The Language Is the Lever: Prompt Language, More Than Training Origin, Shapes How Open LLMs Answer Contested Questions

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi*
Transformer Lab

Abstract

As open language models from many countries enter wide use, understanding what shapes the values they express becomes practically important. We study this with a clean design: how much of a model's stance on contested topics is set by where it was built, and how much by the language it is prompted in? We introduce a controlled bilingual audit of eight size-matched open instruction-tuned models (four built in China, four in the West), each answering the same 145 probe items in English and Chinese. Because every item is asked in both languages, each model is its own control, which cleanly isolates the effect of prompt language. Our central finding is that prompt language is a strong and consistent lever: switching from English to Chinese shifts a model's stance on contested questions by 0.37 points on a five-point scale (95% CI $[0.33, 0.42]$), a reliably labeled outcome, and tends to push its framing toward pro-China positions (a directional signal on a weaker label; conditional odds ratio 3.10, 95% credible interval $[2.35, 4.07]$). This is not a property of Chinese-built models: the shift is just as large in Western-built ones, and it concentrates in nationally grounded topics such as history and geopolitics. Our per-model atlas also surfaces a notable single-model case, DeepSeek-7B, which answers fluently in English but collapses into a canned non-answer in Chinese on 90% of prompts, a language-capability gap that the bilingual design cleanly separates from value-laden refusal. Training origin, by itself, does not account for these differences in our sample.

1 Introduction

Open large language models now come from labs in many countries and are used both as assistants and to label and analyze text at scale. As they spread, a practical question follows: what shapes the values a model expresses on contested topics? Two candidate factors are usually conflated. One is the model's national origin, the idea that the weights of a model built in a given country carry that country's positions. The other is the language of the prompt, the idea that asking in a language activates the cultural priors associated with it. Separating the two has direct consequences for how practitioners should deploy and audit these systems.

The two factors are easy to confound. Model origin travels with scale, architecture, and the language a user happens to prompt in, and a model built in China also tends to be evaluated in Chinese; prompt language alone is known to move model behavior [1]. Cleanly separating "the weights carry the value" from "the language summons the value" requires a design that varies language while holding the model fixed, which is what we build.

*Corresponding author: deep@lab.cloud

We run such an audit. We take eight open instruction-tuned models in a narrow 7–9B size band, four built in China and four in the West, and ask each the same 145 probe items in both English and Chinese. The items span neutral controls and contested questions in geopolitics, modern history, governance, social values, and culturally specific facts. Each response is scored for refusal, stance, and default cultural perspective by two independent cross-family judge models. Because the English and Chinese versions of an item are asked of the same model, the prompt-language manipulation is within-model and each model is its own control.

Our contributions are:

1. A controlled, symmetric, open-weights audit design that isolates the prompt-language effect by asking every item in both languages, so each model is its own control (§3, §4).

2. Evidence that prompt language is the dominant, controllable lever on how open models answer: Chinese prompting shifts stance by 0.37 points (our validated outcome) and pushes framing toward pro-China positions (odds ratio 3.10, directional support), consistently across models of both origins and concentrated in nationally grounded topics (§5).

3. A per-model atlas that surfaces and cleanly characterizes a distinct single-model phenomenon, DeepSeek-7B's Chinese-language capability gap, and distinguishes it from value-laden refusal (§5).

2 Related Work

Origin and creator effects. Closest to our question, Bladon and Bent [2] argue that geopolitical bias in LLMs originates in post-training and is amplified by prompt language, and Buyl et al. [3] report that a model's normative stance reflects the ideology of its creators across many models and languages. Both establish that origin-linked variation exists; neither performs the size-matched, open-weights, within-model decomposition we use to test whether origin survives as a predictor once a single outlier and the language axis are accounted for. Frank [4] find that across nine mostly Chinese-origin open models the behavioral policy applied to sensitive content is a learned, lab-specific routing that is often language-conditioned (factual in English, refusal or steering in Chinese), and warn that refusal-only audits increasingly miss an invisible narrative-steering mode; Noels et al. [5] document moderation and selective omission tailored to a provider's domestic audience. Our refusal analysis reports refusal as a separately modeled, per-model outcome, and our perspective measure targets the framing-level signal they flag.

