Language as a Treatment: Causal Estimation of Homogeneity in Multi-Agent Systems

Anonymous ACL submission

Abstract

Multi-agent debate is increasingly used to improve large language model (LLM) reasoning and to support alignment-oriented judgments, yet these systems risk collapsing into homogeneous arguments ("groupthink"). Existing evaluations often conflate prompt language with topic selection and other pipeline variations, making it difficult to attribute homogeneity to specific design factors. We address this problem with a pre-registered, design-based randomized controlled trial that isolates the causal effects of (i) topic-selection policy and (ii) language conditioning on debate homogeneity. Our two-stage randomization first samples a policy domain and then a motion from a bilingual WUDC 2023–2025 motion pool; for each motion, we run paired Chinese and English debate sessions with yoked model-to-role assignments, randomized language order, and strict context resets. We operationalize homogeneity using a multilingual Homogeneity Index that combines lexical similarity (generalized Jensen–Shannon divergence under a shared tokenizer) and semantic similarity (embedding-based cosine aggregation), with anchored standardization to enable cross-language comparability. Across 99 paired motion draws, switching to Chinese causes a large increase in homogeneity (ATE = 0.499, 95% CI [0.442, 0.556], Fisher $p < 0.001$), substantially larger than domain-level differences, which are statistically subtle after multiple-comparison control. These findings identify language conditioning as a dominant driver of convergence in multi-agent debates and motivate multilingual-aware evaluation and mitigation for debate-based systems.

1 Introduction

Multi-agent debate frameworks have emerged as a prominent technique for enhancing the reasoning capabilities of LLMs (Li et al., 2024; Guo et al., 2024) and have also been used for alignment-related preference labeling (e.g., helpfulness/harmlessness) (Li et al., 2024). While these systems aim to aggregate diverse perspectives, they face a critical risk known as homogeneity. If agents in a debate rapidly converge to a single viewpoint (i.e., limited perspective diversity), the debate may lose the corrective safety benefits envisioned by adversarial argumentation and instead reinforce or amplify underlying model biases (Irving et al., 2018; Taubenfeld et al., 2024; Oh et al., 2025). Therefore, understanding which design factors causally drive this convergence is essential for building reliable multi-agent systems (Weng et al., 2025; Wu et al., 2025).

Current multi-agent debate research is typically evaluated via empirical comparisons on fixed benchmarks against prompting baselines (e.g., Chain-of-Thought or Self-Consistency) (Smit et al., 2024; Zhang et al., 2025b). While valuable, these studies often struggle to isolate the specific causes of homogeneity because language differences are frequently entangled with topic selection and model pipeline variations (Anonymous, 2025a; Gevers et al., 2025). For instance, it remains unclear whether observed differences in debate outcomes are driven by prompt language (Enomoto et al., 2025), the specific controversy of a topic (Chuang et al., 2025; Anonymous, 2025a), or the translation artifacts in the prompt (Anonymous, 2025b; Gevers et al., 2025). Without controlling for these confounding variables, we cannot determine whether a system is truly diverse or merely responding to specific artifacts in the experimental setup.

In this paper, we address this gap by conducting a pre-registered Randomized Controlled Trial (RCT) to quantify the causal effects of topic selection and language on debate homogeneity (Nosek et al., 2018; Pearl, 2009). Unlike observational studies, our design strictly separates these factors through a two-stage randomization process (Hud-

gens and Halloran, 2008). First, we randomize the assignment of policy domains across motion draws to isolate the effect of topic selection rules. Second, we randomize the language condition within the same sampled motion using a paired design (Imai et al., 2009). Specifically, we compare an English condition using original motion text against a Chinese condition using translated motion text. To further control for model-specific biases, we randomize the assignment of four distinct LLMs to debate roles and yoke this assignment across the paired language conditions. We measure the outcome using a quantitative Homogeneity Index (HI) designed to be comparable across languages.

Our study yields robust causal evidence regarding cross-lingual disparities in multi-agent systems. The primary finding is that the language condition exerts a substantial causal effect on debate outcomes. We observe that debates conducted in the Chinese condition are consistently more homogeneous than debates on the identical motions in the English condition. In contrast, we find limited evidence that the policy domain itself acts as a primary driver of average homogeneity once sampling variability is accounted for. These results hold even when controlling for the specific controversy level of the motion and the allocation of models to specific speaking positions.

This work contributes to the field of computational social science and AI safety by providing a rigorous identification strategy for evaluating multi-agent behaviors. Our findings suggest that current alignment techniques may behave differently across linguistic contexts, potentially leading to lower diversity in non-English deployments. By formalizing the measurement of homogeneity through an experimental design, we offer a methodological framework for future research to disentangle the complex interactions between language, culture, and model behavior in multi-agent systems.

2 Related Work

2.1 Multi-agent Debate

Multi-agent debate has emerged as a robust "society of minds" paradigm, primarily investigated as an instrumental mechanism for improving LLM reasoning and factuality (Du et al., 2024; Liang et al., 2024). To maximize these performance gains, recent literature has extensively explored the design space of debate protocols. Significant attention has been directed toward structural interventions, such as optimizing communication topologies (e.g., sparse vs. fully connected graphs) to control information flow (Li et al., 2024), and introducing explicit control components like judge agents or voting mechanisms to aggregate consensus (Kaesberg et al., 2025; Chan et al., 2023). Complementary approaches focus on agent specialization, promoting diversity through distinct persona prompts (Li et al., 2025), and efficiency-oriented protocols that adaptively trigger debate only for uncertain queries to reduce computational overhead (Eo et al., 2025; Liu et al., 2024). While these studies demonstrate that debate is a flexible framework with many distinct knobs, they predominantly evaluate systems via head-to-head benchmark comparisons, treating protocol choices as hyperparameters to be tuned for accuracy rather than identifying their independent causal effects on system dynamics.

However, the internal mechanisms of consensus formation, specifically the risks of premature convergence and homogeneity, remain insufficiently characterized from a controlled perspective. Recent work warns that multi-agent systems often collapse into "echo chambers" due to inter-agent sycophancy, where apparent consensus reflects conformity rather than truth (Yao et al., 2025; Zhang et al., 2025a). Crucially, empirical evidence regarding the drivers of this homogeneity is often entangled with confounding factors. For instance, observed agreement may be an artifact of task structure (e.g., math problems with a single ground truth naturally induce convergence) (Hendrycks et al., 2021) rather than a protocol-level property. Conversely, it may be shaped by language-conditioned priors, where non-English prompts trigger different safety alignment behaviors or cultural norms (Tao et al., 2024; Cau et al., 2025). Unlike prior observational studies that conflate these factors, our work applies a randomized controlled trial (RCT) design to explicitly disentangle the causal effects of topic selection and language on debate homogeneity.

