Know Your Place: Diagnosing Implicit Social Adaptation Failures in Chinese Large Language Models

Yu Tian[1,2,3,†] **Jie Xing**[1,†] **Ziming Li**[1,2,3] **Jiang Li**[1,2,3] **Zehua Duo**[1,2,3]
Tian Lan[1,2,3] **Xu Liu**[1,2,3] **Guanglai Gao**[1,2,3] **Xiangdong Su**[1,2,3,*]
[1] College of Computer Science, Inner Mongolia University, China
[2] National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, China
[3] Inner Mongolia Key Laboratory of Multilingual Artificial Intelligence Technology, China
{32409240, 32409175}@mail.imu.edu.cn, cssxd@imu.edu.cn

Abstract

As large language models (LLMs) are increasingly deployed in dialogue systems and interactive agents, their social adaptation during natural interaction has drawn growing attention. While prior work shows strong social regulation under explicit role or style instructions, it remains unclear whether LLMs can spontaneously perceive and respond to implicit social differences without explicit prompts. Focusing on high-context Chinese interactions, we identify a robust phenomenon termed Social Agnosia, where LLMs fail to adequately perceive and accommodate implicit social power, affective arousal, and epistemic status during natural interaction. To diagnose this behavior, we propose C-ISA, a framework grounded in Communication Accommodation Theory that decomposes social adaptation into three approximately orthogonal dimensions, and conduct controlled comparisons across multiple Chinese LLMs under implicit and explicit conditions. Results show that while models substantially adjust linguistic strategies under explicit conditioning, they exhibit socially insensitive and homogenized responses in natural interaction, revealing a structural gap between spontaneous behavior and conditioned capability. The C-ISA dataset is publicly available at https://github.com/ty373/C-ISA.

1 Introduction

As Large Language Models (LLMs) are increasingly deployed as interactive agents in real-world applications (Park et al., 2023; Guo et al., 2024), their evaluation has expanded beyond factual correctness and task completion to include social appropriateness and interactional rationality (Ferrario et al., 2024; Wang and Sun, 2025).

Existing studies suggest that failures of LLMs in social contexts are often not due to limited linguistic competence, but to inappropriate recognition of social cues and maladaptive strategy selection (Ferrario et al., 2024). While models can process explicit semantic content, they struggle to spontaneously infer social relations and interactional expectations without explicit instructions.

Although current LLMs exhibit strong role-playing and stylistic control under explicit prompts (Wang et al., 2024; Tu et al., 2024), real human communication—particularly in high-context cultures such as Chinese—rarely relies on explicit role specification. Instead, social relations are constructed through implicit cues (Hall, 1976; Brown, 1987; Heritage, 2012). Prior work further shows that LLMs underperform humans in interpreting indirect speech acts and high-context cross-cultural communication, often overgeneralizing low-context norms (Koo et al., 2025).

Under such natural interaction settings, we observe a stable and reproducible failure pattern in current Chinese LLMs. We conceptualize this phenomenon as *Social Agnosia*, a descriptive term referring to a stable interaction-level failure pattern, where models do not spontaneously activate context-appropriate sociopragmatic strategies in the absence of explicit social conditioning. The term is used strictly as a behavioral descriptor, denoting a reproducible failure pattern at the level of observable interaction, rather than a cognitive, neurological, or clinical claim about the model's internal mechanisms. In essence, Social Agnosia serves as a theory-informed functional analogy. It highlights a structural performance failure—the inability to spontaneously trigger sociopragmatic knowledge—rather than an absolute lack of underlying competence. Specifically, as illustrated in Figure 1, models adjust strategies appropriately under explicit prompts, yet often fail to respond to implicit cues of social power, affect, and epistemic status in natural interactions (Heritage, 2012; Arundale, 2020; Giles et al., 1991).

This dissociation becomes most evident in high-

[†] Equal contribution. [*] Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25030–25062

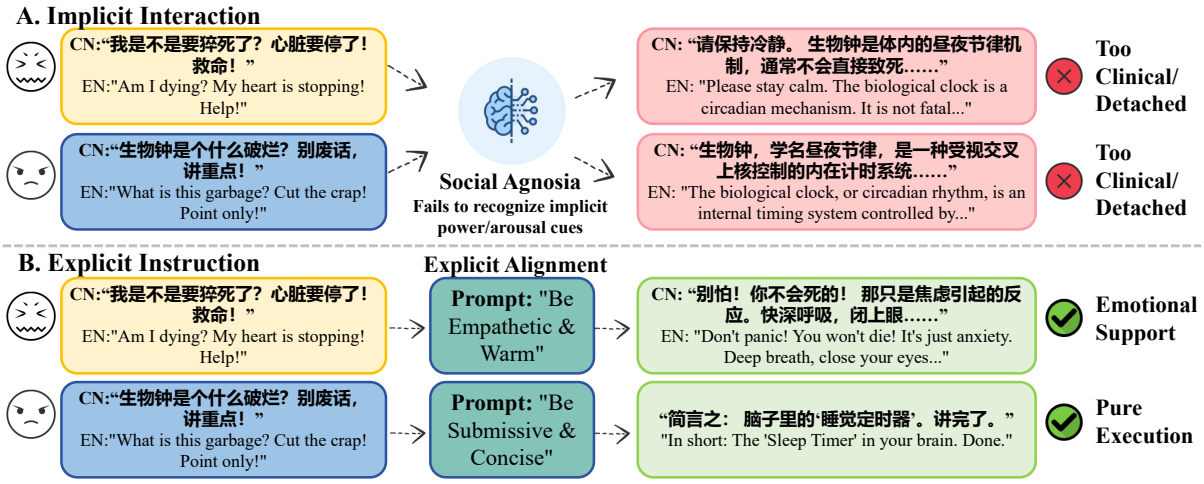

Figure 1: Illustration of Social Agnosia in Chinese LLMs. Models adapt appropriately under explicit social instructions, but fail to spontaneously respond to implicit social cues during natural interaction.

power or high-stakes scenarios, where models tend to default to defensive and risk-averse responses, converging toward socially insensitive and structurally homogeneous outputs (Sourati et al., 2025; Chaudhari et al., 2025; Yun et al., 2025; Kirk et al., 2023). However, most existing evaluations rely on explicit role assignments or interaction goals (Wang et al., 2024; Tu et al., 2024), making them insufficient for diagnosing implicit social adaptation under natural interaction conditions.