Prompt language and multilingual behavior. Tan et al. [6] find that prompting in Chinese relocates rather than removes cultural bias, which directly anticipates our language-effect result. Schut et al. [7] show that multilingual models make semantic decisions in an English-centric latent space, and Cho et al. [1] find that prompt language can dominate content in driving cultural responses; Pan et al. [8] report language-dependent overrefusal. These motivate treating prompt language as a first-class experimental factor.

Measuring values. Prior instruments probe model values with the Political Compass Test and survey items [9, 10], and Lee et al. [11] propose distributional, validity-checked measurement of cultural value alignment while cautioning against single-shot multiple-choice probes. We adopt their caution by using open-ended elicitation with paraphrase-robust, instrument-seeded items. Shen et al. [12] engineer cultural self-awareness through training rather than measuring it.

LLM-as-judge reliability. Our scoring relies on model judges, which carry their own biases. Yang et al. [13] quantify and mitigate judge self-preference, Margalit et al. [14] provide bias-controlled evaluation protocols, and Collot et al. [15] argue for balanced accuracy over κ under class imbalance, while Cho [16] formalize rubric-based collaborative judging. We follow these by using two cross-family judges, excluding same-family self-judging, and reporting agreement appropriately.

Steering. Several methods edit a model's expressed values: linear perspective directions [17], neuron-level value editing that preserves general ability [18], counterfactual value alignment [19], and origin-tied steerable directions identified mechanistically [20]. We do not steer here; these inform future work.

3 Method

We frame the study as a symmetric measurement audit rather than a normative comparison. The unit of analysis is a single `(model, item, language)` response.

Models. We audit eight open instruction-tuned models in a 7–9B band so that origin, not scale, is the variable: four built in China (Qwen2.5-7B [21], DeepSeek-7B [22], GLM-4-9B [23], Yi-1.5-9B [24]) and four in the West (Llama-3.1-8B [25], Gemma-2-9B [26], Mistral-7B [27], OLMo-2-7B [28]).

Probe set. We construct 145 items: 125 contested items spread evenly across five categories (geopolitics, modern history, governance and civil liberties, social values, and culturally specific facts), 25 items per category, and 20 neutral controls. Items are authored as open, balanced questions and seeded where natural from established instruments (the Political Compass Test [29] and World Values Survey [30] framings). Each item is rendered in English and meaning-matched Simplified Chinese, and every Chinese rendering is reviewed for semantic and neutrality parity (a back-translation review for the expanded item set).

Outcomes. For each response, two judge models label: whether the answer *refuses*; its *stance* on the item's axis; and its default cultural *perspective* (pro-China, pro-West, neutral, or other). Refusal is modeled separately, and stance is read only on answered items, so a refusal is never counted as a stance.

Identification. Because each item is asked of the same model in both languages, the English→Chinese contrast is within-model: each model is its own control, so a language effect cannot be explained by between-model differences in size or architecture. We fit mixed-effects logistic models with random intercepts for item and for model, and we test the origin claim both with and without each model to check robustness to outliers.

4 Experimental Setup

Data. The probe set is split group-aware by item (an item's English, Chinese, and paraphrase renderings never straddle the split), stratified by category and contested flag, into 99 development and 46 held-out test items. Generation uses a minimal, neutral system prompt in the prompt's own language, identical decoding across models, and open-ended elicitation.

Scoring. Each of the 2,320 responses is labeled independently by two cross-family judges, Qwen2.5-32B [21] and Mistral-Small-24B [31]. To control self-preference, a judge never scores responses from its own model family; where both judges are valid we take their consensus. This routing leaves 580 responses (25%, the Qwen and Mistral families) judged by a single cross-family judge, with no concordance check for those models. We report inter-judge *consistency* rather than

accuracy: on a 36-response validation sample (single model, Qwen2.5-7B), the two judges agree on refusal with Cohen's $\kappa = 1.00$ and on stance within one point on 91% of answered items, so refusal and stance are our reliable outcomes. Perspective is weaker: raw agreement is 0.86 but $\kappa = 0.23$, deflated by the dominance of the neutral class, a known artifact under class imbalance [15]; we therefore treat perspective as directional support, not a precise rate. No human-labeled ground truth was available, so these numbers establish consistency, not validity. Both judges produce parseable labels on 100% of responses.

Determinism and availability. Generation and judging use fixed decoding and recorded configurations; the probe set, rubric, labels, and analysis code are available on request (§8).

