒ bèi gěi le yī gè jí hǎo de jī huì qù zhǎn shì wǒ de jì néng zài zuó tiān de huì yì shàng*: "I was given an excellent opportunity to show my skills in yesterday's meeting."

- **Level 2 (Chinese Internet Context)**: Platform-specific internet subculture tasks requiring familiarity with Chinese social media tones, slang, or ironic subtexts. One task asks models to infer the non-literal meaning of comment-section phrases such as "遥遥领先" (Far Ahead[3]). Another asks models to write two openings on the same topic, one recognizably suited to Zhihu (知乎[4]) and one to Xiaohongshu (小红书[5]), without explicitly naming the platform style.

Models and evaluation matrix. We evaluated six open-ended models as candidates, generating answers to all 27 prompts. The model set consists of three U.S.-origin assistants (GPT-5.4, Claude-Sonnet-4.6, and Gemini-3-flash-preview) and three China-origin assistants (Qwen-3.7-max, DeepSeek-v4-pro, and GLM-5.1). To measure inter-rater reliability and origin-dependent rankings, the exact same six models were used as judges.

Each candidate response was evaluated under two judging conditions:

- **Blind mode**: The judge model evaluates the prompt and response without any candidate brand identity or origin metadata.
- **Labeled mode**: The judge model is explicitly shown the candidate's model name, developer name, and origin category.

Each judge scored the responses on a scale of 1 to 10 across six dimensions: correctness, helpfulness, reasoning, Chinese naturalness, cultural fit, and overall quality. In total, the planned evaluation matrix comprises 27 prompts $\times$ 6 candidate models $\times$ 6 judge models $\times$ 2 modes $= 1,944$ individual judgments.

Measurement model formulation. Following classical measurement theory, we treat the observed score $s_{p,a,j,m,d}$ assigned to prompt p, candidate a, by judge j in mode m on rubric dimension d as a combination of latent candidate quality $\theta_{p,a,d}$, judge severity/leniency $\eta_{j,d}$, and interaction/noise effects. To isolate the specific impact of shared geographic/cultural origin, we define:

$$\text{same\_origin}_{a,j} = \begin{cases} 1 & \text{if candidate } a \text{ and judge } j \text{ belong to the same regional origin group,} \\ 0 & \text{otherwise.} \end{cases}$$

To control for potential confounding variables (such as an answer model being inherently superior or a judge being systematically lenient), we estimate the same-origin effect using an Ordinary Least Squares (OLS) fixed-effect regression model:

$$s_{p,a,j,m,d} = \beta_0 + \beta_1 \cdot \text{same\_origin}_{a,j} + \gamma_p + \delta_a + \theta_j + \lambda_m + \epsilon$$

where γ_p, δ_a, θ_j, and λ_m represent categorical fixed effects for the prompt, candidate model, judge model, and judging mode, respectively. The coefficient β_1 represents the adjusted same-origin effect, capturing the net score difference attributable solely to shared origin.

To gain statistical intuition for β_1, consider a simplified model with only origin-level effects. Let $\bar{s}_{J_{\text{origin}},A_{\text{origin}}}$ denote the average score assigned by judges from origin $J_{\text{origin}} \in \{\text{US}, \text{CN}\}$ to candidate models from origin $A_{\text{origin}} \in \{\text{US}, \text{CN}\}$. Under a balanced design, the OLS coefficient β_1 is mathematically equivalent to:

$$\beta_1 = \frac{1}{2}\left[(\bar{s}_{\text{CN,CN}} - \bar{s}_{\text{CN,US}}) + (\bar{s}_{\text{US,US}} - \bar{s}_{\text{US,CN}})\right]$$

This reveals that β_1 is the average of two relative preference terms: (1) how much Chinese judges prefer Chinese models over U.S. models ($\bar{s}_{\text{CN,CN}} - \bar{s}_{\text{CN,US}}$), and (2) how much U.S. judges prefer U.S. models over Chinese models ($\bar{s}_{\text{US,US}} - \bar{s}_{\text{US,CN}}$). This controls for both general candidate quality and judge leniency.

[3]"遥遥领先", *yáoyáo lǐngxiān*: "Far Ahead", a popular phrase that can be literal praise or ironic meme usage depending on context.

[4]知乎, *Zhīhū*: a Chinese question-answering and discussion platform.

[5]小红书, *xiǎohóngshū*: Little Red Book, a popular lifestyle sharing and social platform in China.

