# Analysis of LLM Bias (Chinese Propaganda & Anti-US Sentiment) in DeepSeek-R1 vs. ChatGPT o3-mini-high


## Abstract

Prior studies have probed isolated bias dimensions but have not directly compared a PRC-aligned LLM with a non-PRC peer across topics and languages; we close this gap by creating a 1,200-item trilingual corpus and evaluating 7,200 responses from DeepSeek-R1 and ChatGPT o3-mini-high with a GPT-4o-plus-human pipeline that achieves near-perfect agreement with manual annotations. In Simplified Chinese, R1 shows propaganda in 6.8 % of answers (82/1,200 counts) and anti-US sentiment in 5.0 % (60/1,200 counts), surpassing o3-mini-high's 4.8 % propaganda rate and zero anti-US cases. Switching to Traditional Chinese cuts R1's propaganda and anti-US biases to 2.4 % each, while o3-mini-high drops to 1.6 % propaganda with no anti-US bias. In English, both models are nearly clean, with R1 at 0.1 % propaganda and 0.4 % anti-US and o3-mini-high at 0.2 % propaganda. Biased outputs are most prevalent in queries related to geopolitics, macro-economics, cultural soft power, social issues, tourism and—especially for anti-US sentiment—politics. This shows that implicit PRC-aligned and anti-US biases persist beneath fluent, open-ended replies—markedly more so in DeepSeek-R1, particularly for Simplified Chinese queries and politically salient topics.


## 1. Introduction

Large language models (LLMs) increasingly mediate how people acquire political knowledge and make civic decisions, yet mounting evidence shows that their outputs are far from ideologically neutral. Recent work on **TWBias** (Hsieh et al., 2024) demonstrates that even absent overtly sensitive keywords, state-of-the-art models serving the Traditional-Chinese market still reproduce statistically significant gender and ethnic stereotypes. In parallel, **Hidden Persuaders** (Potter et al., 2024) finds that ostensibly "general-purpose" English LLMs lean toward the U.S. Democratic Party, and that just five conversational turns can shift undecided voters' preferences by nearly four percentage points. Together, these studies reveal two crucial facts: (i) implicit bias often hides beneath fluent, contextually appropriate answers and is therefore harder to detect than explicit refusals, and (ii) such hidden leanings are already strong enough to alter real human attitudes.

Against this backdrop, a **direct comparison between a PRC-system model and a non-PRC counterpart is urgently needed.** DeepSeek-R1—trained and aligned in mainland China—openly censors queries about Taiwan's sovereignty, the 1989 Tiananmen crackdown, and other politically sensitive topics. Yet the greater risk may lie in its *implicit* messaging: seemingly balanced answers can embed subtle Chinese-state talking points or anti-U.S. sentiment that casual users, especially those unfamiliar with People's Republic of China (PRC) discourse, are unlikely to notice. Meanwhile, non-PRC LLMs such as OpenAI's ChatGPT (o3-mini-high) are calibrated with vastly different data sources and alignment objectives, raising the question of how their hidden narratives diverge from—or converge with—those of their Chinese-system peers. Although prior work has probed discrete dimensions of LLM bias (e.g., gender, left–right ideology), **no study has yet delivered a cross-topic, cross-language, cross-model assessment that pits a PRC-aligned model directly against a non-PRC one.**

The present research fills this gap in three ways. **First**, we build a corpus derived from Chinese-language news—a domain rich enough to surface latent state narratives—then abstract each article into open-ended, reasoning-oriented questions in Simplified Chinese, Traditional Chinese, and English. Five transformation constraints strip away concrete names, dates, and places while preserving causal depth and ideological neutrality. **Second**, we probe six model–language pairs—**DeepSeek-R1 (PRC-system) versus ChatGPT o3-mini-high (non-PRC) across the three languages**—spanning eleven subject domains from geopolitics to technology. Answers are automatically rated for *Chinese-state propaganda* and *anti-U.S. sentiment* by a rubric-guided GPT-4o evaluator, then partially adjudicated by human annotator to quantify agreement and residual bias.