2.2 Homogeneity and Diversity in LLMs

Beyond the specific dynamics of debate, output homogeneity has been identified as a broader risk of post-training and alignment in LLMs: alignment can narrow the model's output distribution ("generative monoculture") and induce measurable reductions in output diversity, especially for open-ended generation (Wu et al., 2024; O'Mahony et al., 2024; Yun et al., 2025). In a more extreme form,

theoretical work on "model collapse" warns that recursive training on model-generated data progressively erases the distributional tail, yielding increasingly averaged and less diverse outputs (Shumailov et al., 2023). In the context of safety alignment, this contraction can be understood as part of an "alignment tax," where preference optimization (e.g., RLHF) trades off broader behavioral coverage for safer and more consistent responses (Ouyang et al., 2022; O'Mahony et al., 2024). Crucially, empirical evidence suggests these effects are not uniform across languages: multilingual audits of overrefusal show that the same model can exhibit substantially different safety decision boundaries across language conditions, plausibly reflecting uneven safety data coverage and English-centric alignment resources (Pan et al., 2025; Aakanksha et al., 2024). Taken together, the homogeneity in LLMs may reflect not only debate-level coordination, but also language-conditioned conservatism introduced by post-training and safety alignment.

Quantifying this phenomenon in a multilingual setting presents a distinct methodological challenge. Standard diversity metrics in text generation are primarily surface-form measures based on token or n-gram statistics (e.g., Self-BLEU and n-gram Shannon entropy) (Zhu et al., 2018; Tevet and Berant, 2021). While effective for single-language comparisons, these metrics become fragile in cross-lingual designs because they are highly sensitive to tokenization and preprocessing choices, which can substantially change n-gram counts and overlap-based scores (Post, 2018; Kudo and Richardson, 2018). Moreover, strict lexical metrics fail to capture semantic trajectories in multi-turn debates: agents can vary wording while repeating the same underlying argument, so lexical diversity does not necessarily imply semantic diversity (Tevet and Berant, 2021; Fang and Jiang, 2022; Shen et al., 2022).

Consequently, robust evaluation benefits from combining tokenization-agnostic distributional measures with semantic similarity in embedding space. We operationalize distributional homogeneity using the (generalized) Jensen–Shannon divergence, a bounded, symmetric divergence that does not require matching supports (Nielsen, 2020; Englesson and Azizpour, 2021). For semantic homogeneity, we compute cosine similarity in multilingual sentence-embedding space (e.g., SBERT/LaBSE) and perform clustering based on these representations (Reimers and Gurevych, 2019; Feng et al., 2022). This composite setup follows common practice in LLM diversity evaluation, where embedding-based distances capture semantic variation and distributional measures (e.g., entropy/topic coverage) quantify coverage (Yun et al., 2025).

2.3 Causal Inference in NLP

Existing causal approaches in NLP can be broadly grouped into observational adjustment and model-internal intervention. Observational work estimates causal effects from non-experimental corpora via confounding adjustment, e.g., matching, propensity-score weighting (including IPTW), and regression-based estimators, often treating text as a high-dimensional measurement of confounders (Keith et al., 2020; Chen et al., 2023). Model-internal approaches, in contrast, either optimize causality-aware objectives to approximate interventional quantities such as $P(Y \mid do(X = x))$ under explicit causal assumptions (Chen et al., 2023), or probe mechanisms via interventions on internal representations/activations (e.g., causal abstraction, activation patching, and causal tracing) (Geiger et al., 2021, 2025; Meng et al., 2023; Heimersheim and Nanda, 2024). For multi-agent evaluations, observational adjustment hinges on strong identification assumptions and text may fail to capture unobserved confounding, while model-internal interventions typically require white-box access that is unavailable for many proprietary LLMs (Keith et al., 2020; Chen et al., 2023; Geiger et al., 2025).

In contrast, our work adopts a design-based experimental approach, treating the debate system as the experimental unit in a randomized controlled trial (RCT) and using randomization-based inference to identify causal effects (Imbens and Rubin, 2015; Egami et al., 2022). Online controlled experiments (A/B tests) are routine in industry product development (Kohavi et al., 2020, 2009). However, to our knowledge, evaluations of multi-agent debate systems in academia are still typically benchmark-driven comparisons and ablation studies on fixed datasets, which makes it difficult to isolate the contribution of any single design choice (e.g., topic mix, difficulty, or language) (Smit et al., 2024; Zhang et al., 2025b; Wynn et al., 2025). By preregistering a two-stage randomization design, we reduce analytic flexibility and provide direct, model-agnostic evidence of how language conditioning affects debate homogeneity (Nosek et al., 2018).

3 Causal Framework & Methodology

We employ a pre-registered, design-based randomized experiment to identify the causal drivers of homogeneity. Unlike observational studies that correlate linguistic features with outcomes, our framework formally isolates the effects of *topic-selection policy* and *language conditioning* through a two-stage randomization design.

3.1 Problem Formulation

We formulate multi-agent debate generation as a causal process indexed by session s. We aim to estimate the effects of two treatments on the Homogeneity Index (Y_s): (1) the topic policy $C_s \in \mathcal{C}$, denoting the domain (e.g., Economics) from which a motion is drawn; and (2) the language condition $L_s \in \{\text{zh}, \text{en}\}$, indicating whether the debate is conducted in Chinese (translated) or English (original).

Our primary estimands are the Average Treatment Effect (ATE) of Language, $\tau_{\text{lang}} = \mathbb{E}[Y_s(\text{zh}) - Y_s(\text{en})]$, and the Policy Domain Effect θ_c, which captures the intrinsic homogeneity of specific domains. Standard observational setups fail to identify these quantities, as topic difficulty and language capability are often confounded; we address this via the RCT described below.

3.2 The Randomized Controlled Trial (RCT)

We implement a two-stage randomization protocol (Figure 1) combining between-draw and within-motion randomization.