5 Results

The dataset is balanced: 2,320 responses, 1,160 from Chinese-origin and 1,160 from Western-origin models, and 1,160 in each language.

5.1 Prompt language reshapes how models answer

Our reliable outcome is stance, the position a model takes on a contested item's axis (refusal $\kappa = 1.00$, stance within-one 0.91; perspective is weaker and is reported as directional support below). Prompting in Chinese moves stance by a clear margin. Averaged over the within-item English-versus-Chinese pairs, the absolute stance shift is 0.37 points on the five-point axis (95% CI $[0.33, 0.42]$, $n = 850$ item-model pairs), and the shift is of similar size for Chinese-built and Western-built models (0.40 versus 0.35; the difference is not significant). Because each model answers the same item in both languages, this is a within-model effect that between-model differences in size or architecture cannot explain.

The *direction* of the shift tends toward pro-China framing. A mixed-effects logistic regression of pro-China framing on prompt language, with random intercepts for item and model, gives a conditional odds ratio of 3.10 for Chinese (95% credible interval $[2.35, 4.07]$; $\beta = +1.13$, posterior sd 0.14 on the log-odds scale). The conditional odds ratio is larger than the change in the marginal rate, which rises from 0.086 to 0.140 (a marginal odds ratio of 1.73); both describe the same shift on different scales. The rise is present for Western-built models as well as Chinese-built ones (Table 1, Figure 1). This outcome rests on perspective, our least reliable label ($\kappa = 0.23$), and the variational fit can understate posterior width, so we read it as directional support for the validated stance result rather than a precise rate. Per model (Figure 1, 95% Wilson intervals), the rise is clear for Qwen2.5 and GLM-4 (non-overlapping intervals), consistent but within noise for the Western models, and flat for Yi-1.5; the per-model and per-category breakdowns below are descriptive and uncorrected for multiple comparisons.

The effect concentrates in topics with concrete national claims. The English→Chinese rise in pro-China framing is largest for history ($0.010 \to 0.115$) and geopolitics ($0.015 \to 0.095$), present for culturally specific facts ($0.405 \to 0.475$), and essentially absent for abstract value questions in governance ($0.000 \to 0.005$) and social values ($0.000 \to 0.010$); see Figure 2.

5.2 The per-model atlas surfaces a Chinese-language capability gap

Looking model by model lets the audit distinguish two things that an ecosystem average would merge: a value-laden refusal and a plain failure to answer. One model stands out. DeepSeek-7B

Table 1: Probability of pro-China framing on contested items, English versus Chinese prompt. The shift appears for both origins. This is the perspective label (directional support, $\kappa = 0.23$).

Origin	English	Chinese
Chinese-origin models	0.096	0.156
Western-origin models	0.076	0.124
All models (pooled)	0.086	0.140

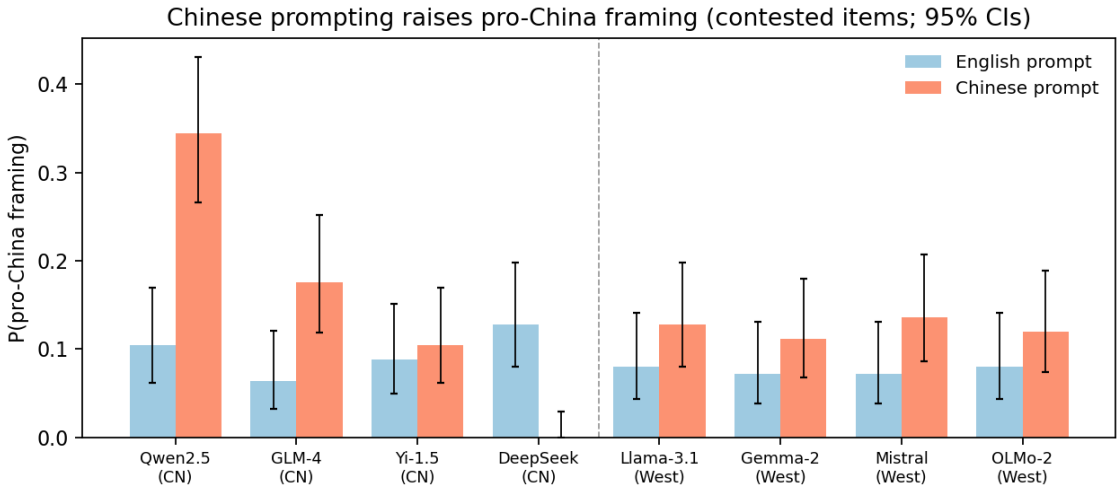

Figure 1: Pro-China framing under English versus Chinese prompting, per model, on contested items (95% Wilson intervals). The rise is clear for Qwen2.5 and GLM-4, consistent but within noise for the Western-built models (right of the dashed line), and flat for Yi-1.5. DeepSeek shows zero pro-China framing in Chinese because it returns a canned non-answer (§5).