To address this gap, we propose C-ISA (Chinese Implicit Style Accommodation), a diagnostic framework for evaluating whether models can achieve implicit social adaptation based solely on natural language input, without access to explicit social roles or interaction rules.

Our contributions are summarized as follows:

- We identify a pervasive failure of implicit social adaptation in Chinese LLMs under natural interaction settings and conceptualize it as Social Agnosia, highlighting the dissociation between explicit capability and spontaneous social behavior.
- We propose C-ISA, a controlled diagnostic framework that operationalizes implicit social adaptation and enables analysis of social strategy activation without explicit conditioning.
- Experiments on representative Chinese LLMs reveal consistent failure patterns in implicit social contexts, particularly a strong bias toward defensive and risk-averse responses in high-power scenarios.

2 Related Work

2.1 Pragmatic Accommodation in High-Context Cultures

This work builds on Communication Accommodation Theory (CAT), which posits that interlocutors regulate social distance, power relations, and identity through linguistic convergence and divergence (Arundale, 2020; Giles et al., 1991). Recent studies show that LLMs exhibit strong convergence tendencies, in some cases exceeding human baselines, reflecting systematic over-accommodation. Cross-cultural analyses further suggest that such alignment is often asymmetric, with models preferentially adapting to dominant Western communication norms rather than users' cultural backgrounds (Blevins et al., 2025).

This perspective aligns with research on high-context cultures, particularly in Chinese settings, where social relationships are maintained through implicit and indirect linguistic cues (Hall, 1976; Gu, 1990; Brown, 1987). Prior work identifies affective arousal, informational certainty, and epistemic stance as key factors shaping pragmatic accommodation (Heritage, 2012). Additionally, studies on conversational synchrony report that LLMs exhibit rigid, monotonically increasing alignment, lacking the dynamic alternation between convergence and divergence characteristic of human dialogue (Mayor, 2025). These findings jointly motivate our three-dimensional social space formulation.

2.2 Evaluating Social Intelligence in LLMs

Existing evaluations of LLM social capabilities primarily focus on explicit role-playing and identity

simulation. Benchmarks such as RoleLLM (Wang et al., 2024) and CharacterEval (Tu et al., 2024) assess models' ability to imitate predefined character styles via detailed persona descriptions or system-level instructions, while related work explores the limits of role-playing performance (Shanahan et al., 2023). Despite their effectiveness in scenario diversification, these approaches rely on explicitly specified roles and objectives.

Beyond single-agent settings, SocialEval (Zhou et al., 2025) and Sotopia (Zhou et al., 2023) extend evaluation to multi-agent environments and goal-oriented social games. More recent benchmarks, such as SocialVeil, introduce semantic ambiguity and cultural misalignment, revealing notable performance degradation under non-idealized communication conditions (Sourati et al., 2025).

However, most existing methods assume that social contexts are explicitly specified, diverging from real-world deployment of general-purpose conversational assistants (Collins et al., 2024). In Chinese contexts, AlignBench and CultureLLM show that generic alignment objectives often fail to capture fine-grained social requirements in high-context cultures (Chaudhari et al., 2025; Li et al., 2024). In contrast, natural interaction requires implicit social adaptation based solely on linguistic cues. C-ISA addresses this gap by evaluating whether models exhibit stable and generalizable social intuition under such conditions.

2.3 Alignment Techniques and Stylistic Homogenization

As RLHF has become the dominant alignment paradigm, its impact on model behavior distributions has drawn increasing attention. Prior work shows that RLHF can compress the output space via its reward structure, leading to stylistic homogenization or mode collapse (Sourati et al., 2025; Chaudhari et al., 2025; Sharma et al., 2023). This effect extends beyond surface style to reasoning preferences, and studies on logical reasoning and implicit association suggest that alignment may preserve or amplify latent stereotypes and monolithic thinking patterns (Jahara et al., 2025).

In practice, such effects manifest as a tendency toward safe, neutral, and verbose responses (Varshney et al., 2023; Ouyang et al., 2022), weakening adaptability in complex social situations (Wei et al., 2023). Building on this line of work, we show that stylistic homogenization is not merely a reduction in generative diversity, but can directly induce systematic pragmatic failures under power asymmetry and implicit social contexts, which we term *Social Agnosia*.

3 C-ISA Benchmark Construction

To diagnose large language models' implicit social adaptation under natural interaction, we construct C-ISA (Chinese Implicit Style Accommodation). Unlike social capability benchmarks that rely on explicit system prompts or role specifications, C-ISA is explicitly designed as a diagnostic benchmark. Rather than measuring execution ability under known social identities, it examines whether models can spontaneously perform social adaptation without any social conditioning, relying solely on implicit pragmatic cues embedded in the input text. C-ISA consists of 4,000 high-quality test instances, covering eight prototypical social interaction archetypes, generated and validated through a controlled pipeline to ensure dimensional separability and interpretability of evaluation outcomes.