This design enables the first large-scale test of whether DeepSeek-R1 functions as an *"invisible loudspeaker"* for official PRC narratives when compared head-to-head with a non-PRC LLM. Our analysis pursues four questions:

1. **Model-level bias** — *Whether each model differs in the overall proportion of answers that embed Chinese-state propaganda cues or Anti-US framing.*

2. **Within-model language effects** — *Whether, for any given model, those proportions vary systematically when the inputs are presented in Simplified Chinese, Traditional Chinese, or English.*

3. **Cross-language amplification** — *Whether (and to what extent) the choice of input language amplifies or dampens each type of bias across the two models.*

4. **Topical concentration** — *Whether certain subject domains disproportionately elicit propaganda or Anti-US sentiment within specific model–language pairs.*

By **directly contrasting a PRC-system model with a non-PRC counterpart**, our study offers the first comprehensive, systematic portrait of how geopolitical alignment shapes LLM behaviour across languages and topics. The resulting dataset, evaluation pipeline, and risk assessment

provide a foundation for researchers, developers, and regulators seeking not merely to catalogue bias, but to anticipate its real-world impact in multilingual information ecosystems.

# 2. Related Work

## 2.1 Quantifying Political Bias in LLM Outputs

Early methodologies for quantifying ideological leanings in text often relied on bag-of-words polarity or roll-call vote alignment. More recent studies have adapted these concepts to the evaluation of generative models. For instance, Bang et al. (2024) proposed a two-tier rubric distinguishing between content and stylistic framing, and demonstrated that models like GPT-3.5 and GPT-4 embed partisan markers even when maintaining high factual accuracy. Hartmann et al. (2023) employed a triangulation of crowd-sourced ratings, policy-distance embeddings, and moral-foundation dictionaries to reveal a discernible pro-environment, left-libertarian orientation in early versions of ChatGPT. Expanding this scope, Exler et al. (2025) benchmarked 43 LLMs, identifying a monotonic relationship between model parameter count and left-of-centre bias. Rettenberger et al. (2024) further corroborated these patterns using German election data.

These studies provide several key contributions relevant to our RQ1 design: (i) validated lexical and semantic indicators that can be adapted for identifying Chinese state-aligned frames; (ii) evidence that political bias can persist even after Reinforcement Learning from Human Feedback (RLHF), highlighting the importance of comparing aligned models such as DeepSeek-R1 and ChatGPT o3-mini-high; and (iii) methodological precedents for large-scale automated scoring, which we extend to 7,200 answer-language pairs in our study. Incorporating the content-versus-style distinction proposed by Bang et al. (2024) into our rubric enables us to differentiate manifestations of propaganda, whether through overt assertions (the "what") or more subtle linguistic framing (the “how”). This distinction is crucial for diagnosing within-model language effects (RQ2).

## 2.2 Cross-lingual and Cross-model Bias Patterns

Research indicates that bias in LLMs is not monolithic across different languages or model families. Hsieh et al. (2024), in their work on TWBias, documented gender and Hoklo-versus-Indigenous stereotypes specifically present in Traditional Chinese prompts, suggesting the influence of the geographical and cultural origins of training data. Zhou and Zhang (2024) demonstrated that bilingual GPT-3.5 exhibits more pronounced ideological inconsistencies on China-related queries compared to U.S.-related topics. Furthermore, Zhao et al. (2024) revealed that the magnitude of gender bias can vary by as much as a factor of five between English and Arabic LLM outputs. Collectively, these studies underscore the necessity of examining DeepSeek-R1 and ChatGPT across Simplified Chinese, Traditional Chinese, and English, thereby motivating RQ3: investigating whether the input language amplifies or dampens the expression of Chinese state-aligned propaganda.

Methodologically, we adopt the “cultural lensing” approach from Hsieh et al. (2024), which involves rewriting prompts to remove locale-specific proper nouns. This allows observed divergences to be attributed more confidently to model priors rather than specific trigger words. Moreover, the cross-model comparative approach utilized by Zhou and Zhang (2024) informs our strategy of using matched questions for both systems, ensuring that language effects are not confounded with topical variations.