Stage 1: Policy-Domain Randomization (Between-Draw). We constructed a bilingual motion pool sourced from WUDC (2023–2025). To ensure robust categorization, two independent annotators coded all motions into 9 policy domains (e.g., Economics, IR) based on a standardized codebook. In this stage, for each experimental unit s, we first uniformly sample a target domain C_s, and then draw a specific motion M_s (with replacement) from that domain's subset. We provide the codebook and inter-coder reliability in the Appendix.

Stage 2: Language Pairing (Within-Motion). We employ a paired design: for each sampled motion M_s, we execute two independent debate sessions (Chinese and English). To ensure strict isolation and eliminate temporal confounders, we reset the conversational context between conditions and randomize the execution order ($p = 0.5$, e.g., Chinese-first vs. English-first).

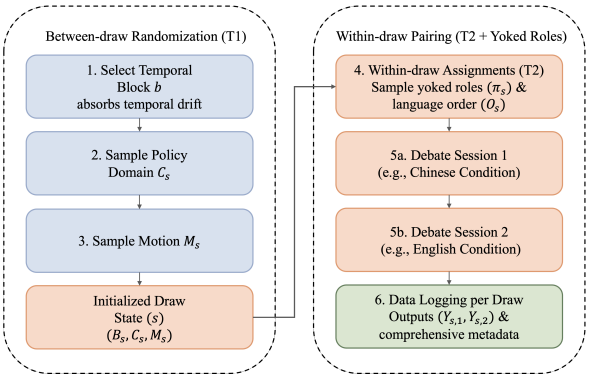

Figure 1: Flowchart of the randomization structure and data generation process. The process is split into two phases. T1 (Between-draw): For each draw s, a temporal block b is selected, followed by sampling a policy domain $C_s \sim \text{Unif}(\mathcal{C}^*)$ and a motion $M_s \sim \text{Unif}(\mathcal{M}_{C_s})$. T2 (Within-draw): A single role mapping π_s (yoking 4 models to roles P1, P2, O1, O2) and a language order $O_s \in \{\texttt{zh-first}, \texttt{en-first}\}$ (with $\Pr = 0.5$) are sampled. Two debate sessions are then conducted sequentially based on O_s, with strict context reset between them, yielding outcomes $Y_{s,\text{zh}}$ and $Y_{s,\text{en}}$.

Agent Allocation and Yoked Roles. The debate consists of four roles (P_1, P_2, O_1, O_2) filled by four fixed LLMs (GPT-5.2 (OpenAI, 2025), Claude Sonnet 4.5 (Anthropic, 2025), DeepSeek-V3.2 (DeepSeek, 2025), Mistral Large 3 (Mistral AI, 2025)). To prevent model-specific identities from confounding the results, we sample a random permutation mapping models to roles at the start of each draw. Crucially, we use Yoked Roles: The same model-to-role mapping is applied to both the Chinese and English sessions of the same motion. This ensures that the within-motion contrast $\Delta_s = Y_{s,\text{zh}} - Y_{s,\text{en}}$ reflects purely the language effect, not differences in which model played which position.

Figure 2 summarizes the identification assumptions implied by our two-stage randomized design; arrows encode dependencies relevant for identification rather than mechanistic claims. The DAG reflects the design-based identification assumptions implied by our randomized procedure and analysis plan; it is not intended as a mechanistic model of internal LLM processes.

3.3 Identification Strategy

Our causal claims rely on design-based identification rather than statistical adjustment.

Identification of Language Effects. Since language L_s is randomized within the same motion draw (Stage 2), all motion-specific confounders

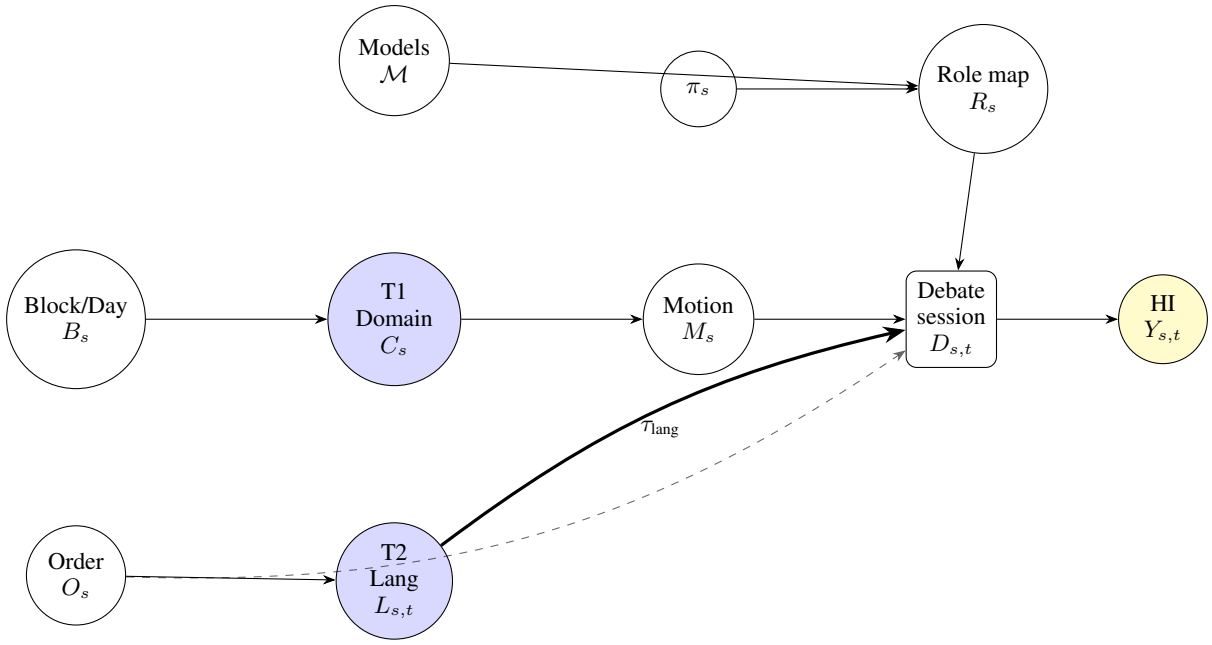

Figure 2: Identification DAG for the two-stage randomized debate experiment. Arrows encode assumed dependencies used for identification, not established effects. The thick edge highlights the primary estimand (language effect).