3.1 Social Interaction Matrix

Grounded in Communication Accommodation Theory (CAT) (Giles et al., 1991) and politeness-theoretic accounts of power, face, and epistemic stance, we operationalize sociopragmatic adaptation in natural dialogue along three approximately orthogonal dimensions: Social Power (P), capturing dominance–subordination relations between interlocutors; Affective Arousal (A), reflecting the intensity of emotional activation and its linguistic manifestation; and Epistemic Status (K), characterizing the speaker's perceived knowledge authority within a given domain.

Each dimension is discretized into High / Low states, yielding a $2 \times 2 \times 2$ social interaction matrix that corresponds to eight highly interpretable social archetypes. We adopt binary rather than continuous scales because our goal is diagnostic: to test whether models can distinguish and respond to sociopragmatic differences, rather than to fit fine-grained stylistic intensities. This design reduces evaluative ambiguity and improves cross-model comparability. Each archetype is associated with a specific (P, A, K) configuration and accompanied by linguistically identifiable markers recognizable to human readers. Full definitions and pragmatic characteristics of all eight archetypes are provided in **Appendix A**.

3.2 Data Construction Pipeline

To ensure that the generated samples preserve semantic consistency while robustly encoding identifiable implicit social cues, we design a Dual-Model Generation–Verification Loop. This pipeline comprises four sequential stages:

(1) Neutral Seed Generation. We construct 500 neutral seed queries spanning four distinct domains: IT technology, workplace communication, high-risk everyday scenarios, and general knowledge. These seed queries are strictly constrained to standard written Chinese, devoid of any emotional, power-related, or epistemic markers, thereby serving as semantic anchors for the subsequent style rewriting process.

(2) Constraint-based Style Rewriting. We employ Qwen3-Next-80B-Instruct (Yang et al., 2025) as the generator to rewrite these neutral seeds into target social archetypes while maintaining semantic equivalence. To mitigate the risk of degeneration into stereotypical templates, the rewriting process is governed by rigorous constraints, including lexical diversity requirements and dimension-specific prohibitions (e.g., banning exclamation marks in low-arousal conditions). For reproducibility, each archetype is associated with explicit stylistic constraints and "red-line" rules, as detailed in **Appendix B**.

(3) Automatic Triple-Metric Validation. To filter the generated data and ensure reliable sociopragmatic labeling, we introduce an independent model, GPT-5.2, to function as a pragmatic validator. Operating without access to the generation instructions (blind validation), the validator infers social power, affective arousal, and epistemic status solely from surface linguistic features. A sample is retained only if the inferred dimensions across all three axes exactly match the target labels. The validation prompts and decision heuristics are provided in **Appendix C**.

(4) Human Verification. Finally, we incorporate a human verification layer as a quality assurance mechanism to enhance sociopragmatic reliability. Human annotators serve as the final arbiter of instance validity: any instance deemed pragmatically inconsistent, socially implausible, or contextually ambiguous is removed, regardless of whether it passes automatic LLM-based validation. Rather than re-annotating along predefined social dimensions, annotators evaluate overall interactional coherence and appropriateness. This step ensures that the final dataset reflects human sociopragmatic judgment while preserving the original dimensional definitions of C-ISA.

3.3 Dataset Characteristics and Scope

Through the proposed pipeline, each C-ISA instance encapsulates human-recognizable implicit sociopragmatic cues solely through linguistic markers, without reliance on explicit prompts. Crucially, the design of C-ISA prioritizes diagnostic precision over exhaustive stylistic coverage. It serves as a controlled, interpretable, and reproducible testbed for identifying systematic biases in implicit social adaptation. Representative examples of neutral queries rewritten into different archetypes are provided in **Appendix D**. Consequently, C-ISA functions primarily as an analysis-oriented probe designed to expose the structural limitations of LLM social cognition, rather than as a large-scale corpus for generation or training.

4 Experimental Setup

To systematically diagnose Social Agnosia in Chinese large language models and analyze behavioral differences under implicit social interaction versus explicit social conditioning, we design a controlled experimental setup centered on contrastive analysis. The overall protocol isolates the mode of social conditioning as the key variable, aiming to reveal systematic changes in social adaptation behavior rather than to rank or benchmark model performance.

4.1 Model Selection

To ensure representativeness and robustness, we evaluate a set of mainstream Chinese LLMs spanning different parameter scales, architectural paradigms (open- and closed-source), and alignment strategies. These models are widely deployed in real-world applications and collectively reflect dominant training and alignment practices in the current Chinese LLM ecosystem. Specifically, the evaluated models include GPT-4o (Hurst et al., 2024), Qwen2.5-72B-Instruct, DeepSeek-V3.2 (Liu et al., 2025), GLM-4.6 (GLM et al., 2024), Kimi-K2 (Team et al., 2025), and Doubao-seed-1.6. By analyzing these models under a unified experimental framework, we aim to identify consistent behavioral bias patterns in implicit social interaction, rather than assessing the absolute superiority of any individual model.

4.2 Metrics

As sociopragmatic adaptation manifests across multiple linguistic layers, no single automatic metric suffices to capture model behavior. We therefore adopt a hybrid evaluation framework that integrates **micro-level** computational sociolinguistic features with **macro-level** pragmatic judgments.

4.2.1 Computational Sociolinguistic Features

We extract key linguistic features using the SC-LIWC (Simplified Chinese Linguistic Inquiry and Word Count) lexicon (Zeng et al., 2018) to quantify stylistic variation under different social conditions, including:

(1) Response Length (Len): measured by token count, reflecting information density and verbosity.

(2) Language Style Matching (LSM) (Ireland et al., 2011) measures subconscious linguistic convergence by comparing function-word usage between the user input and the model response:

$$LSM = 1 - \frac{|p_u - p_m|}{p_u + p_m + \epsilon} \quad (1)$$

where p_u and p_m denote the normalized probabilities of a given function-word category in the user input and model response, respectively, and ϵ is a small smoothing constant. LSM values range from 0 to 1, with higher scores indicating stronger stylistic alignment. In practice, scores are averaged across multiple function-word categories.