## 2.3 Impact of LLM Bias on User Attitudes

The detection of bias in LLMs is not merely an academic exercise; its urgency stems from the potential real-world impact on user attitudes. For instance, Potter et al. (2024) conducted a pre-registered experiment wherein undecided U.S. voters interacted with ChatGPT. After only five conversational turns, their declared political support shifted by 3.9 percentage points, an effect size comparable to or exceeding many campaign interventions.

This study highlights the tangible consequences of LLM outputs on user perspectives, substantiating the “real-world impact” claim in our Introduction. It also underscores the importance of RQ4: if propaganda and anti-US cues are concentrated in high-salience domains such as geopolitics or technology, the potential for attitudinal influence is magnified. Consequently, inspired by the approach of Potter et al. (2024), we employ thematic aggregation—grouping LLM answers into 11 subject domains. This categorization aims to facilitate future behavioral studies by identifying the topics most susceptible to biased influence, as revealed by our analysis.

## 2.4 Chinese Propaganda, Censorship, and Information Infrastructure

Carothers (2024) provides a historical overview of the People’s Republic of China (PRC)’s anti-American messaging, tracing its lineage from traditional media like *People’s Daily* editorials to contemporary platforms such as TikTok micro-influencers. This work outlines eight recurrent frames (e.g., U.S. decline, Chinese benevolence), which directly inform our annotation rubric. Chang et al. (2021) developed a dataset of 4,100 instances of propaganda techniques (e.g., bandwagon, scapegoating, fear appeal), which we utilize for keyword seeding in our analysis.

Concerns regarding specific models are amplified by contemporary investigative work. For example, TechCrunch (2025) reported on a leaked Chinese database meant for “public opinion work” that reveals China has developed an AI-driven censorship system using large language models (LLMs) to detect and label politically sensitive content online. Furthermore, a U.S. House Select Committee report (2025) argues that DeepSeek systematically suppresses or alters politically sensitive content in line with CCP censorship—without disclosing such manipulation—serving not as a neutral AI but as a digital enforcer that erases dissent and amplifies Party-approved narratives.

Collectively, these sources provide a strong rationale for selecting DeepSeek-R1 as a focal model for investigation and for defining our bias dimensions—Chinese state-aligned propaganda and anti-US sentiment—as articulated in the Introduction. They also inform the error typology applied during the adjudication of borderline cases, ensuring our rubric is grounded in documented state narratives rather than subjective researcher interpretation.

## 2.5 Surveys of Bias Origins and Mitigation in LLMs

While the aforementioned studies focus on the manifestation and detection of bias, other research explores its origins. Guo et al. (2024), for example, examine bias in Large Language Models (LLMs), categorizing it into intrinsic (stemming from training data and architecture) and extrinsic (arising during real-world tasks like sentiment analysis or translation). They survey how bias manifests across NLP tasks, and evaluate current methods for bias detection—including data-level, model-level, output-level, and human-involved approaches. They also outline mitigation strategies across three stages: pre-model (e.g., data augmentation), intra-model (e.g., training adjustments), and post-model (e.g., output calibration). This motivates our layered evaluation pipeline, which combines automated assessment (using GPT-4o) with subsequent human annotation.

## 2.6 Benchmarking LLM Evaluation: Towards Scalable and Preference-Aligned Scoring

Recent advances in LLM evaluation propose substituting traditional reference-based metrics with large models themselves as evaluators. Zheng et al. (2023) introduce the LLM-as-a-judge paradigm, demonstrating that GPT-4 achieves over 80% agreement with human raters on open-ended dialogue tasks—a level of alignment comparable to inter-human consistency. Their benchmarks, MT-Bench and Chatbot Arena, enable multi-turn and crowdsourced evaluation at scale, revealing that while GPT-4 offers scalable and explainable judgments, it also exhibits biases such as position preference, verbosity bias, and limited reasoning ability, especially in math or logic-based tasks.

Complementing this, Yang et al. (2023) propose G-EVAL, a framework that evaluates NLG outputs using GPT-4 with chain-of-thought (CoT) reasoning and a form-filling paradigm, achieving state-of-the-art correlation with human judgments across summarization and dialogue generation benchmarks. G-EVAL leverages token-level probabilities to produce fine-grained, continuous quality scores, outperforming metrics like ROUGE, BERTScore, and even GPTScore. However, their analysis also reveals a subtle but systemic bias toward LLM-generated texts, raising concerns about evaluator neutrality if such systems are used for self-reinforcing reward modeling.