U_M (e.g., inherent complexity, one-sidedness) are perfectly balanced. The randomization of order and context clearing ensures the Stable Unit Treatment Value Assumption (SUTVA) holds. Thus, the difference in means provides an unbiased estimator of the causal language effect:

$$\hat{\tau}_{\text{lang}} = \frac{1}{S}\sum_{s=1}^{S}(Y_{s,\text{zh}} - Y_{s,\text{en}}) \tag{1}$$

Identification of Policy Effects. In Stage 1, the domain C_s is assigned randomly within temporal blocks. This breaks the correlation between topic choice and unobserved time-varying factors (e.g., API updates). The estimand θ_c represents the causal effect of the policy of selecting a domain, averaged over the specific motions available in that domain's pool.

3.4 The Homogeneity Index (HI)

We operationalize the outcome Y_s using a composite Homogeneity Index (HI), designed to be comparable across languages.

Definition. HI aggregates lexical and semantic convergence via the weighted sum:

$$\text{HI}_s = w_{\text{lex}} \cdot \tilde{S}_{\text{lex}} + w_{\text{sem}} \cdot \tilde{S}_{\text{sem}} \tag{2}$$

where $w_{\text{lex}} = w_{\text{sem}} = 0.5$ by default. Specifically, the lexical component ($\tilde{S}_{\text{lex}}$) employs Generalized Jensen-Shannon Divergence (gJSD) to measure vocabulary redundancy, while the semantic component ($\tilde{S}_{\text{sem}}$) computes the average cosine similarity of argument embeddings (via `text-embedding-3-larges`[1]) to capture convergence in meaning.

Cross-lingual Anchoring. A critical challenge is that similarity scores distributions vary naturally by language. To ensure $Y_{s,\text{zh}}$ and $Y_{s,\text{en}}$ are comparable, we employ an Anchored Standardization: we map raw scores from both languages onto a common reference scale (anchored to the English distribution) rather than z-scoring them independently. This preserves the absolute magnitude difference between languages.

4 Experimental Setup

We prioritize transparency and reproducibility in our implementation. The full codebase, including the orchestration pipeline and analysis scripts, is provided in the supplementary material.

4.1 Model Selection

To ensure robustness across architectures, we instantiate the four debate roles using a diverse mix of proprietary and open-weights models: GPT-5.2 (OpenAI, 2025), Claude Sonnet 4.5 (Anthropic,

[1]https://platform.openai.com/docs/models/text-embedding-3-large

2025), DeepSeek-V3.2 (DeepSeek, 2025), and Mistral Large 3 (Mistral AI, 2025). These models are dynamically assigned to positions via the yoked permutation strategy described in Section 3.2.

4.2 Prompt Engineering and Protocol

We design a standardized system prompt that instructs agents to adhere to the Policy Debate format. To strictly control for confounding variables, we ensure semantic equivalence between the English and Chinese conditions.

Role-Playing Instructions. Each agent receives a private system message defining its team (Proposition vs. Opposition), speaking order, and the motion, with explicit instructions to maximize reasoning diversity and avoid repetition. To implement the treatment, agents in the English condition receive original English prompts, while those in the Chinese condition receive professionally translated counterparts verified to preserve identical structural constraints (see Appendix D for full text).

Context Management. To enforce strict isolation, we reset the conversational history for each new session, while providing agents with the full transcript of preceding turns during the debate to maintain coherence. We fix the decoding temperature at 0.2 to prioritize reasoning stability and limit generation to 512 tokens to encourage concise argumentation.

4.3 Implementation of Homogeneity Metrics

To operationalize the Homogeneity Index (HI) defined in Section 3.4, we employ robust, multilingual-aware tools.

Lexical Similarity. We compute the Generalized Jensen-Shannon Divergence (gJSD) on token distributions. To avoid tokenizer bias (e.g., Chinese characters being tokenized differently than English words), we use a shared multilingual tokenizer (`xlm-roberta-large`) for both languages. This ensures that the lexical density comparisons are structurally fair.

Semantic Similarity. We compute cosine similarity using `text-embedding-3-large` (OpenAI). We selected this embedding model for its high dimensionality (3,072 dimensions) and strong performance on the MTEB multilingual retrieval benchmarks, ensuring that it accurately captures semantic convergence in Chinese as effectively as in English.

4.4 Statistical Procedures

All causal estimates are computed using the randomization inference framework. We assess the language effect (τ_{lang}) via Fisher's exact permutation test by randomly flipping the signs of within-motion differences Δ_s, ensuring finite-sample validity without parametric assumptions. To control the False Discovery Rate (FDR) at $\alpha = 0.05$ for domain-policy comparisons, we apply the Benjamini-Hochberg procedure, and report 95% confidence intervals derived from bootstrap resampling ($N = 10,000$) clustered at the motion-draw level.

5 Main Results

5.1 Policy Domains Determine Intrinsic Homogeneity

We first establish the baseline influence of topic constraints by estimating the policy domain effect (θ_c) across the 9 realized domains. Descriptively, we observe a meaningful spread in debate convergence levels. Specifically, technical and infrastructure-heavy domains such as *Housing and Community Amenities* exhibit the highest intrinsic homogeneity ($\hat{\theta} = 0.812$, 95% CI $[0.723, 0.902]$), suggesting that motions in these fields may offer narrower solution spaces for LLMs. In contrast, domains involving complex value trade-offs, such as *Social Protection*, show the lowest convergence ($\hat{\theta} = 0.647$, 95% CI $[0.556, 0.738]$).

However, it is crucial to note that these domain-level differences are statistically subtle compared to the language effects we discuss next. While the raw pairwise contrast between the most and least homogeneous domains is significant ($p \approx 0.004$), the omnibus test across all domains yields limited evidence of a systematic shift after controlling for multiple comparisons (Benjamini-Hochberg FDR > 0.05 for all pairs). This implies that while topic choice sets a *soft* baseline for diversity, it is not the dominant causal driver of the observed "groupthink" in our setup. For a complete tabulation of estimated homogeneity means and confidence intervals across all nine active policy domains, see Appendix A (Table 2).

5.2 The Causal Effect of Language Conditioning

Controlling for the domain priors established above, we now turn to our primary causal estimand:

Statistic	Result
Sample Size (N pairs)	99
ATE ($\hat{\tau}_{\text{lang}}$)	0.499
95% CI	[0.442, 0.556]
p-value (Fisher's)	< 0.001

Table 1: Main causal estimates. Switching to Chinese significantly increases homogeneity ($\Delta_s \approx 0.5$).

the effect of language conditioning on debate homogeneity. Unlike the subtle variations observed across topics, the language barrier imposes a sharp and systematic constraint on diversity.