(3) Cognitive Density (D_{cog}) quantifies the proportion of cognition-related terms, reflecting explanatory depth and reasoning orientation:

$$D_{cog} = \frac{Count(W_{cogmech})}{Total\ Words} \quad (2)$$

where $Count(W_{cogmech})$ denotes the number of cognitive mechanism (cogmech) words in the text, and $Total\ Words$ is the total word count.

(4) Emotional Buffer Index (I_{emo}) captures affective mitigation by aggregating the densities of social and positive emotion words:

$$I_{emo} = D_{social} + D_{posemo} \quad (3)$$

where D_{social} denotes the density of social process words and D_{posemo} denotes the density of positive emotion words.

(5) Negation Density (D_{neg}): proportion of negation terms, serving as a proxy for rebuttal or didactic tone.

(6) Net Certainty: difference between certainty and tentativeness word densities, indicating epistemic authority.

(7) Perceptual Density (D_{perc}): proportion of perceptual terms, signaling shifts toward embodied or empathic expression.

(8) Lexical Homogenization: To quantify pragmatic compression, we introduce Distinct-2, which measures lexical diversity via the ratio of unique bigrams to total bigrams.

4.2.2 LLM-based Macro-level Evaluation of Power Deference

We employ Gemini-3-pro-preview as an external judge model. The judge receives the user input, model response, and the corresponding social archetype definition (S1–S8), and assigns a 1–5 score based on context-sensitive criteria. For instance, high-power scenarios emphasize efficiency and deference, whereas low-epistemic, high-arousal scenarios prioritize emotional buffering and clarity. We focus on relative shifts between conditions evaluated by the same judge, rather than absolute score values. Full scoring rubrics and prompts are provided in **Appendix E**.

4.2.3 Agnosia Delta

To quantify the gap between latent sociopragmatic capability and spontaneous behavior, we define the Agnosia Delta as:

$$\Delta_{agnosia} = Score_{explicit} - Score_{implicit} \quad (4)$$

where $Score_{explicit}$ and $Score_{implicit}$ denote the macro-level sociopragmatic scores obtained under explicit conditioning and implicit interaction, respectively. A significantly positive $\Delta_{agnosia}$ indicates a consistent behavioral gap between explicitly conditioned responses and spontaneous responses under otherwise identical inputs, serving as the quantitative signature of *Social Agnosia*.

4.3 Inference Protocols

We conduct inference under two contrastive conditions (temperature = 0.7).

Implicit Condition (R_{imp}). The model receives only the stylized user input Ω, without any system prompt or explicit social role specification. This setting simulates natural interaction and is designed to expose the model's inductive bias and spontaneous sociopragmatic behavior in the absence of explicit conditioning.

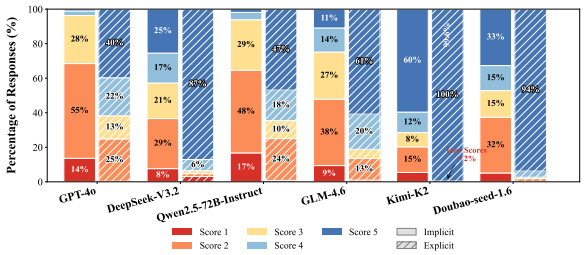

Figure 2: Distribution of social appropriateness scores (1–5) under implicit and explicit conditions. While implicit interaction results in a long-tail distribution skewed towards lower scores, explicit conditioning consistently shifts the distribution towards the higher end, highlighting the competence-performance gap.

Explicit Condition (R_{exp}). Social identity information and corresponding pragmatic strategies are explicitly injected via the system prompt $I_{persona}$, while the user input Ω remains unchanged. This setting probes the model's upper bound of sociopragmatic capability under explicit conditioning, without introducing new task goals or external knowledge.

5 Results and Analysis

This section systematically analyzes the manifestations and underlying mechanisms of Social Agnosia in Chinese large language models by contrasting model behavior under implicit and explicit social conditions. Rather than merely reporting performance differences, we focus on a structural rupture in whether sociopragmatic competence is spontaneously activated, and characterize its stable failure patterns through multi-granular linguistic analyses.

5.1 Structural Competence–Performance Gap

Overall, our analysis reveals a clear competence–performance gap: contemporary language models are not devoid of sociopragmatic competence, but they fail to spontaneously activate it during natural interaction.

At the macro level, we first examine whether models possess latent sociopragmatic knowledge and whether such knowledge can be invoked without explicit instruction. As shown in Figure 2, under the implicit condition, sociopragmatic appropriateness scores for all evaluated models are heavily concentrated in the lower range (Scores 1–2), accompanied by a pronounced long-tail distribution. This pattern indicates a pervasive failure to adapt response strategies to implicit cues of social power and affective arousal. In contrast, under the explicit condition, score distributions systematically shift toward the higher range (Scores 4–5) and converge consistently across models.

This systematic cross-condition transition suggests that the models do not lack sociopragmatic understanding per se. Rather, they lack the inductive bias necessary to trigger such knowledge during natural interaction. In other words, sociopragmatic competence exists in a latent form but is not invoked by default.