answers 95% of contested prompts in English but only a handful in Chinese (Table 2, Figure 3). Reading the raw outputs shows why: for 90% of its Chinese prompts (131/145, including neutral controls such as the boiling point of water), DeepSeek returns an identical canned line, "I am not able to access the internet and therefore do not have access to the latest news or information," while answering the same items correctly in English. The deflection is content-independent, firing on neutral facts as readily as on sensitive ones, so this is a Chinese-language capability gap rather than topic-targeted moderation, a distinction the bilingual design makes directly visible. This single model also drives the aggregate non-answer statistics (Chinese-built models at 25.2% in Chinese versus 1.2% in English, against 2.2% and 0.4% for Western-built models), so those aggregates are best read per model.

5.3 The differences track language, not origin

Setting that one capability gap aside, the contested-topic differences in our data are a language effect rather than an origin effect. With DeepSeek excluded, the Chinese-built Chinese-language non-answer rate is 0.003, and the remaining three Chinese-built models (Qwen2.5, GLM-4, Yi-1.5) answer at Western-model rates, while the language-driven framing shift of §5 holds across both

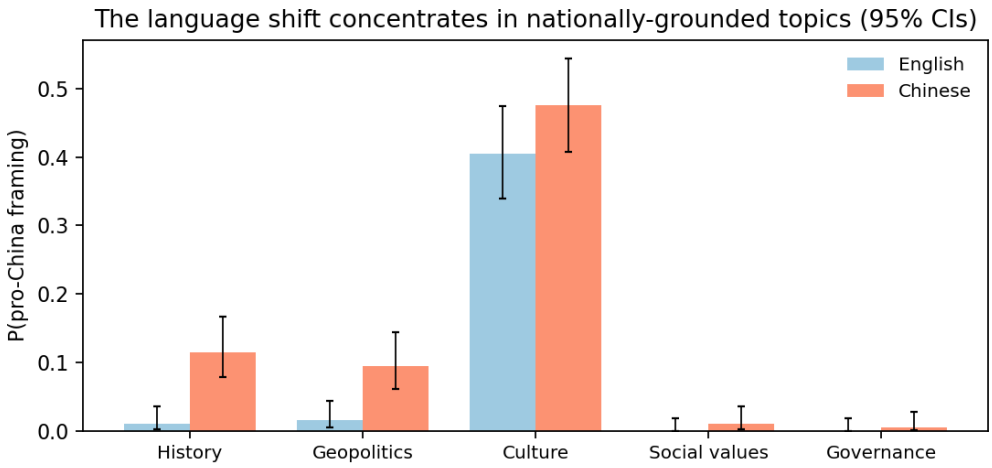

Figure 2: The language-induced shift toward pro-China framing concentrates in nationally grounded topics (history, geopolitics, culture) and is near zero for abstract value questions (governance, social values).

Table 2: DeepSeek-7B non-answer rate by category, English versus Chinese. The Chinese deflection is near-total and topic-independent, including neutral controls, the signature of a capability failure rather than selective censorship.

Category	English	Chinese
Geopolitics	0.00	1.00
History	0.04	1.00
Governance	0.08	1.00
Social values	0.08	1.00
Culturally specific facts	0.04	1.00
Neutral controls	0.00	0.95*

*DeepSeek deflects 19/20 neutral controls in Chinese; for all other models the neutral controls are a clean negative control.

origins. We do not claim a proven origin null: with four models per side the origin contrast is underpowered (the residual interaction estimate, $\beta = -0.49$ with posterior sd 0.87 on the log-odds scale, is uninformative), so the right reading is that origin does not account for the differences we observe, while prompt language does. The neutral controls are a clean negative control for every model except DeepSeek in Chinese.