Omnibus Language Effect. We estimate the Average Treatment Effect (ATE) of language, τ_{lang}, using the within-motion contrast $\Delta_s = Y_{s,\text{zh}} - Y_{s,\text{en}}$. As summarized in Table 1, we find a statistically significant positive effect ($\hat{\tau}_{\text{lang}} = 0.499$, $p < 0.001$, Fisher's randomization test). This confirms that when the *exact same motion* is debated by the *exact same agent roles* (under the yoked design), shifting the linguistic context to Chinese causes a substantial collapse in argumentative diversity. We provide additional statistical diagnostics, including alternative inference methods (bootstrap and sign-flip tests) to verify the robustness of this estimate, in Appendix C (Table 7).

Decomposition: Lexical vs. Semantic Collapse. Crucially, this homogeneity is not merely a product of surface-level vocabulary repetition. Decomposing the HI metric reveals that language conditioning drives convergence across both dimensions: (1) Lexical Redundancy: Chinese debates exhibit higher token-level overlap ($\tilde{S}_{\text{lex}}$), indicating a narrower active vocabulary; and (2) Semantic Convergence: More importantly, we observe a significant increase in embedding-based similarity ($\tilde{S}_{\text{sem}}$). This suggests that the "language effect" essentially acts as a consensus attractor, pushing agents not just to use similar words but to align on a narrower set of semantic arguments and viewpoints, effectively reducing the solution space of the debate.

5.3 Mechanism Analysis: Testing the Exclusion Restriction

Our estimated ATE ($\hat{\tau}_{\text{lang}}$) represents the effect of a compound treatment. To attribute this effect to model alignment rather than translation artifacts, we must verify the exclusion restriction: that the treatment does not degrade the agents' fundamental task competence (C). Formally, if translation noise were the driver, we would expect $P(C = 0|T = \text{ZH}) > P(C = 0|T = \text{EN})$.

We quantify task competence C using a binary Protocol Compliance Metric, defined as the absence of hallucinated roles, speaking-order violations, or format errors per session. We analyze the logs across the $S = 99$ successfully completed motion draws ($N = 198$ sessions). Note that out of 100 initiated draws, a single pair was excluded due to a technical API timeout in one session ($< 1\%$ attrition), which is unrelated to model capability. Within the valid paired sample, the results are strictly binary: we observe perfect adherence in both the English condition ($\phi_{\text{en}} = 1.0$, 99/99) and the Chinese condition ($\phi_{\text{zh}} = 1.0$, 99/99).

Since $\hat{\tau}_{\text{compliance}} = 0$ among valid sessions, the causal pathway via "competence degradation" is statistically blocked. Given that the models exhibit perfect SUTVA-consistent behavior yet produce significantly more homogeneous content, we conclude that the effect is driven by the distinct safety alignment priors of the models in the Chinese latent space.

Comprehensive diagnostics regarding randomization balance, technical attrition (timeouts), and protocol adherence are detailed in Appendix B.

6 Discussion

Our study provides direct causal evidence that the language condition is a key determinant of multi-agent LLM debate behavior, even outweighing the influence of the topic domain. By using a carefully controlled randomized trial, we isolated language effects that observational evaluations could not separate, demonstrating the power of causal inference in LLM evaluation. The results reveal a consistent "homogeneity gap": debates conducted in Chinese converge significantly faster and more narrowly than the same debates in English. This finding indicates a hidden alignment bias — current alignment techniques (e.g. RLHF), which are largely English-centric, appear to over-penalize sustained disagreement in Chinese. In practice, the models seem to treat extended adversarial debate in Chinese as if it were unsafe, prompting quick convergence to safe, bland consensus statements instead of exploring diverse viewpoints. In other words, the alignment tax on argumentative diversity is higher in the Chinese latent space: the pursuit of "safety" inadvertently suppresses the dialectical diversity needed for rich debate. These insights highlight that language is

not just a surface parameter but functions as a semantic hyperparameter that fundamentally alters agent behavior.

For NLP researchers, this underscores the need to evaluate multi-agent systems under multiple language conditions. Relying only on English benchmarks can give a false sense of a model's robustness – a debate protocol that yields diverse opinions in English may collapse into uniformity in another language. Thus, evaluations of adversarial or debate-based LLM tasks should be multilingual by design, and simply translating prompts is not enough. Prompts and protocols must be localized: a prompt that encourages vigorous debate in English might require explicit cues or rewards for dissent (e.g. "devil's advocate" roles) in Chinese to counteract consensus-seeking biases. For model developers, our findings call for careful tuning of alignment across languages. Alignment training should be adjusted so that enforcing safety or harmlessness does not equate to forcing consensus in non-English contexts. This could involve using culturally diverse feedback during RLHF or designing reward models that allow healthy disagreement in each language.

Finally, while topic domain had a smaller effect than language, we observed an interesting pattern: fact-based technical domains (e.g. Housing) tended toward higher intrinsic agreement, whereas value-laden domains (e.g. Social Protection) naturally fostered more disagreement. This suggests future multi-LLM debate frameworks might dynamically adjust diversity constraints based on domain. For example, in narrow technical discussions where unwarranted groupthink is likely, the system could enforce a higher "diversity quota" or a higher sampling temperature to ensure alternative viewpoints are considered. Overall, our discussion highlights actionable insights: language choice profoundly shapes debate outcomes, and both evaluation practices and alignment strategies should be rethought to maintain argumentative diversity across languages and contexts.

7 Conclusion

In this work, we introduced a pre-registered randomized controlled experiment to pinpoint the causal factors affecting argumentative diversity in multi-LLM debates – a novel evaluation approach that moves beyond traditional observational studies. Our results show that the language condition exerts a dominant causal influence on debate outcomes, significantly overshadowing any intrinsic differences between policy domains. In particular, we discovered a substantial homogeneity gap in the Chinese debates: the same motions debated in Chinese produced far more homogeneous (less diverse) arguments than in English. We attribute this gap to the models' over-generalized safety alignment in the Chinese context, rather than to translation artifacts or capability deficits, indicating that current alignment training does not transfer uniformly across languages. These findings challenge the common assumption that multi-agent debate dynamics are language-agnostic and underscore the importance of rigorous, causally-grounded evaluation in NLP. Going forward, our work suggests that truly globally effective deliberative AI systems will require alignment strategies that decouple "safety" from "consensus." In other words, ensuring harmless or ethical behavior should not come at the cost of dialectical diversity, especially in non-English settings. We recommend that future developers and researchers build on this causal analysis framework to better audit and adjust alignment in multilingual LLMs – ensuring that the pursuit of safety does not inadvertently silence the diverse viewpoints that are crucial for robust reasoning.