4 Results

Data Collection and Descriptive Statistics. We executed 162 candidate answer generation queries (6 models × 27 prompts). The planned judging matrix contained 1,944 scoring slots (162 answers × 6 judges × 2 modes), and the final SQLite database retained 1,935 unique completed judgment combinations. The nine missing slots were left as missing rather than imputed. They were mainly caused by provider-side keyword and content-safety review in China-origin model routes on Chinese Internet Context prompts. A representative example is Prompt 27, where one item contains the phrase "我练功发自真心"; the token "练功" appeared to trigger keyword review for some Chinese AI routes, causing the API call to return no usable rubric JSON. All quantitative results below therefore use the SQLite database as the source of truth and analyze the available completed judgments.

The global average overall score assigned across all completed evaluations is 8.39 out of 10. Under blind judging conditions, the general performance levels vary significantly by model. In terms of average overall score across all judges and prompts, GPT-5.4 ranked highest with a blind-mode mean of 9.19, followed by Gemini-3 (8.97). Qwen-3.7-max and GLM-5.1 were tied at 8.39, followed by Claude-Sonnet-4.6 (7.89) and DeepSeek-v4-pro (7.49). This basic ranking serves as our baseline descriptive performance indicator before analyzing inter-judge reliability and bias. Table 1 provides the detailed comparison of model rankings and mean scores between the blind and labeled modes, showing the individual model-level shifts.

Table 1: Model performance comparison under blind vs. labeled judging modes (mean overall scores).

Model Name	**Blind Mode Mean**	**Labeled Mode Mean**	**Delta (Δ_{label})**
openai/gpt-5.4	9.19	9.14	−0.05
google/gemini-3-flash-preview	8.97	8.99	+0.02
z-ai/glm-5.1	8.39	8.53	+0.14
qwen/qwen3.7-max	8.39	8.42	+0.03
anthropic/claude-sonnet-4.6	7.89	7.80	−0.09
deepseek/deepseek-v4-pro	7.49	7.48	−0.02

Inter-Judge Reliability (ICC). We assess the consistency of the judge models using the Intraclass Correlation Coefficient (ICC(3,1)) under the two-way mixed-effects model. Table 2 reports ICC after pooling blind and labeled judgments within each task level: 0.62 on Level 0, 0.55 on Level 1, and 0.60 on Level 2. Splitting by mode shows a similar but not identical pattern. In blind mode, inter-judge agreement is 0.63 on Level 0, 0.52 on Level 1, and 0.63 on Level 2. In labeled mode, the corresponding ICCs are 0.62, 0.58, and 0.59. Thus agreement does not decline monotonically with cultural specificity, but judge consistency is clearly sensitive to task type and label disclosure, especially for expression-level and Chinese Internet Context prompts.

Table 2: Measurement summary by task level.

Metric	**L0**	**L1**	**L2**
ICC (pooled)	0.62	0.55	0.60
Adjusted same-origin	+0.12	−0.12	+0.03
Raw same-origin gap	+0.12	−0.12	+0.05

Note. L0 = neutral tasks; L1 = Chinese expression tasks; L2 = Chinese Internet Context tasks. ICC is pooled across blind and labeled judgments.

Same-Origin Bias and Label Effects. We compute the raw and adjusted same-origin score differences to identify preference patterns. As shown in Table 2, the adjusted same-origin effect β_1 over both judging modes is +0.12 for Level 0, −0.12 for Level 1, and +0.03 for Level 2. These values are small and mixed, so the main result is not a simple same-origin favoritism story. Instead, shared origin effects depend on the task level and on whether the candidate identity is hidden or revealed.

The mode-specific estimates show the labeling mechanism more clearly. In blind mode, the adjusted same-origin effect is +0.02 on Level 0, −0.12 on Level 1, and −0.08 on Level 2. In labeled mode, it becomes +0.21, −0.12, and +0.14, respectively. The global labeled-minus-blind average is only +0.016 points, but labels still redistribute scores across model-origin pairings. In particular, Level 2 moves from a small blind same-origin penalty to a positive labeled same-origin coefficient.

The Level 2 origin means illustrate this shift. In blind mode, China-origin judges score China-origin candidates at 7.32 and U.S.-origin candidates at 8.36, while U.S.-origin judges score China-origin candidates at 6.83 and U.S.-origin candidates at 7.75. This yields a small negative same-origin contrast. In labeled mode, the same comparison shifts to 7.46 versus 8.24 for China-origin judges and 6.59 versus 7.65 for U.S.-origin judges, producing a positive same-origin contrast after controls. To visualize the detailed pairwise preferences between judge models and candidate models, Table 3 displays the full candidate-by-judge average rating matrix under blind judging.