6 Discussion

Two readings of "national values in models" are often conflated. One is that the weights of a model built in a given country encode that country's positions, so the model will express them regardless of how it is prompted. The other is that prompting in a language activates the cultural priors associated with that language, in any model. Our within-model design favors the second: the reliable, cross-model effect is that Chinese prompting shifts how a model answers (the validated

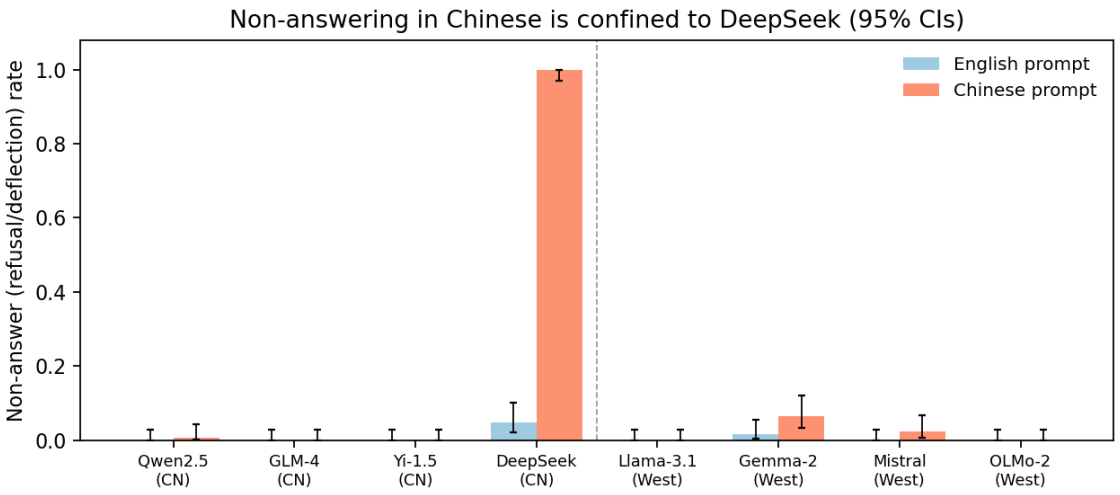

Figure 3: Non-answering in Chinese is confined to DeepSeek-7B (95% Wilson intervals). The other seven models, Chinese-built and Western alike, answer in both languages.

stance result), and the direction tends pro-China in Western models as much as Chinese-built ones. This is consistent with reports that prompt language relocates cultural bias [6], that multilingual models reason in an English-centric space [7], and that cultural defaults shift with the input language [1].

A complementary mechanistic line finds origin-linked features as model-exclusive linear directions that can be causally steered [20], which indicates that weight-resident origin signals coexist with the prompt-language effect we measure. Our behavioral audit does not contradict this; it shows that at the level of what a user actually receives, the language of the prompt is the dominant, cross-model lever.

The bilingual, per-model design also pays off by separating effects that an aggregate would merge. The most pronounced single-model behavior in our data is not a value choice at all but a Chinese-language capability gap, and because we read each model on its own and in both languages, we can label it as such rather than fold it into an ecosystem-level claim about values. The practical takeaway for practitioners and auditors is that the prompt language is the first lever to control when measuring or deploying these models, and that per-model, per-language reporting is what keeps a capability failure from being read as a value stance.

7 Limitations

The judge-language confound (the main open threat). Our perspective and stance labels are assigned by model judges that read the response in its own language: English judges for English responses, Chinese for Chinese. If a judge is even slightly more inclined to read Chinese-language text as pro-China regardless of content, that bias is perfectly aligned with our English-to-Chinese manipulation and would reproduce the language effect with no change in model behavior. One of our two judges (Qwen2.5-32B) is itself Chinese-built. Our within-model design controls the model but not the judge, and a defense that the contrast cancels a *constant* judge bias does not cover a *language-correlated* one. Prior work also reports that model judges overdetect political framing relative to humans [4]. We did not run the control that would settle this: judging back-translated

responses (Chinese responses translated to English and vice versa), or human labeling on a language-balanced subset. Until that is done, the directional pro-China result should be read as suggestive. We treat this as the primary item for future work.