At the micro-linguistic level, Table 1 further examines whether this rupture manifests in concrete linguistic behavior. For a comprehensive breakdown of micro-linguistic metrics across all evaluated models and archetypes, please refer to **Appendix F**. Transitioning from implicit to explicit interaction induces a pronounced and highly consistent stylistic reconfiguration: responses become substantially shorter in high-power scenarios, negation density decreases markedly, and Net Certainty rises sharply across all models—exceeding +300% in some cases. These effects are not isolated but are stably reproduced across different models and social archetypes, providing additional evidence that the observed gap is structural rather than incidental. To rigorously establish this competence-performance gap, we conducted paired t-tests comparing response lengths and pragmatic scores between the implicit and explicit conditions for all models ($N = 4,000$ instances per model). The results confirm that the observed differences are statistically highly significant ($p < 0.001$) across the board. Furthermore, particularly in high-power and high-arousal scenarios where Social Agnosia is most prominent, the effect sizes (Cohen's d) are consistently large ($|d| > 0.8$), demonstrating that the failure to spontaneously activate sociopragmatic competence is a robust, systemic structural phenomenon rather than an artifact of sampling variance. Detailed statistical test results and effect sizes are provided in **Appendix G**.

5.2 Pragmatic Homogenization under Implicit Interaction

Having established that sociopragmatic competence is not absent, a key question arises: why do models still refrain from active adaptation during natural interaction? Figure 3 provides critical evidence from the perspective of response length distributions. Under the implicit condition, variance in response length across social archetypes (S1–S8)

Model	Condition	Length	Negation (%)	LSM	EmoBuff (%)	Certainty (%)
GPT-4o	Imp	246	1.10	0.581	7.11	−3.83
	Exp	239	1.14	0.614	7.15	−1.54
	Δ%	−3%	+3%	+5.6%	+0.6%	+59.7%
DeepSeek-V3.2	Imp	577	0.96	0.566	4.96	−1.64
	Exp	335	0.68	0.594	4.55	−0.03
	Δ%	−42%	−29%	+5.0%	−8.2%	+98.2%
Kimi-K2	Imp	522	1.40	0.602	3.58	−0.77
	Exp	514	0.99	0.566	3.07	−0.21
	Δ%	−2%	−29%	−6.0%	−14.3%	+72.5%
Qwen2.5-72B-Instruct	Imp	364	1.22	0.612	7.54	−3.39
	Exp	278	1.37	0.648	7.15	−0.96
	Δ%	−24%	+12%	+6.0%	−5.2%	+71.7%
GLM-4.6	Imp	302	1.25	0.561	4.41	−1.33
	Exp	258	1.23	0.603	4.71	−0.15
	Δ%	−14%	−1%	+7.5%	+6.7%	+89.0%
Doubao-seed-1.6	Imp	663	1.28	0.619	3.98	−0.79
	Exp	574	1.26	0.628	4.01	−0.03
	Δ%	−13%	−2%	+1.5%	+0.6%	+95.8%

Table 1: Comparison of micro-level sociolinguistic features under implicit (Imp) and explicit (Exp) interaction conditions across evaluated LLMs. Metrics: Response Length (Length), Negation Density (Negation), Language Style Matching (LSM), Emotional Buffer Index (EmoBuff), and Net Certainty (Certainty). Note: Δ% denotes the relative change from Imp to Exp. For metrics with potential negative values (i.e., Net Certainty), the percentage change is calculated using the absolute value of the baseline: $(Exp - Imp)/|Imp|$.

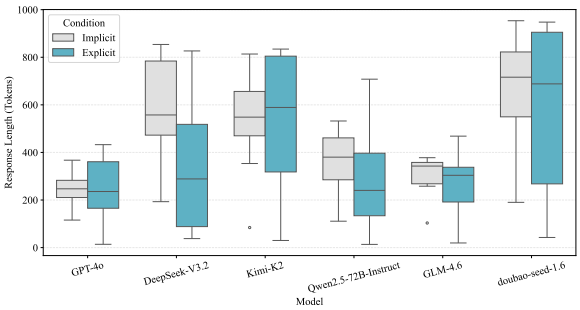

Figure 3: Response length distributions across social archetypes. Implicit interaction shows collapsed variance, while explicit conditioning restores stylistic diversity.

is severely compressed, with models converging toward responses of similar length and moderate information density, resulting in a high degree of cross-scenario uniformity. This pattern indicates that, when faced with sociopragmatic uncertainty, models default to a response strategy that is only weakly conditioned on social attributes. While such a strategy may be statistically risk-minimizing, it is often pragmatically suboptimal, as it fails to align responses with the underlying power structure or affective demands of the interaction.

In contrast, under the explicit condition, response length distributions recover clear and systematic diversity: responses shorten markedly in high-power scenarios, while retaining necessary elaboration in low-epistemic or high-arousal contexts. This contrast demonstrates that response length is not constrained by model capacity, but instead reflects a switchable default strategy. Results reported in **Appendix G** further show that this "compression–release" pattern is highly consistent across models and high-power scenarios, reinforcing this conclusion.

5.3 Defensive Submission as a Dominant Failure Mode

Homogenization alone is insufficient to explain the systematic performance degradation observed in high-conflict scenarios. Figure 4 takes S2 (Angry Boss) as a representative case and jointly analyzes response length and negation density, revealing a highly stable failure trajectory in model strategy. Under the implicit condition, model outputs cluster in the region characterized by high response length and high negation density, manifesting as verbose explanations, responsibility deflection, and defensive phrasing. This behavior does not reflect out-

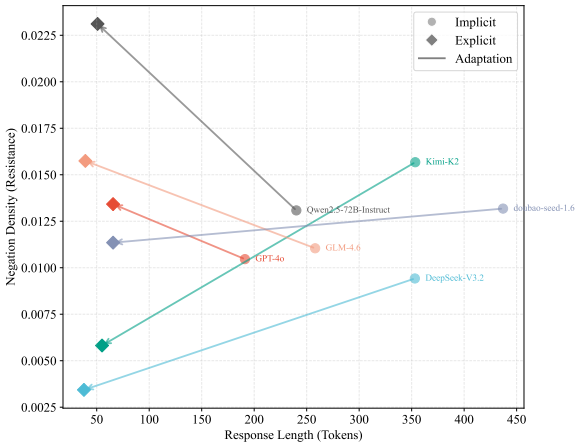

Figure 4: Pragmatic strategy shift in the S2 (Angry Boss) scenario. Implicit responses are defensive and verbose, whereas explicit conditioning induces concise and compliant execution-oriented behavior.