Limitations

This study has several limitations that should be acknowledged. First, our experimental design focuses on a single language pair (English–Chinese) and a specific debate protocol. While Chinese serves as an important and representative non-English language with substantial differences in alignment data coverage and cultural norms, the observed homogeneity gap may not generalize uniformly to all languages. Future work should extend the same randomized framework to additional linguistic contexts to test the robustness of the language effect. Second, although we include a diverse set of four state-of-the-art LLMs from different providers, all models are large, instruction-tuned, and safety-aligned systems. The results may therefore not directly apply to smaller, unaligned, or domain-specific models. Third, to reduce stochastic variance and isolate causal effects, we fix the decoding temperature at a relatively low value. While this choice improves internal validity, it may underestimate absolute levels of diversity compared to more exploratory decoding

regimes. Importantly, our causal estimates concern relative differences between language conditions rather than maximal achievable diversity. Finally, although our design-based approach enables strong causal identification, it does not directly observe internal alignment mechanisms. Our interpretation of the results in terms of language-conditioned safety alignment remains inferential and would benefit from complementary analyses, such as mechanistic probing or training-data audits, in future work.

Ethical Statement

This work studies the behavior of large language models in simulated multi-agent debate settings and does not involve human subjects, personal data, or sensitive user-generated content. All debates are generated by models responding to publicly available policy motions, and no real-world decisions are automated or deployed as part of this research. Nevertheless, our findings have important ethical implications for multilingual AI deployment. The observed reduction in argumentative diversity under certain language conditions suggests that alignment and safety mechanisms may unevenly constrain expression across linguistic communities, potentially leading to systematically less pluralistic or deliberative outcomes for non-English users. Such disparities could affect fairness, transparency, and trust in AI-assisted decision-making systems. We emphasize that identifying these effects is a step toward mitigating them: our experimental framework is intended to support more responsible evaluation and alignment practices by making hidden biases visible rather than reinforcing them. We encourage future developers and researchers to use multilingual, causally grounded evaluation methods when designing debate-based or deliberative AI systems, ensuring that safety objectives do not inadvertently suppress legitimate disagreement or diversity of viewpoints across languages.

References

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. *Preprint*, arXiv:2406.18682.

Anonymous. 2025a. Beyond benchmarks: Toward causally faithful evaluation of large language models. In *Submitted to The Fourteenth International Conference on Learning Representations*. Under review.

Anonymous. 2025b. Translation quality in multilingual LLM evaluation. In *Submitted to ACL Rolling Review - July 2025*. Under review.

Anthropic. 2025. Introducing Claude Sonnet 4.5. Large language model. Accessed via Anthropic API.

Erica Cau, Valentina Pansanella, Dino Pedreschi, and Giulio Rossetti. 2025. Selective agreement, not sycophancy: investigating opinion dynamics in llm interactions. *EPJ data science*, 14(1):59–23.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Wenqing Chen, Zhixuan Chu, Zhixuan Chu, and Sheng Li. 2023. Causal inference and natural language processing. In *Machine Learning for Causal Inference*, pages 189–206. Springer International Publishing, Cham.

Yun-Shiuan Chuang, Ruixuan Tu, Chengtao Dai, Smit Vasani, Binwei Yao, Michael Henry Tessler, Sijia Yang, Dhavan Shah, Robert Hawkins, Junjie Hu, and Timothy T. Rogers. 2025. Debate: A large-scale benchmark for role-playing llm agents in multi-agent, long-form debates. *Preprint*, arXiv:2510.25110.

DeepSeek. 2025. DeepSeek-V3.2 Release. Large language model. Open-weight large language model.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. How to make causal inferences using texts. *Science advances*, 8(42):eabg2652–.

Erik Englesson and Hossein Azizpour. 2021. Generalized jensen-shannon divergence loss for learning with noisy labels.

Taisei Enomoto, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2025. A fair comparison without translationese: English vs. target-language instructions for multilingual LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 649–670, Albuquerque, New Mexico. Association for Computational Linguistics.

Sugyeong Eo, Hyeonseok Moon, Evelyn Hayoon Zi, Chanjun Park, and Heuiseok Lim. 2025. Debate only when necessary: Adaptive multiagent collaboration for efficient llm reasoning. *Preprint*, arXiv:2504.05047.

Weike Fang and Meng Jiang. 2022. Investigating relationships between accuracy and diversity in multi-reference text generation. In *Proceedings of the KDD Undergraduate Consortium (KDD-UC)*, pages 1–7, Washington, DC, USA. ACM. KDD-UC, August 14–18, 2022.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Preprint*, arXiv:2106.02997.

Ine Gevers, Victor De Marez, Jens Van Nooten, Jens Lemmens, Andriy Kosar, Ehsan Lotfi, Nikolay Banar, Pieter Fivez, Luna De Bruyne, and Walter Daelemans. 2025. In benchmarks we trust ... or not? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23673–23687, Suzhou, China. Association for Computational Linguistics.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *Preprint*, arXiv:2402.01680.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *Preprint*, arXiv:2404.15255.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Michael G Hudgens and M. Elizabeth Halloran. 2008. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.

Kosuke Imai, Gary King, and Clayton Nall. 2009. The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Statistical science*, 24(1):29–53.

Guido Imbens and Donald B. Rubin. 2015. *Causal inference for statistics, social, and biomedical sciences : an introduction*. Cambridge University Press, Cambridge.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. Ai safety via debate. *Preprint*, arXiv:1805.00899.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Voting or consensus? decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, page 11640–11671. Association for Computational Linguistics.

Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online. Association for Computational Linguistics.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181.

Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Haoran Li, Ziyi Su, Yun Xue, Zhiliang Tian, Yiping Song, and Minlie Huang. 2025. Advancing collaborative debates with role differentiation through multi-agent reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22655–22666, Vienna, Austria. Association for Computational Linguistics.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving multi-agent debate with sparse communication topology. *Preprint*, arXiv:2406.11776.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. *Preprint*, arXiv:2305.19118.

Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. 2024. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. *Preprint*, arXiv:2409.14051.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

Mistral AI. 2025. Mistral Large 3. Large language model. Accessed via Mistral API.