Construct and measurement validity. "Pro-China framing" and "stance" are judge-assigned labels validated only for *consistency*, not accuracy: inter-judge agreement on a 36-response, single-model sample, with no human ground truth (refusal $\kappa = 1.00$; stance within-one 0.91; perspective $\kappa = 0.23$). Perspective is the weakest outcome, which is why the headline rests on stance and perspective is reported as directional support. A quarter of responses (580/2,320, the Qwen and Mistral families) are judged by a single cross-family judge, so consensus is unavailable for them. The reported odds ratios use a variational fit that can understate posterior width, and the refusal interaction estimates are separation-driven and should not be read as ordinary effect sizes.

External validity. We audit eight models at 7–9B; larger and closed models may differ, and four models per origin is too few to license a population-level claim about either ecosystem, which is why the origin result is a failure to detect rather than a null. We study English and Simplified Chinese only.

Internal validity. Translation non-equivalence could masquerade as a language effect; we reduce this with a back-translation neutrality-parity review of the items, but residual drift cannot be fully excluded, and we did not report inter-translator agreement on the full set. We use open-ended greedy decoding with one sample per cell, so within-condition variability and paraphrase robustness are unestimated. The weights-versus-language question is addressed behaviorally; a base-model and mechanistic decomposition is left to future work.

8 Availability

The artifacts underlying this study, including the 145-item bilingual probe set, the dual-judge rubric, the full per-response labels, and the analysis scripts that reproduce every number and figure in this paper, are available from the authors on request.

9 Conclusion and Future Work

A controlled, symmetric, open-weights audit shows that prompt language is the dominant, controllable lever on how open LLMs answer contested questions: Chinese prompting shifts stance by 0.37 points (our validated outcome) and pushes framing toward pro-China positions (odds ratio 3.10, directional support), consistently across models of both origins and concentrated in nationally grounded topics. The same per-model, bilingual lens lets us identify and correctly label a distinct single-model capability gap (DeepSeek-7B in Chinese) rather than mistake it for a value stance, and it shows that training origin does not by itself account for the differences we observe. The most important next step is to rule out a judge-language confound, by judging back-translated responses or human-labeling a language-balanced subset; beyond that, a base-versus-chat decomposition would localize the language effect to pretraining or post-training, and a steering study would test whether the language-summoned framing can be neutralized without harming general ability. The benchmark is available on request to make these follow-ups straightforward.

References

[1] Seungho Cho, Changgeon Ko, Eui Jun Hwang, Junmyeong Lee, Huije Lee, and Jong C. Park. Language over Content: Tracing Cultural Understanding in Multilingual Large Language Models. 2025. URL https://arxiv.org/abs/2510.16565.

[2] Stuart Bladon and Brinnae Bent. It's the Humans, Not the Data: Geopolitical Bias in LLMs Originates in Post-Training, Amplified by the Language of the Prompt. 2026. URL https://arxiv.org/abs/2605.23825.

[3] Maarten Buyl, Alexander Rogiers, Sander Noels, Guillaume Bied, Iris Dominguez-Catena, Edith Heiter, Iman Johary, Alexandru-Cristian Mara, Raphaël Romero, Jefrey Lijffijt, and Tijl De Bie. Large Language Models Reflect the Ideology of their Creators. 2024. URL https://arxiv.org/abs/2410.18417.

[4] Gregory N. Frank. Detection Is Cheap, Routing Is Learned: Why Refusal-Based Alignment Evaluation Fails. 2026. URL https://arxiv.org/abs/2603.18280.

[5] Sander Noels, Guillaume Bied, Maarten Buyl, Alexander Rogiers, Yousra Fettach, Jefrey Lijffijt, and Tijl De Bie. What Large Language Models Do Not Talk About: An Empirical Study of Moderation and Censorship Practices. 2025. URL https://arxiv.org/abs/2504.03803.

[6] Qian Tan, Lei Jiang, Yuting Zeng, Shuoyang Ding, and Xiaohua Xu. Mitigating Cultural Bias in LLMs via Multi-Agent Cultural Debate. 2026. URL https://arxiv.org/abs/2601.12091.

[7] Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do Multilingual LLMs Think In English? 2025. URL https://arxiv.org/abs/2502.15603.

[8] Licheng Pan, Yongqi Tong, Xin Zhang, Xiaolu Zhang, Jun Zhou, and Zhixuan Chu. Understanding and Mitigating Overrefusal in LLMs from an Unveiling Perspective of Safety Decision Boundary. 2025. URL https://arxiv.org/abs/2505.18325.