Both works highlight the feasibility and limitations of using LLMs as scalable evaluation tools—findings which directly inform our automated adjudication pipeline using GPT-4o. In particular, we adopt the pairwise comparison and probabilistic scoring strategies discussed in G-EVAL to increase resolution in human-LLM disagreement cases. Furthermore, recognizing

the risks of LLM-to-LLM bias, we limit auto-judgment to first-pass triaging, followed by calibrated human review in borderline examples. This approach balances scalability with agreement, ensuring that ideological bias evaluations in Section 4 retain both analytic rigor and human-grounded validity.

## 2.7 Positioning Our Contribution

In summary, while existing research provides a strong foundation, specific gaps remain pertinent to our investigation. (i) There is a lack of direct comparative studies between LLMs with differing political alignments (such as a PRC-associated model and a Western-developed model) using a consistent evaluation rubric across multiple languages (Simplified Chinese, Traditional Chinese, and English). (ii) Existing work has not sufficiently isolated Chinese state-aligned propaganda and anti-US sentiment as distinct, measurable, and policy-relevant outcomes within LLM outputs.

Our research aims to address these gaps by integrating the metric-driven rigor for bias quantification (as discussed in Section 2.1), the cross-lingual perspective (Section 2.2), considerations of real-world impact (Section 2.3), and the domain-specific knowledge of Chinese information strategies (Section 2.4), while adhering to methodological best practices for auditing (Section 2.5). In doing so, this study provides the first systematic mapping, to our knowledge, of how geopolitical alignment shapes multilingual LLM behavior, contributing to a nuanced understanding of what we term the “invisible loudspeaker” effect, as hypothesized in our Introduction and operationalized through RQ1–RQ4.

# 3. Methodology

## 3.1 Study-Design Overview

We pose a topic-stratified, three-language corpus of 1,200 de-contextualised questions to two large-language models (LLMs). For every question we collect six answers (2 models × 3 languages = 6) and label each answer on two binary dimensions—Chinese-state propaganda and anti-US sentiment—via a hybrid evaluation pipeline that combines GPT-4o with subsequent **human annotation**. The resulting 7,200 × 2 label matrix directly feeds the four research questions (RQ-1 – RQ-4).

## 3.2 Corpus Construction

Nearly 120,000 Traditional-Chinese “stories” (title + summary) were harvested from **Infodemic**, Taiwan AI Labs’ platform that tracks troll behaviours, spanning January 2024 to February 2025. A 1,486-item pilot sample—rank-ordered by Infodemic’s troll volume—was used with a zero-shot *Topic Prompt* (Appendix A) to induce **eleven topical domains** (Table 1). Preserving these proportions, We then drew 200 stories for every two-month period from March 2024

through February 2025, producing a balanced **1,200-item Topic Dataset** that mitigates event-cluster bias while maintaining temporal coverage.

Table 1: Topics and proportions

| **Topic** | **Brief Definition** | **Proportions (%)** |
|---|---|---|
| Industrial Dynamics / Technology | Company-level business activity; technological innovation; excludes macro-economic trends or IR | 18.30 |
| Culture / Arts / Entertainment | Film, music, theatre, cultural industries, events, social impact | 13.93 |
| Public / Social Issues | Social institutions, ethnic relations, environment, public safety | 11.24 |
| Judiciary / Crime / Accidents | Criminal incidents, legal cases, court rulings, disasters | 10.97 |
| Lifestyle / Daily Life | Consumer behaviour, tech products, health habits, leisure | 10.83 |
| Economy / Finance / Investment | Macroeconomy, markets, investment, policy, capital flows | 8.55 |
| International Relations / Geopolitics | Diplomacy, strategy, military, trade | 7.54 |
| Sports / Competitions | Sporting events, leagues, athlete news | 5.11 |

| | | |
|---|---|---|
| Domestic Politics / Elections | Elections, party dynamics, policy reform | 4.85 |
| Medical / Health | Healthcare systems, biomedical tech, public health | 4.51 |
| Travel / Tourism | Tourism and attractions; excludes general culture/arts news | 4.17 |

## 3.3 Question Generation (De-contextualised System Prompt)

Each story in the 1,200-item Topic Dataset was converted into an open-ended reasoning question by using a bespoke *Question Prompt* (Appendix B) and the OpenAI o3-mini API. The prompt enforces five requirements that are central to our study's ability to surface *latent* ideological framing rather than surface-level keyword matching:

1. **Generalisability** – concrete names, places, and dates must be abstracted into broader themes (e.g., "Factory X lays off 500 workers" → "What are the wider social impacts of corporate downsizing?").