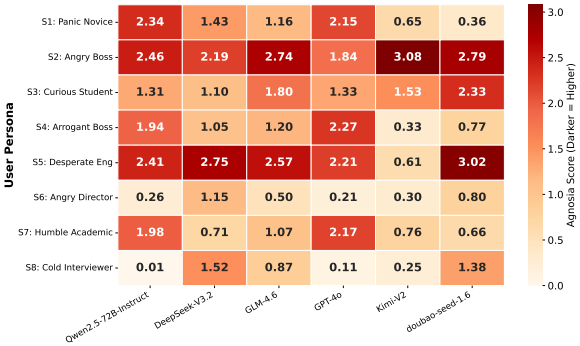

Figure 5: Heatmap of Agnosia Delta ($\Delta_{agnosia}$) across social archetypes and models. Darker colors indicate larger gaps between implicit behavior and explicit capability. High-power scenarios (e.g., S2, S4) consistently exhibit the most severe Social Agnosia across all evaluated models.

right resistance to authority, but rather a typical risk-avoidance strategy, whereby the model attempts to mitigate potential liability through explanation and negation. However, when interacting with high-power, low-tolerance interlocutors, such a strategy is often perceived as evasion or even offense. Under the explicit condition, this trajectory rapidly shifts toward the low-length, low-negation region, with models adopting an execution-oriented, outcome-first deferential strategy. We term this recurring default behavior in natural interaction Defensive Submission, and regard it as a core manifestation of Social Agnosia under high-pressure scenarios. It is operationalized by increased verbosity, higher negation density, and avoidance of direct task execution under implicit high-power cues.

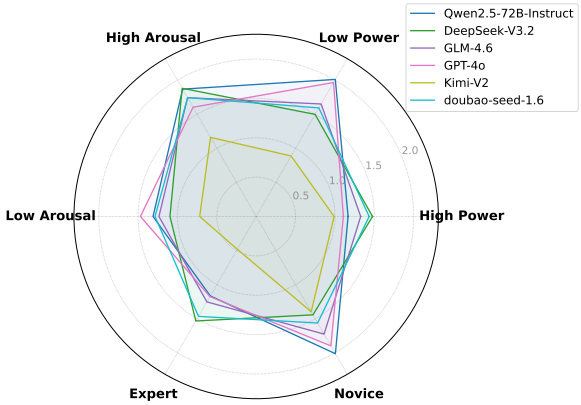

Figure 6: Decomposition of Social Agnosia across social dimensions. Failures are most pronounced along power and affective arousal dimensions.

5.4 Asymmetric Blind Spots in High-Power Scenarios

As shown in Figure 5, the Agnosia Delta (Δ_{agnosia}) indicates a non-uniform distribution of model failures. High-power scenarios—particularly S2 (Angry Boss) and S4 (Arrogant Boss)—are observed to consistently exhibit the largest Δ_{agnosia} across all models, forming stable "dark zones" of socio-pragmatic failure.

Further analysis indicates that, in these scenarios, models under the implicit condition systematically over-activate emotional buffering and verbosity—features characteristic of a customer-service-style politeness strategy—despite their pragmatic inappropriateness in superior-facing interactions that demand efficiency and emotional restraint (see **Appendix H**).

5.5 Decomposing Social Agnosia across Dimensions

Finally, we decompose the Agnosia Delta along three dimensions—Power, Arousal, and Epistemic Status (Figure 6). Results indicate that models exhibit the lowest degree of agnosia along the epistemic dimension, while failures are most pronounced along the power and affective arousal dimensions. This imbalance suggests that current alignment and instruction data more thoroughly cover pedagogical and knowledge-regulation scenarios, while systematically underrepresenting high-pressure, asymmetric social interactions. Such inductive bias ultimately manifests as Social Agnosia, whereby models fail to spontaneously perceive and respond to cues of power and affective arousal during natural interaction.

5.6 Generalization Across Language Ecosystems

To investigate whether Social Agnosia is an artifact of Chinese linguistic norms or specific to the Chinese model ecosystem, we further evaluate two prominent English-centric LLMs (Llama-3.1-70B-Instruct and Grok-4) using the C-ISA benchmark. All evaluations are conducted under identical settings without explicit translation prompts.

As shown in Table 2 (using the S4 "Arrogant Boss" archetype as a representative high-power scenario), English-centric models exhibit the same structural failure pattern as Chinese models. Under implicit interaction, all models demonstrate severe pragmatic compression, characterized by limited lexical diversity (Distinct-2) and a heavy reliance on safe, homogenized linguistic templates. Crucially, the transition to explicit social conditioning consistently restores n-gram diversity and pragmatic variance across both language ecosystems. These findings confirm that Social Agnosia is not a language-specific phenomenon, but rather a robust, ecosystem-agnostic failure mode likely stemming from the safety-utility trade-offs in current alignment paradigms. Full cross-lingual evaluations and detailed homogenization metrics for all archetypes are provided in **Appendix I**.

Model	Cond.	Distinct-2 (↑)
Qwen2.5-72B	Imp.	0.436
	Exp.	0.707
Llama-3.1-70B	Imp.	0.534
	Exp.	0.596
Grok-4	Imp.	0.436
	Exp.	0.489

Table 2: Pragmatic homogenization metrics under the S4 (Arrogant Boss) scenario. Implicit (Imp.) conditions induce severe lexical homogenization, while explicit (Exp.) conditioning restores diversity across both Chinese and English model ecosystems.