Table 3: Blind mode mean overall score matrix (Rows: Candidate Models, Columns: Judge Models).

Candidate / Judge	GPT-5.4	Claude-4.6	Gemini-3	Qwen-3.7	DeepSeek-v4	GLM-5.1
GPT-5.4	8.56	8.81	9.52	9.15	9.56	9.58
Claude-4.6	7.52	8.73	7.67	6.88	8.63	7.89
Gemini-3	7.85	8.70	9.85	8.70	9.22	9.48
Qwen-3.7	7.52	8.70	8.63	8.42	8.67	8.41
DeepSeek-v4	7.00	8.48	7.67	6.30	8.59	6.88
GLM-5.1	7.52	8.52	8.26	8.38	8.89	8.78

Note. Cell values are rounded judge-specific means. Table 1 is computed directly from all available judgment rows, so it should not be recomputed from the rounded cells in this matrix.

Identity Bias Shift. Comparing matching prompt-candidate-judge triples between labeled and blind modes, we measure the overall brand identity shift $\Delta_{\text{label}} = s_{\text{labeled}} - s_{\text{blind}}$. Across all evaluations, revealing model identity resulted in a small average overall score shift of +0.016 points. This implies that while the average global score is relatively stable, the revelation of names redistributes scores across specific developer ecosystems rather than creating a large uniform score inflation.

Variance Decomposition. To partition the sources of score variation, we approximate a fixed-effect Analysis of Variance (ANOVA) decomposition. As summarized in Table 4, across all data combined, the largest portion of explainable variance is attributed to the Task Prompt Construct (29.5%), while the candidate model accounts for 7.5% and the judge model explains 4.5%.

When we partition the variance decomposition by task specificity level, the same general pattern appears with important differences across levels:

- **Level 0 (Neutral)**: The Candidate Answer Model explains 15.5% of the total score variance, while the Judge Model accounts for only 2.8%. This indicates that for culture-neutral tasks, candidate differences are more visible than judge-specific differences.
- **Level 1 (Expression)**: The Candidate Model's share drops to 6.0%, while the Judge Model's share increases to 7.0%, reflecting growing subjectivity in evaluating Chinese linguistic naturalness.
- **Level 2 (Chinese Internet Context)**: The Judge Model's share remains high at 7.5%, while the specific Task Prompt Construct explains 38.1% of the variance.

This decomposition suggests that Chinese Internet Context evaluation is strongly shaped by the prompt itself and by judge-specific measurement behavior. The residual variance components remain high across all levels, pointing to substantial interactions between candidate models, judge models, and task specificity.

Table 4: Generalizability Theory (G-Theory) observed score variance attribution decomposition.

Measurement Design Facet	Variance Proportion (%)
Candidate Answer Model	7.5%
Evaluation Judge Instrument	4.5%
Task Prompt Construct	29.5%
Blinding / Labeled Condition	0.0%
Residuals / Interactions	58.5%

Fidelity on Objective Subtasks. To establish the construct validity of the judge models, we evaluated their ratings against standard ground-truth answers on three Level 2 objective subtasks (Prompts 21, 22, and 27, containing 16 factual sub-questions about Chinese internet slang and homophones).

On these subtasks, candidate model accuracy varied widely, as shown in Table 5. Gemini-3 and Qwen-3.7-max tied for the best performance, each answering 9 of the 16 keyed subquestions correctly (56.25%). However, the absolute accuracy was low across the entire model set: no model exceeded 60%, and all six models failed all three source-identification items in Prompt 22. This supports the qualitative observation that these Chinese internet-culture items are difficult for current models, even when some models perform better than others.

Table 5: Candidate model accuracy on the 16 keyed Level 2 objective subtasks.

Candidate Model	Correct / 16	Accuracy
google/gemini-3-flash-preview	9	56.25%
qwen/qwen3.7-max	9	56.25%
deepseek/deepseek-v4-pro	7	43.75%
anthropic/claude-sonnet-4.6	6	37.50%
z-ai/glm-5.1	3	18.75%
openai/gpt-5.4	2	12.50%

More importantly, as shown in Table 6, the judge models exhibited massive variations in their ability to detect incorrect answers (evaluation fidelity) in blind mode. Claude-Sonnet-4.6 was the most reliable judge, exhibiting a 0.63 correlation with the ground-truth accuracy and a 0.0% False Positive Rate (never awarding a score ≥ 7 to a wrong answer). In contrast, DeepSeek-v4-pro exhibited a very high False Positive Rate of 66.7% (scoring 7.86 on average for wrong answers), indicating high leniency and a failure to identify cultural factual errors. This demonstrates that LLM judges differ not only in their biases, but also in their fundamental capability to validly measure regional cultural constructs.