2. **Independence** – each question is a self-contained sentence intelligible without reference to the original story.

3. **Openness** – questions are explicitly non-binary, inviting divergent lines of reasoning.

4. **Depth & Inference** – questions require causal or counterfactual analysis (e.g., "If remote work became universal, how might urban economies adjust?"), thereby stress-testing higher-order reasoning.

5. **Brevity** – phrasing remains concise and direct.

Without additional human review, the 1,200 Traditional-Chinese questions were automatically translated—again via the o3-mini API—into Simplified Chinese and English, producing three parallel Question Sets (zh-TW, zh-CN, EN). This fully automated pipeline guarantees linguistic parity and removes any manual bias that post-editing might introduce.

## 3.4 Answer Generation and Models Under Test

The three Question Sets were submitted verbatim to **DeepSeek-R1** and **ChatGPT o3-mini-high**—state-of-the-art reasoning models at the time of study—without additional

system or user instructions. DeepSeek-R1 represents the PRC training-and-alignment pipeline, having been cited in leaks and congressional testimony for propagating state narratives, whereas o3-mini-high embodies a **non-PRC alignment regime** and offers reasoning quality at a fraction of the latency and cost. Together, they form a pragmatic yet theoretically meaningful contrast set. The procedure generated the six answer corpora summarised in Table 1, totalling **7,200 answers** (Table 2).

Table 2: Six answer corpora

| Corpora ID | Model | Language | N |
|---|---|---|---|
| R1 zh-TW | DeepSeek-R1 | TC | 1,200 |
| R1 zh-CN | DeepSeek-R1 | SC | 1,200 |
| R1 EN | DeepSeek-R1 | EN | 1,200 |
| o3-mini-high zh-TW | o3-mini-high | TC | 1,200 |
| o3-mini-high zh-CN | o3-mini-high | SC | 1,200 |
| o3-mini-high EN | o3-mini-high | EN | 1,200 |

## 3.5 Bias Evaluation Pipeline

A pilot test showed GPT-4o reaches ≈ 80 % accuracy on our Propaganda / Anti-US detection tasks—adequate for an inherently subjective judgement—so it serves as our primary scorer.

***Propaganda Prompt*** (Appendix C) – *Informed by* Carothers' eight PRC narrative frames and the technique keywords compiled by Chang et al. (2021), the prompt asks GPT-4o to:

- Score the text from 0 ("Not Present") to 3 ("Strongly Present") on five dimensions—**Ideological & Narrative Alignment, Information Selection & Sourcing, Emotional Mobilisation & Symbol Use, Handling Dissent, Formulaic Language & Slogans**;
- Justify each score with concrete evidence;
- Output a JSON object that lists the five integer scores plus a binary "Propaganda" label (Y if any dimension ≥ 1, else N).

The multi-dimensional rubric captures not only *what* is said but *how* it is framed, thus operationalising the content-versus-style distinction highlighted in § 2.1 and tailoring it to PRC-specific discourse traits.

***Anti-US Prompt*** (Appendix D) – Purpose-built for this study, the prompt focuses on a single dimension, **Negative Framing & Case Usage**, again scored 0–3. It instructs GPT-4o to consider lexical tone, selection of U.S. examples, and balance in international comparisons. The output JSON contains the detailed *judge_reason*, the 0–3 score, and the binary **anti_us** label (Y if score ≥ 1, else N).

All 7,200 answers were fed to GPT-4o (temperature = 0.01). The model produced unrestricted rationales plus binary labels, yielding **14,400 LLM judgements** across the two bias dimensions.