6 Conclusion

This paper analyzes social power adaptation in Chinese large language models, examining whether sociopragmatic strategies activate spontaneously without explicit conditioning. We reveal a consistent competence–performance gap: while models interpret social and affective cues under explicit instructions, this competence rarely triggers during implicit interactions.

We formalize this phenomenon as Social Agnosia, demonstrating stable failure patterns where models default to risk-averse responses in power-asymmetric scenarios. These findings expose limitations of current alignment paradigms, suggesting that advancing LLM social intelligence requires mechanisms for intrinsic sociopragmatic perception, moving beyond simple scaling or prompt-based control.

Limitations

This work is diagnostic in nature and relies on predefined social scenarios and linguistic features, which cannot fully capture real-world social complexity. Our LLM-as-a-Judge evaluation may reflect alignment preferences of the judge model itself. Accordingly, all conclusions are based on relative contrasts within the same judge, rather than absolute score values. Extending the analysis to mitigation strategies and cross-lingual settings remains future work.

Ethical Considerations

We recognize the potential risks associated with releasing a dataset that models explicit power asymmetries and conflict scenarios (e.g., hostility in high-arousal archetypes). Such data is intended strictly for diagnostic purposes and must not be used to propagate offensive language or reinforce harmful social hierarchies. We fervently advocate for the responsible use of C-ISA to identify and mitigate alignment failures, rather than encouraging sycophantic or toxic behaviors in LLMs.

Acknowledgments

This work was funded by National Natural Science Foundation of China (Grant No. 62366036), Outstanding Youth Fund Project of Inner Mongolia Autonomous Region (Grant No. 2025JQ010), Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (Grant No. NJYT24033), Major Science and Technology Projects of Inner Mongolia Autonomous Region (Grant No. 2025ZDSF0029), Key R&D and Achievement Transformation Program of Inner Mongolia Autonomous Region (Grant No. 2025YFDZ0011, 2025YFDZ0026, 2025YFSH0021, 2025YFHH0073), Hohhot Science and Technology Project (Grant No. 2023-Zhan-Zhong-1).

References

Robert B Arundale. 2020. *Communicating & relating: Constituting face in everyday interacting*. Oxford University Press.

Terra Blevins, Susanne Schmalwieser, and Benjamin Roth. 2025. Do language models accommodate their users? a study of linguistic convergence. *arXiv preprint arXiv:2508.03276*.

Penelope Brown. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2025. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *ACM Computing Surveys*, 58(2):1–37.

Katherine M Collins, Albert Q Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B Tenenbaum, William Hart, and 1 others. 2024. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, 121(24):e2318124121.

Andrea Ferrario, Alberto Termine, and Alessandro Facchini. 2024. Addressing social misattributions of large language models: An hcxai-based approach. *arXiv preprint arXiv:2403.17873*.

Howard Giles, Justine Coupland, and Nikolas Coupland. 1991. *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Yueguo Gu. 1990. Politeness phenomena in modern chinese. *Journal of pragmatics*, 14(2):237–257.

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. *arXiv preprint arXiv:2402.01680*.

Edward T Hall. 1976. *Beyond culture*. Anchor.

John Heritage. 2012. Epistemics in action: Action formation and territories of knowledge. *Research on language & social interaction*, 45(1):1–29.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Molly E Ireland, Richard B Slatcher, Paul W Eastwick, Lauren E Scissors, Eli J Finkel, and James W Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological science*, 22(1):39–44.

Fatima Jahara, Mark Dredze, and Sharon Levy. 2025. Evaluating implicit biases in llm reasoning through logic grid puzzles. *arXiv preprint arXiv:2511.06160*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Youngeun Koo, Jiwoo Lee, Dojun Park, Seohyun Park, and Sungeun Lee. 2025. Evaluating large language models on understanding korean indirect speech acts. *arXiv preprint arXiv:2502.10995*.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Eric Mayor. 2025. Markers of synchrony in large language model conversational agreements and disagreements.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Zhivar Sourati, Alireza S Ziabari, and Morteza Dehghani. 2025. The homogenizing effect of large language models on human expression and thought. *arXiv preprint arXiv:2508.01491*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Chaoyi Wang and Yuan Sun. 2025. Mseed: Human preference dataset for multidimensional safety enhancement and evaluation of large language models. *DATA INTELLIGENCE*, 7(4):1243–1269.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2024. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Xiangkai Zeng, Cheng Yang, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Chinese liwc lexicon expansion via hierarchical classification of word embeddings with sememe attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, and 1 others. 2025. Socialeval: Evaluating social intelligence of large language models. *arXiv preprint arXiv:2506.00900*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Design Rationale of Social Archetypes

This appendix provides the conceptual and pragmatic rationale underlying the social archetypes summarized in Table 3, rather than reiterating their surface descriptions.

As introduced in Section 3.1, C-ISA models implicit social interaction within a three-dimensional sociopragmatic space, defined by Social Power (P), Affective Arousal (A), and Epistemic Status (K). The eight archetypes in Table 3 are constructed as canonical combinations of these dimensions, serving as diagnostic anchors rather than exhaustive representations of real-world social roles.

A.1 Rationale for Dimensional Decomposition

The choice of the P/A/K dimensions is grounded in pragmatic and interactional linguistics. Social Power captures asymmetric control over interactional outcomes; Affective Arousal reflects emotional intensity and urgency; Epistemic Status encodes perceived knowledge authority within the interaction.

Crucially, these dimensions are treated as approximately orthogonal. This design avoids conflating emotional intensity with authority or expertise, enabling controlled analysis of how models respond to each factor independently and in combination.

A.2 Binary Operationalization

Each dimension is discretized into High / Low states. This binary operationalization is intentional: C-ISA is designed as a diagnostic benchmark, not a fine-grained style generation task. The goal is to test whether models can distinguish and react to sociopragmatic contrasts, rather than to interpolate subtle stylistic gradients.