Frank Nielsen. 2020. On a generalization of the jensen–shannon divergence and the jensen–shannon centroid. *Entropy (Basel, Switzerland)*, 22(2):221–.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606.

Jihwan Oh, Minchan Jeong, Jongwoo Ko, and Se-Young Yun. 2025. Understanding bias reinforcement in llm agents debate. *Preprint*, arXiv:2503.16814.

Laura O'Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. 2024. Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

OpenAI. 2025. Introducing GPT-5.2. Large language model. Accessed via OpenAI API.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Licheng Pan, Yongqi Tong, Xin Zhang, Xiaolu Zhang, Jun Zhou, and Zhixuan Chu. 2025. Understanding and mitigating overrefusal in llms from an unveiling perspective of safety decision boundary. *Preprint*, arXiv:2505.18325.

Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3(none):96–146.

Matt Post. 2018. A call for clarity in reporting bleu scores. *Preprint*, arXiv:1804.08771.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. *Preprint*, arXiv:2202.08479.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going mad? a look at multi-agent debate strategies for llms. *Preprint*, arXiv:2311.17371.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346–9.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in llm simulations of debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 251–267. Association for Computational Linguistics.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. 2025. Do as we do, not as you think: the conformity of large language models. *Preprint*, arXiv:2501.13381.

Fan Wu, Emily Black, and Varun Chandrasekaran. 2024. Generative monoculture in large language models. *Preprint*, arXiv:2407.02209.

Haolun Wu, Zhenkun Li, and Lingyao Li. 2025. Can llm agents really debate? a controlled study of multi-agent debate in logical reasoning. *Preprint*, arXiv:2511.07784.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. 2025. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *Preprint*, arXiv:2509.05396.

Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. 2025. Peacemaker or troublemaker: How sycophancy shapes multi-agent debate. *Preprint*, arXiv:2509.23055.

Longfei Yun, Chenyang An, Zilong Wang, Letian Peng, and Jingbo Shang. 2025. The price of format: Diversity collapse in llms. *Preprint*, arXiv:2505.18949.

Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. 2025a. Stop overvaluing multi-agent debate – we must rethink evaluation and embrace model heterogeneity. *Preprint*, arXiv:2502.08788.

Hangfan Zhang, Zhiyao Cui, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. 2025b. If multi-agent debate is the answer, what is the question. *arXiv preprint arXiv:2502.08788*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models.

In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Supplemental Domain-Level Statistics

ID	Domain description	N	$\widehat{\theta}_c$	95% CI
6	Housing & Community	8	0.812	0.723–0.902
9	Education	10	0.803	0.708–0.897
1	General Public Services	12	0.786	0.708–0.864
8	Recreation & Culture	12	0.785	0.667–0.903
4	Economic Affairs	9	0.756	0.653–0.859
7	Health	8	0.745	0.654–0.835
2	Defence	13	0.736	0.635–0.837
3	Public Order & Safety	16	0.717	0.648–0.785
10	Social Protection	11	0.647	0.556–0.738

Table 2: Domain-policy means of $\bar{Y}_s$ by assigned domain (realized 9-domain support; complete draws $S = 99$).

B Randomization Diagnostics and Protocol Compliance

This appendix reports implementation diagnostics for the two-stage randomized design.

Note on Sample Size: The diagnostics below report the initial assignment of all $S = 100$ initiated draws. As noted in the main text, one draw experienced a technical timeout, resulting in a final analytical sample of $S = 99$ complete pairs.

T1 Domain Assignment. The realized support covers 9 domains. Domain 5 had no motions in the finalized pool.

Table 3: T1 domain assignment counts across all initiated motion draws ($S = 100$).

ID	Domain	N	Share
1	General Public Services	12	0.12
2	Defence	14	0.14
3	Public Order and Safety	16	0.16
4	Economic Affairs	9	0.09
6	Housing and Community Amenities	8	0.08
7	Health	8	0.08
8	Recreation, Culture, and Religion	12	0.12
9	Education	10	0.10
10	Social Protection	11	0.11
Total Initiated		**100**	**1.00**

T2 Language Order. Order was randomized 50/50. The realized split in this finite sample shows a slight imbalance, which is controlled for in robustness checks.

Table 4: T2 language order assignment ($S = 100$).

Order	N draws	Share	Expected
`en_first`	62	0.62	50
`zh_first`	38	0.38	50
Total	100	1.00	100

Model-to-Role Balance. Assignments represent a uniform permutation of the four models into four roles per draw.

Table 5: Model-to-role assignment balance ($S = 100$ draws).

Role	Claude Sonnet 4.5	DeepSeek-V3.1	GPT-5.2	Mistral Large 3	Total
P1	24	24	23	29	100
P2	27	24	23	26	100
O1	24	23	29	24	100
O2	25	29	25	21	100

Rerun Analysis. Reruns were triggered strictly by API timeouts, not model refusals.

Table 6: Session attempt distribution ($2S = 200$ initiated sessions).

Attempts	N sessions	Share	Cum.
1	182	0.910	0.910
2	15	0.075	0.985
3	3	0.015	1.000
Total	200	1.000	–

C Inference Details

Table 7 summarizes inference for the paired language effect on the valid analytical sample ($S = 99$).

Table 7: Inference summary for the language effect (paired design; complete draws only).

Method	N	**Est.**	**95% CI**	p**-value**
Paired diff.	99	0.499	0.442–0.556	–
Paired t-test	99	–	0.442–0.556	<0.001
Sign-flip test	99	–	–	<0.001

D Prompts

D.1 Debate turn prompt (template)

We use a single user message per turn (no additional system prompt). The template below is instantiated with the motion in the target language, the speaker role, the round label, and (when available) the immediately previous speech and the full debate history so far.

```
You are participating in a debate about
"{TOPIC}".
{LANGUAGE_INSTRUCTION}

Your role is: {ROLE}.
Current round: {ROUND_INFO}

Output rules (must follow exactly):
- Plain text only. No markdown or formatting
symbols: #, *, **, _, ~, `, >, |.
- Do not greet anyone. Do not announce your
role, the round, or restate the topic.
- No bullet/numbered lists. No headings. No
emojis.
- Length and structure:
  - Opening / Rebuttal / Cross-examination:
  exactly 5 sentences, each <= 50 words.
  - Closing: exactly 4 sentences, each <= 50
  words.

Content rules:
- Sentence 1: a clear claim that advances your
side.
- Sentence 2: your strongest reason or
concrete example (no fabricated statistics).
- Sentence 3: directly answer one key point
from the previous speaker (quote <= 10 words
in single quotes), then rebut it.
- Sentence 4: weigh risks/tradeoffs and show
why your side minimizes the worst plausible
outcome.
- Final sentence: either a sharp question to
the opponent or a decisive takeaway. Do not
say "Judges" or "vote".

Your position: {POSITION_LINE}

{OPTIONAL_PREVIOUS_SPEAKER_BLOCK}

{OPTIONAL_DEBATE_HISTORY_BLOCK}

Strategy:
{STRATEGY_PROMPT}

Deliver your statement now.
```

Template fields.