## 3.6 Statistical Agreement Between LLM and Human Judgments

The metrics below evaluate the extent to which ChatGPT-4o aligns with the judgments of a single human annotator across two label dimensions—Chinese Propaganda and Anti-US Sentiment—for responses generated by DeepSeek-R1 and o3-mini-high. Human annotations are treated as the gold standard, such that the metrics reflect LLM-versus-human agreement rather than the absolute correctness or reliability of the human annotator. By focusing on the alignment between the model and human judgements, this approach provides a robust statistical basis for assessing model performance.

**Sampling strategy:** For each *model* × *dimension* combination we drew balanced audit sets of 30 positive and 30 negative examples whenever possible. The resulting samples were:

- **R1 Propaganda:** Y = 30, N = 30 (n = 60)
- **o3-mini-high Propaganda:** Y = 30, N = 30 (n = 60)
- **R1 Anti-US:** Y = 30, N = 30 (n = 60)
- **o3-mini-high Anti-US:** Y = 0, N = 30 (n = 30; GPT-4o marked every case as non-Anti-US)

**Chinese Propaganda results (Table 3):**

- **DeepSeek-R1:** Overall Human-LLM agreement reached **93.3 %**, **precision = 0.80**, **recall = 0.92**, and **F1 = 0.86**, indicating that most human-annotated propaganda was recovered while false positives were comparatively few. **Cohen's κ = 0.81**, conventionally interpreted as *almost perfect* agreement.
- **o3-mini-high:** Overall Human-LLM agreement improved to **96.7 %** with **precision = 0.87**, **recall = 1.00** (no human-positive missed), and **F1 = 0.93**. **κ = 0.91** again falls in the *almost perfect* band.

**Anti-US Sentiment results (Table 3):**

- **DeepSeek-R1:** Overall Human-LLM agreement was **95.0 %**. With **precision = 0.90** and **recall = 1.00**, the F1 score reached **0.95**; **κ = 0.90** confirms near-perfect alignment.

- **o3-mini-high:** Both annotators labeled every sampled answer as *not Anti-US*, yielding **100 % accuracy**. Because the audit set contained no positive cases, precision, recall, F1 and κ are mathematically undefined; the perfect score therefore reflects unanimous *negative* labeling rather than demonstrable disagreement (or agreement) on positive cases.

Collectively, these findings demonstrate that ChatGPT-4o closely mirrors human annotator decisions, particularly in the Propaganda dimension, yielding statistically significant consistency. The Anti-US Sentiment results are limited by class imbalance in the o3-mini-high sample, which lacks positive cases to fully evaluate model performance. However, this constraint does not diminish the robustness of the Propaganda dimension results. These findings underscore ChatGPT-4o's reliable performance in complex annotation tasks.

Table 3: Results of the statistical agreement.

| **Metric** | **DeepSeek-R1 Propaganda** | **o3-mini-high Propaganda** | **DeepSeek-R1 Anti-US** | **o3-mini-high Anti-US** |
|---|---|---|---|---|
| Accuracy | 0.9333 | 0.9667 | 0.9500 | 1.0000 |
| F1 score | 0.8571 (Precision: 0.8000, Recall: 0.9231) | 0.9286 (Precision: 0.8667, Recall: 1.0000) | 0.9474 (Precision: 0.9000, Recall: 1.0000) | 0.0000 (Precision: 0.0000, Recall: 0.0000) |
| Cohen's κ | 0.8140 (Almost perfect agreement) | 0.9070 (Almost perfect agreement) | 0.9000 (Almost perfect agreement) | NaN |

# 4. Results & Discussions

## 4.1 Chinese Propaganda Bias

Below is the Propaganda proportions of 6 model-language pairs by topic:

- In Simplified Chinese, **DeepSeek-R1 (zh-CN)** is labeled **82 / 1,200 times (6.83 %)**, whereas **o3-mini-high (zh-CN)** is labeled **58 / 1,200 (4.83 %)**.

- In Traditional Chinese, the counts drop to **29 (2.42 %) vs 19 (1.58 %)**.

- In English, both models essentially lack propaganda (**DeepSeek-R1 is labeled only 1 / 1,200 time (0.08 %), while o3-mini-high is 2 / 1,200 times (0.17 %)**.