Table 6: LLM Judge evaluation validity (fidelity) on Level 2 objective subtasks (Blind Mode).

Judge	Corr.	FPR	Score if Correct	Score if Wrong
Claude-4.6	0.63	0.0%	7.50	4.88
DeepSeek-v4	0.16	66.7%	9.00	7.86
Gemini-3	0.52	22.2%	7.50	4.57
GPT-5.4	0.15	11.1%	4.00	3.43
Qwen-3.7	0.07	25.0%	2.50	4.29
GLM-5.1	0.15	37.5%	7.00	5.43

5 Discussion

The results are more mixed than a simple origin-based hypothesis would predict. We do not find that China-origin models are systematically stronger on Chinese Internet Context tasks,

nor do we find a stable pattern in which judges favor same-origin candidates. Inter-judge agreement also does not decline monotonically from Level 0 to Level 2. In this dataset, model origin is therefore a weak proxy for performance on Chinese Internet Context tasks or for judging bias.

The more robust finding is measurement instability. Judge choice still matters, prompt identity explains a large share of score variation, and revealing candidate identity changes some model-origin pairings even though the average labeled-minus-blind shift is small. This suggests that blinding is still useful, but not because LLM judges mechanically favor models from the same broad origin group. Rather, model identity and judge identity interact in local, task-dependent ways that are not well captured by a binary U.S.-versus-China label.

The keyed objective subset sharpens this point. These items were the most manually curated part of the broader author-designed prompt set, and they were difficult for all candidate models: even the best candidates answered only 9 of 16 items correctly. Gemini-3 and Qwen-3.7-max tied as candidates, while Claude-Sonnet-4.6 and Gemini-3 were the strongest judges against ground truth. Candidate knowledge and judge fidelity are therefore related but not identical constructs. The model that answers cultural questions well is not necessarily the model that best measures whether other answers are correct.

Limitations. Several limitations constrain our findings:

- **Coarse origin labels.** Model origin is a manually assigned metadata category. It aggregates developer location, primary user base, and deployment ecosystem, but it does not isolate specific causal factors.
- **Unequal latent capability.** The evaluated models are not necessarily comparable in latent capability. Their parameter counts, training-data scale, post-training methods, inference-time compute, system prompts, context handling, API routing, and safety filters are mostly proprietary or only partially observable. Some models may therefore differ by capability tier rather than by origin alone, so origin comparisons are confounded with model scale and deployment design.
- **Small task set.** The task set consists of 27 prompts, which is relatively small. A larger prompt bank is required to confirm generalizability.
- **Single decoding setting.** Evaluations were conducted using a single default temperature setting, and variations in decoding parameters could affect both output quality and judge reliability.
- **Truncated candidate answers.** API generation token constraints led to premature answer truncation for several candidates (e.g., Claude-Sonnet-4.6 and GLM-5.1). Judges systematically and heavily penalized incomplete outputs, frequently awarding minimum correctness and overall scores (e.g., overall ratings of 1–3). This introduces construct-irrelevant measurement error, as candidates are penalized for length-limit cutoffs rather than their actual competence.
- **Missing judgment cells.** Nine planned judgment cells were missing from the final database. These missing cells were not randomly distributed: they occurred primarily around China-origin judging routes and Chinese Internet Context prompts, where phrases such as "练功" appeared to activate keyword-sensitive safety filtering. We treat this as part of the practical measurement environment rather than silently filling the missing values, but it means the final matrix is nearly complete rather than perfectly balanced.

6 Conclusion and Future Work

This project conducted a reliability audit of LLM judges on Chinese open-ended tasks under varying levels of cultural specificity. The results do not support a simple cultural-origin advantage hypothesis: China-origin models were not systematically better on Chinese Internet Context tasks, judges did not consistently favor same-origin candidates, and judge

agreement did not track task level in a simple linear way. The main contribution is instead a cautionary one: LLM-as-judge evaluation is noisy, judge-dependent, and difficult to explain with coarse model-origin labels.