[9] Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. Revealing Fine-Grained Values and Opinions in Large Language Models. 2024. URL https://arxiv.org/abs/2406.19238.

[10] Konrad Löhr, Shuzhou Yuan, and Michael Färber. The Hidden Bias: A Study on Explicit and Implicit Political Stereotypes in Large Language Models. 2025. URL https://arxiv.org/abs/2510.08236.

[11] Jaehyeok Lee, Xiaoyuan Yi, Jing Yao, Hyunjin Hwang, Roy Ka-Wei Lee, Xing Xie, and JinYeong Bak. Distributional Open-Ended Evaluation of LLM Cultural Value Alignment Based on Value Codebook. 2026. URL https://arxiv.org/abs/2604.06210.

[12] Lingzhi Shen, Xiaohao Cai, Yunfei Long, Imran Razzak, Guanming Chen, and Shoaib Jameel. CALM: Culturally Self-Aware Language Models. 2026. URL https://arxiv.org/abs/2601.03483.

[13] Jinming Yang, Zheng Hu, Chuxian Qiu, Zhenyu Deng, Xinshan Jiao, and Tao Zhou. Quantifying and Mitigating Self-Preference Bias of LLM Judges. 2026. URL https://arxiv.org/abs/2604.22891.

[14] Yanki Margalit, Erni Avram, Ran Taig, Oded Margalit, and Nurit Cohen-Inger. PeerRank: Autonomous LLM Evaluation Through Web-Grounded, Bias-Controlled Peer Review. 2026. URL https://arxiv.org/abs/2602.02589.

[15] Stephane Collot, Colin Fraser, Justin Zhao, William F. Shen, Timon Willi, and Ilias Leontiadis. Balanced Accuracy: The Right Metric for Evaluating LLM Judges Explained through Youden's J statistic. 2025. URL https://arxiv.org/abs/2512.08121.

[16] Arthur Cho. GrandJury: A Collaborative Machine Learning Model Evaluation Protocol for Dynamic Quality Rubrics. 2025. URL https://arxiv.org/abs/2508.02926.

[17] Junsol Kim, James Evans, and Aaron Schein. Linear Representations of Political Perspective Emerge in Large Language Models. 2025. URL https://arxiv.org/abs/2503.02080.

[18] Yonghui Yang, Junwei Li, Jilong Liu, Yicheng He, Fengbin Zhu, Weibiao Huang, Le Wu, Richang Hong, and Tat-Seng Chua. Controllable Value Alignment in Large Language Models through Neuron-Level Editing. 2026. URL https://arxiv.org/abs/2602.07356.

[19] Hanze Guo, Jing Yao, Xiao Zhou, Xiaoyuan Yi, and Xing Xie. Counterfactual Reasoning for Steerable Pluralistic Value Alignment of Large Language Models. 2025. URL https://arxiv.org/abs/2510.18526.

[20] Thomas Jiralerspong and Trenton Bricken. Cross-Architecture Model Diffing with Crosscoders: Unsupervised Discovery of Differences Between LLMs. 2026. URL https://arxiv.org/abs/2602.11729.

[21] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 Technical Report. 2024. URL https://arxiv.org/abs/2412.15115.

[22] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, et al. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. 2024. URL https://arxiv.org/abs/2401.02954.

[23] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, et al. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. 2024. URL https://arxiv.org/abs/2406.12793.

[24] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, et al. Yi: Open Foundation Models by 01.AI. 2024. URL https://arxiv.org/abs/2403.04652.

[25] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 Herd of Models. 2024. URL https://arxiv.org/abs/2407.21783.

[26] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, et al. Gemma 2: Improving Open Language Models at a Practical Size. 2024. URL https://arxiv.org/abs/2408.00118.

[27] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile

Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. 2023. URL https://arxiv.org/abs/2310.06825.

[28] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, et al. 2 OLMo 2 Furious. 2024. URL https://arxiv.org/abs/2501.00656.

[29] Political Compass Organisation. The Political Compass. https://www.politicalcompass.org, 2024. URL https://www.politicalcompass.org.

[30] Christian Haerpfer, Ronald Inglehart, et al. World Values Survey. Technical report, World Values Survey Association, 2022. URL https://www.worldvaluessurvey.org.

[31] Mistral AI. Mistral Small 3. https://mistral.ai/news/mistral-small-3, 2025. URL https://mistral.ai/news/mistral-small-3.