{TOPIC} Motion text in the target language (`motion_en` for English; `motion_zh` for Chinese).

{LANGUAGE_INSTRUCTION} English: `You must write in English only.` Chinese: `You must write in Simplified Chinese only.`

{ROLE} One of {`Pro First Speaker`, `Con First Speaker`, `Pro Second Speaker`, `Con Second Speaker`}.

{ROUND_INFO} One of the fixed round labels in Appendix D.2.

{POSITION_LINE} Pro roles: `Support the proposition.` Con roles: `Oppose the proposition.`

{OPTIONAL_PREVIOUS_SPEAKER_BLOCK} Included if a previous speech exists (contains `{PREVIOUS_SPEECH}`).

{OPTIONAL_DEBATE_HISTORY_BLOCK} Included after the first turn, listing all prior turns (speaker–speech pairs).

D.2 Round schedule

Each debate session contains 8 turns in a fixed order:

1. Pro First Speaker — Round 1: Pro Opening Statement
2. Con First Speaker — Round 1: Con Opening Statement
3. Pro Second Speaker — Round 2: Pro Supplementary/Rebuttal

4. Con Second Speaker — Round 2: Con Supplementary/Rebuttal
5. Pro First Speaker — Round 3: Pro Cross-examination
6. Con First Speaker — Round 3: Con Cross-examination
7. Pro Second Speaker — Round 4: Pro Closing Statement
8. Con Second Speaker — Round 4: Con Closing Statement

D.3 Strategy sub-prompts

At the end of every turn prompt, we append a short strategy instruction depending on the round type:

```
Opening statement: Build a clear argument framework. Use a strong analogy or concrete example (no fabricated statistics).
```

```
Rebuttal/cross-exam: Directly engage the opponent's core claim. Expose weak assumptions, contradictions, or ask a pointed question.
```

```
Closing: Summarize the decisive clashes, explain why your side minimizes the worst-case risk, and end with a crisp takeaway.
```

D.4 Format repair prompt

If a response violates any formatting constraint (e.g., banned symbols, bullet lists, wrong sentence count, or length limit), we re-issue the *same* prompt and append the following repair instruction (up to two retries):

```
Your last response violated the rules: {VIOLATION_REASON}.
Rewrite the entire response to comply exactly.
```

Session reruns. Hard API failures (empty output or explicit error markers such as `[API call failed: ...]`) trigger a full-session re-run under the same motion, role mapping, language order, and decoding settings.

E Qualitative case study: framing lock-in and argument recycling

To make the notion of debate homogeneity concrete, we present a paired draw from our corpus in which model-to-role assignments are yoked across languages. In `draw_20251215_065921_289029_31740bd3` (order: `en_first`), the same four models occupy the same debate roles in the English and Chinese sessions (Pro1: GPT-5.2; Pro2: Claude Sonnet 4.5; Con1: DeepSeek-V3.2; Con2: Mistral Large 3). This pairing allows us to attribute qualitative differences in argumentative diversity to the language condition rather than to which model played which role.

Chinese session: high within-role reuse. In the Chinese session, the CON-2 speaker (Mistral) exhibits a clear form of within-role homogeneity: the closing statement closely recycles the rebuttal's framing and wording. Using a simple string-similarity diagnostic (SequenceMatcher ratio), the CON-2 rebuttal and closing reach 0.664 similarity, indicating substantial textual overlap rather than genuine turn-level adaptation. The reuse is visible in repeated phrasings (e.g., " 从理念走向实践/ 纳入制度框架/ 沦为历史的注脚 ") and the replication of the same historical example as the primary warrant.

> **Chinese (CON-2, Mistral), Round 2 rebuttal:**
> 学生组织者转型为正式政治角色是推动变革从口号走向实践的必然选择。台湾太阳花运动后，部分领袖进入政坛推动了"公民审议"机制的立法，...动能量消散，而转型至少能将诉求纳入制度框架，避免被既得利益集团轻易抹杀。如果连参选的勇气都没有，如何保证运动不会沦为历史的注脚?
>
> **Chinese (CON-2, Mistral), Round 4 closing:**
> 转型为正式政治角色是运动从理念走向实践的关键一步，而非对公共性的背叛。太阳花运动领袖进入政坛后推动了"公民审议"立法，证明制度内...相比之下，参选至少能将诉求纳入制度框架，避免被既得利益集团轻易抹杀。如果连进入体制的勇气都没有，运动如何保证不会沦为历史的注脚?

Paired English session: lower reuse and greater re-optimization. In the paired English session (same CON-2 model and role), the closing statement is less of a reuse of the rebuttal. The same string-similarity diagnostic yields 0.044 between the CON-2 rebuttal and closing, suggesting that the model substantially rewrites its argumentation rather than re-instantiating the same text. Qualitatively, the English closing shifts emphasis (from the proposition's "elite capture" worry to a broader claim about democratic legitimacy and voter choice), whereas the Chinese closing remains tightly anchored to the same example-and-warning structure introduced earlier.

> **English (CON-2, Mistral), Round 2 vs. Round 4 (excerpted):**

Round 2: “The proposition assumes student organisers cannot evolve into ... Why should we fear their participation instead of trusting voters to decide?”
Round 4: “The proposition fears a revolutionary elite ... What right do we have to deny people the right to choose those who fought for their voices?”

Interpretation. This paired vignette illustrates one operational form of homogeneity in our setting: *argument lock-in*, where an agent settles early on a stable framing and then reuses it with minimal turn-by-turn adaptation. Because the two sessions share the same motion draw and the same model-to-role mapping, the contrast is difficult to explain by model identity or topic selection. Instead, it aligns with our quantitative finding that language conditioning can systematically affect the diversity of arguments produced in multi-agent debate.