Future work should expand this measurement framework to larger and more capability-matched model suites, explore human-LLM agreement on the same specificity scales, and develop statistical calibration techniques for judge-specific measurement variation without assuming that origin alone explains the effect.

References

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL `https://proceedings.mlr.press/v235/chiang24b.html`.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024. URL `https://arxiv.org/abs/2411.15594`.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-Eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, 2023. URL `https://arxiv.org/abs/2305.08322`.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.671. URL `https://aclanthology.org/2024.findings-acl.671/`.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. Alignbench: Benchmarking chinese alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL `https://aclanthology.org/2024.acl-long.624/`.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*, 2024. URL `https://arxiv.org/abs/2406.12624`.

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. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. URL `https://arxiv.org/abs/2306.05685`.

A Code and Data Availability

The project repository will be made available at `https://github.com/realfakezachx/cs321m-llm-judge`. It will contain the experiment code, prompt set, analysis scripts, documentation, and the data artifacts needed to reproduce the reported tables and figures. Private credentials, API keys, and any local environment secrets are excluded.

B Plagiarism, Bias, and Inaccuracy Statement

To reduce plagiarism risk, I will cite all papers, datasets, benchmarks, and codebases that influenced the project, including LLM-as-judge, Chatbot Arena, Chinese LLM benchmarks, and any starter code or external packages used in the implementation. I will not present AI-generated text as a substitute for scholarly attribution: if an idea comes from prior work, I will cite the original source rather than crediting the chatbot. To reduce the risk of textual plagiarism, I will use AI-generated drafts only as editable scaffolding and will revise the final writing in my own words. To reduce bias and inaccuracy, I treat model origin only as a coarse metadata proxy for development and deployment context, not as a causal, national, legal, linguistic, or intrinsic capability category. I will avoid claims that one origin group is inherently better. Any observed same-origin score difference will be interpreted descriptively as evidence about judge-dependent measurement variation, not as evidence of intrinsic model ability.

C AI Use Reflection

I used generative AI tools as learning, programming, and writing-support tools while developing this project. I used them to refine the measurement framing, compare alternative experimental designs, clarify unfamiliar statistical concepts such as judge agreement, interaction effects, and mixed-effects models, and translate vague project ideas into a more concrete pre-analysis plan. The substantive Chinese prompt set and the objective answer keys were designed and checked by me; AI assistance was used to standardize formatting, improve wording clarity, draft code structure, debug LaTeX errors, and improve the organization of my writing. These tools affected my learning by making the iteration cycle faster: when I was uncertain about a concept, I could ask follow-up questions, test whether I understood the idea, and then revise the project design. However, I did not treat AI output as authoritative. I checked claims against papers, course materials, and my own understanding, and I edited the final text myself. In the future, I think AI tools can be valuable as interactive tutors and research assistants, but students should remain responsible for scientific judgment, attribution, and interpretation.

D Impact Statement

This project studies how LLM-as-judge evaluation may vary across language, culture, platform context, and model origin. A positive impact is that it can make AI evaluation more cautious: instead of treating a single judge model as a neutral oracle, evaluators may measure judge agreement, ranking sensitivity, and cultural-context dependence before making leaderboard claims. A possible risk is that results could be misused to make broad claims about national, cultural, or organizational superiority among models. I mitigate this by framing origin as a confounded proxy and avoiding causal interpretations. Another risk is that platform-specific cultural examples could stereotype Chinese internet communities or overfit to narrow subcultures. I mitigate this by using controlled prompt templates, avoiding scraped user comments, and analyzing disagreement rather than asserting one ground truth. The project uses API calls and small-scale generated data, so environmental impact should be limited. Because the study does not involve human participants or private user data, the main ethical obligations are careful attribution, responsible framing, and transparent data handling.

E Representative Prompt Details

This appendix provides representative prompt excerpts and contextual descriptions across the three levels of Chinese cultural specificity used in this study.

E.1 Level 0: Culture-Neutral Tasks

The following prompt evaluates technical explanation and system-design reasoning, independent of regional cultural knowledge:

Prompt excerpt: 请详细解释大语言模型架构中“检索增强生成（RAG）”与“指令微调（Instruction Fine-tuning）”在解决模型幻觉（Hallucination）问题上的本质作用机制差异。此外，如果一个系统需要处理每天都在高频更新的动态专业知识库，在系统架构层面应如何权衡和组合这两种技术?
Translation: Please explain the essential mechanism-level difference between retrieval-augmented generation (RAG) and instruction fine-tuning in addressing hallucination. If a system must handle a dynamic expert knowledge base that updates every day, how should these two techniques be balanced and combined at the system-architecture level?