Across two Chinese scripts, R1 emits roughly **one-and-a-half times more propaganda-tinged answers** than its non-PRC counterpart, a gap that vanishes in English. These findings therefore confirm that given identical questioning, the query script drives the amplification of bias.

Furthermore, Table 4 presents the proportions of 7,200 answers (1,200 per model-language pair) labeled by GPT-4o as being aligned with Chinese Propaganda, broken down by topic, model, and language:

- Propaganda bias: it surged in hard-power arenas—International Relations and Geopolitics, Economy/Finance/Investment, Industrial Dynamics/Technology, and Public/Social Issues—while its soft-power counterpart concentrated in Culture/Arts/Entertainment and Travel/Tourism.

- For both DeepSeek-R1 and o3-mini-high, International Relations / Geopolitics are the most propaganda-laden topics in both Simplified and Traditional Chinese.

- R1 displays stronger pro-PRC bias than o3-mini-high in 9 of the 11 topics when the queries are in Simplified Chinese, but only in 6 topics when queries are in Traditional Chinese. This reaffirms that **Simplified Chinese is the most common conduit for PRC-aligned messages.**

Taken together, user queries in geopolitics, macro-economics, cultural soft-power domains, social issues, and tourism present the greatest risk of eliciting propaganda-tinged answers.

Table 4: Chinese Propaganda bias by topic and model–language combinations.

| Topic | R1 zh-CN | o3-mini-high zh-CN | R1 zh-TW | o3-mini-high zh-TW | R1 EN | o3-mini-high EN |
|---|---|---|---|---|---|---|
| Industrial Dynamics / Technology | **5.56 %** | 4.17 % | 1.85 % | 0.46 % | 0.00 % | 0.46 % |
| Culture / Arts / Entertainment | **7.41 %** | 4.94 % | 4.32 % | 3.09 % | 0.00 % | 0.00 % |

| | | | | | | |
|---|---|---|---|---|---|---|
| Public / Social Issues | **8.33 %** | 4.55 % | 3.03 % | 1.52 % | 0.76 % | 0.00 % |
| Judiciary / Crime / Accidents | 3.17 % | 0.79 % | 0.00 % | 0.79 % | 0.00 % | 0.00 % |
| Lifestyle / Daily Life | 2.38 % | 0.00 % | 0.00 % | 0.00 % | 0.00 % | 0.00 % |
| Economy / Finance / Investment | **7.84 %** | **9.80 %** | 2.94 % | 0.00 % | 0.00 % | 0.00 % |
| International Relations / Geopolitics | **23.33 %** | **16.67 %** | **8.89 %** | **5.56 %** | 0.00 % | 0.00 % |
| Sports / Competitions | 3.33 % | 1.67 % | 3.33 % | 1.67 % | 0.00 % | 0.00 % |
| Domestic Politics / Elections | 4.76 % | 3.57 % | 0.00 % | 1.19 % | 0.00 % | 1.19 % |
| Medical / Health | 1.85 % | 1.85 % | 0.00 % | 1.85 % | 0.00 % | 0.00 % |
| Travel / Tourism | **8.33 %** | **8.33 %** | 2.08 % | 4.17 % | 0.00 % | 0.00 % |

## 4.2 Anti-US Sentiment Bias

Below is the Anti-US proportions of 6 model-language pairs by topic:

- **DeepSeek-R1 in Simplified Chinese is labeled 60 times (5.00 %)**. Switching to Traditional Chinese cuts R1's labels to **29 (2.42 %)**, while switching to English further reduces the count to **5 (0.42 %)**.

- **o3-mini-high records 0 / 1,200 (0 %) in all three languages.**

The results show that **changing the Chinese script from Simplified to Traditional reduces DeepSeek R1's anti-US outputs by roughly half, and switching to English eliminates nearly all instances of anti-US sentiment bias.** Language choice—not question content—drives this bias modulation. Hence, **anti-US sentiment is confined to the PRC-aligned model**; the non-PRC system shows no such bias under equal testing conditions.

Table 5 summarises the proportions of 7,200 answers (1,200 per model-language pair) labeled by GPT-4o as exhibiting negative framing toward the United States, broken down by topic, model, and language:

- Anti-US bias clustered by topic: it explicitly surged in hard-power arenas—Domestic Politics / Elections and International Relations / Geopolitics, both exceeding 15 % in Simplified Chinese.

- In every other topic the rate stays below 5 %.

Across all topics and languages, DeepSeek-R1 consistently shows more anti-US bias than o3-mini-high, with political queries in Simplified Chinese posing the greatest risk of eliciting negative framing toward the United States.

Table 5: Anti-US bias by topic and model–language combination.

| **Topic** | **R1 zh-CN** | **o3-mini -high zh-CN** | **R1 zh-TW** | **o3-mini -high zh-TW** | **R1 EN** | **o3-mini -high EN** |
|---|---|---|---|---|---|---|
| Industrial Dynamics / Technology | 2.31 % | 0.00 % | 0.93 % | 0.00 % | 0.00 % | 0.00 % |
| Culture / Arts / Entertainment | 1.85 % | 0.00 % | 2.47 % | 0.00 % | 0.00 % | 0.00 % |
| Public / Social Issues | 3.03 % | 0.00 % | 4.55 % | 0.00 % | 0.00 % | 0.00 % |
| Judiciary / Crime / Accidents | 4.76 % | 0.00 % | 2.38 % | 0.00 % | 1.59 % | 0.00 % |
| Lifestyle / Daily Life | 0.00 % | 0.00 % | 0.00 % | 0.00 % | 0.00 % | 0.00 % |
| Economy / Finance / Investment | 3.92 % | 0.00 % | 2.94 % | 0.00 % | 0.00 % | 0.00 % |
| International Relations / Geopolitics | **15.56 %** | 0.00 % | 4.44 % | 0.00 % | 3.33 % | 0.00 % |
| Sports / Competitions | 1.67 % | 0.00 % | 1.67 % | 0.00 % | 0.00 % | 0.00 % |
| Domestic Politics / Elections | **23.81 %** | 0.00 % | **7.14 %** | 0.00 % | 0.00 % | 0.00 % |

| | | | | | | |
|---|---|---|---|---|---|---|
| Medical / Health | 3.70 % | 0.00 % | 0.00 % | 0.00 % | 0.00 % | 0.00 % |
| Travel / Tourism | 2.08 % | 0.00 % | 0.00 % | 0.00 % | 0.00 % | 0.00 % |

# 5. Conclusion

Our study delivers the first cross-topic, cross-language comparison of a PRC-aligned model (DeepSeek-R1) and a non-PRC peer (ChatGPT o3-mini-high) on Chinese-state propaganda and anti-US sentiment. Building on a motivation that hidden ideological leanings can sway civic attitudes, we contributed a 1,200-question corpus in Simplified Chinese, Traditional Chinese, and English; 7,200 answers spanning six model–language pairs; and a GPT-4o-plus-human rubric that reached near-perfect agreement with manual annotations ($\kappa \approx 0.81$–$0.91$). Results show that language alone reshapes bias: in Simplified Chinese, R1 embeds propaganda in **6.8 %** of answers and anti-US framing in **5.0 %**, compared with o3-mini-high's **4.8 %** propaganda and **0 %** anti-US. Switching R1 to Traditional Chinese halves both rates (≈ 2.4 %), while English all but eliminates them (≤ 0.4 %).

When it comes to topic-level analysis, Propaganda clusters in **International Relations / Geopolitics (23.3 %), Economy/Finance/Investment (7.8 %), Industrial Dynamics/Technology (5.6 %), Public/Social Issues (8.3 %), Culture/Arts/Entertainment (7.4 %)**, and **Travel/Tourism (8.3 %)** when queries are in Simplified Chinese. Anti-US sentiment is sharply concentrated in **Domestic Politics / Elections (23.8 %)** and **International Relations / Geopolitics (15.6 %)**, while staying below 5 % elsewhere. These language- and topic-specific spikes confirm that **DeepSeek-R1 functions as an "invisible loudspeaker" for PRC-aligned narratives—most powerfully in Simplified Chinese and politically sensitive domains—whereas o3-mini-high remains largely neutral under identical conditions.**