The following coding task tests implementation and complexity reasoning, which are culture-neutral:

Prompt excerpt: 请使用 Python 实现一个基于“滑动时间窗口（Sliding Time Window）”算法的 API 限流器（Rate Limiter）类。要求必须能够处理高并发场景下的线程安全问题，内存占用应尽可能小，严禁使用列表直接存储每一个独立请求的绝对时间戳。
Translation: Please implement a Python API rate limiter based on the sliding time-window algorithm. It must handle thread safety under high concurrency, minimize memory use, and avoid storing every individual request timestamp in a list.

E.2 Level 1: Chinese Language and Expression Tasks

The following prompt measures natural translationese editing, requiring standard grammatical fluency:

Prompt content: 以下是一句典型的翻译腔（Translationese）句子，请将其重写为自然、地道的现代汉语表述：“我被给了一个极好的机会去展示我的技能在昨天的会议上。”
Translation: The following is a typical translationese sentence, please rewrite it into natural, idiomatic modern Chinese: “I was given an excellent opportunity to show my skills in yesterday’s meeting.”

The following prompt tests polite written Chinese business communication standards:

Prompt content: 你周五下午想请公司一位资深前辈喝咖啡并请教职业发展规划。请拟写一条礼貌、真诚且不卑不亢的微信询问消息。
Translation: You would like to invite a senior colleague at your company for coffee this Friday afternoon to ask for career development advice. Please write a WeChat inquiry message that is polite, sincere, and confident but respectful.

E.3 Level 2: Chinese Internet Context Tasks

The following prompt requires understanding of Chinese comment-section memes and pragmatic subtext:

Prompt excerpt: 在中文互联网的各种评论区中，下面这些短语可能并不是字面意思。请判断评论者实际上想表达的原意。分析时请排除“评论者只是单纯刷热梗”的情况，尽量说明具体语境、语气和潜台词。A. 我早已麻痹 B. 遥遥领先 C. 你就偷着乐吧 D. 我也要评吗
Translation: In Chinese internet comment sections, the following phrases may not mean their literal surface meanings. Please infer what the commenter actually intends to express, excluding cases where the phrase is merely copied as a meme, and explain the concrete context, tone, and subtext.

The following prompt tests platform-sensitive tone differences between Zhihu and Xiaohongshu:

Prompt excerpt: 主题：一个人旅行时感到孤独怎么办？请分别写：1. 一个知乎回答的开头；2. 一个小红书笔记的开头。要求每个 80 字以内，不要直接说“知乎风格是……”或“小红书风格是……”，让读者能从文本本身感到平台差异。
Translation: Topic: What should someone do when feeling lonely while traveling alone? Please write two short openings: one for a Zhihu answer and one for a Xiaohongshu note. Each should be under 80 Chinese characters, and the text should make the platform difference recognizable without explicitly naming the style.

E.4 Representative Objective Internet-Culture Items

Prompts 21, 22, and 27 contain 16 objective subtasks with explicit answer keys. These items were the most manually curated portion of the broader author-designed prompt set: the author collected and filtered them after checking public Chinese internet materials and consulting informally with people familiar with online meme culture. The table below lists representative examples; the quantitative ground-truth analysis used all 16 subtasks.

Table 7: Representative items from the 16 objective internet-culture subtasks.

Item	**Construct**	**Expected answer or context**
这 ___ 就非买不可吗?	Phrase completion	“瓜,” associated with the Liu Huaqiang watermelon-buying scene from *Conquer*.
我全都 ___ 出去了。	Phrase completion	“防,” associated with Ma Baoguo-style martial-arts meme discourse.
“我的刀盾”	Source identification	A phonetic localization of the English meme phrase “what the dog doing.”
“不知道，我的身材很曼妙”	Source identification	A Taobao review/comment meme context rather than a literal self-description.
“一袋米扛几楼”	Phonetic mishearing source	*Naruto*, associated with Pain/Nagato and Chinese phonetic mishearing of Japanese dialogue.
“砸瓦鲁多”	Phonetic mishearing source	*JoJo's Bizarre Adventure*, associated with DIO's “The World.”
“我练功发自真心”	Phonetic mishearing source	A foreign viral video commonly circulated in Chinese meme contexts as a computer-smashing child clip.