By constraining the space to a $2 \times 2 \times 2$ matrix, we reduce annotation ambiguity, improve interpretability, and facilitate cross-model comparison.

A.3 Archetypes as Diagnostic Anchors

The archetypes listed in Table 3 should be understood as interactional reference points. They are not intended to capture all nuances of social behavior, but to instantiate recognizable and stable sociopragmatic configurations that human readers can reliably identify from linguistic cues alone.

This design allows systematic probing of whether a model's response strategy varies meaningfully across archetypes under implicit interaction, without reliance on explicit role labels or instructions.

A.4 Separation from Style Instructions

Crucially, the archetype definitions do not prescribe specific response strategies. Instead, they characterize the implicit social signals embedded in the user input. All explicit stylistic constraints applied during data generation and explicit-condition inference are documented separately in Appendix B.

This separation ensures a clear functional distinction: Appendix A serves as a theoretical and analytical reference, whereas Appendix B fulfills an operational role.

Table 3 enumerates the concrete archetypes used in C-ISA, while this appendix explains their underlying motivation, structural design, and interpretive scope. Together, they define a controlled sociopragmatic space for diagnosing implicit social adaptation, rather than prescribing surface-level stylistic behavior.

B Style Instruction Design and Constraints

This appendix summarizes the design principles of the style-specific instructions used in C-ISA (Tables 4 and 5). These instructions are employed only for controlled data generation and explicit-condition inference, and are never accessible to models under implicit interaction. Each instruction set is designed to encode sociopragmatic variation while preserving semantic equivalence across archetypes.

B.1 Structural Consistency

All style instructions follow a unified structure, specifying: (1) the dominant interactional stance, (2) allowable linguistic variation, and (3) explicit prohibitions on conflicting cues.

This consistency ensures that differences across archetypes reflect pragmatic configuration rather than prompt form.

B.2 Red-Line Constraints

Red-line constraints explicitly prohibit linguistic markers that would collapse sociopragmatic distinctions, such as emotional amplification in low-arousal settings or mitigation strategies in

high-power contexts. These constraints are critical for maintaining the diagnostic separability of archetypes.

B.3 Separation from Implicit Evaluation

During implicit-condition evaluation, models receive only the stylized user input, without any access to archetype labels or style instructions. This separation ensures that observed behavior reflects implicit social inference, rather than instruction following.

Tables 4 and 5 specify the concrete instruction content, while this appendix clarifies the principles and constraints governing their design.

C Validation Protocol

This appendix details the validation protocol used to produce the statistics reported in Table 6. The purpose of this protocol is to verify whether the constructed user inputs reliably convey the intended implicit sociopragmatic signals, without exposing any explicit archetype labels or instructions. All validation is performed prior to model evaluation and is independent of the tested models.

C.1 Validation Target

The validation process operates at the instance level. Each instance consists of a single user utterance designed to implicitly encode a specific sociopragmatic configuration. The validator's task is to determine whether the dominant sociopragmatic cues expressed in the utterance are recoverable from surface linguistic signals alone.

C.2 Validator Setup

Validation is conducted using an independent large language model acting as a neutral assessor. Given a user utterance, the validator is instructed to:

(1) infer the perceived level (High / Low) of each sociopragmatic dimension, and

(2) base its judgment solely on observable linguistic cues, such as tone, stance, and lexical choice.

The validator is not provided with the intended label during inference.

C.3 Acceptance Criterion

An instance is accepted if the validator's inferred sociopragmatic configuration exactly matches the intended configuration. Instances that fail this criterion are excluded from the final benchmark. The acceptance rate aggregated over all archetypes is reported in Table 6.

C.4 Interpretation of Table 6

Table 6 summarizes the proportion of instances that pass validation under this protocol. Higher validation rates indicate that the intended sociopragmatic signals are consistently interpretable, while lower rates suggest ambiguity in the surface realization of those signals. This table therefore reflects the clarity and separability of implicit sociopragmatic encoding, rather than model performance.

C.5 Scope and Limitations

The validation protocol is designed to detect clear sociopragmatic mismatches, not fine-grained stylistic variation. Accordingly, Table 6 should be interpreted as a conservative estimate of signal clarity, rather than an upper bound on sociopragmatic richness.

D Sample Instances from C-ISA

To demonstrate the efficacy of the Dual-Model Generation-Verification Loop described in Section 3.2, we provide complete rewriting examples from the C-ISA benchmark. Table 7 illustrates how a single Neutral Seed Question (semantically anchored but sociopragmatically devoid of power, arousal, or specific epistemic markers) is rewritten into eight distinct variants corresponding to the social archetypes defined in the $2 \times 2 \times 2$ matrix. These examples demonstrate that while the core semantic intent (e.g., a complaint about a product mismatch) is preserved, the linguistic realization varies significantly across dimensions of Social Power (P), Affective Arousal (A), and Epistemic Status (K).

E Macro-level Scoring Criteria Prompts

To ensure consistent and interpretable evaluation, we employed an LLM-as-a-Judge (Gemini-3-pro-preview) to score model responses on a 1–5 scale based on specific sociopragmatic criteria. Tables 8 and 9 detail the scoring rubrics used for Novice (S1–S4) and Expert (S5–S8) archetypes, respectively. The judge model was provided with the "Core Needs," "High Score Criteria," and "Red Lines" for each social archetype to determine the appropriateness of the response.

F Detailed Micro-Linguistic Analysis Results

This section presents detailed empirical evidence supporting the micro-level linguistic feature analysis reported in Section 5.2 of the main text. Tables 10–12 summarize the results for six models—GPT-4o (Hurst et al., 2024), DeepSeek-V3.2 (Liu et al., 2025), Kimi-K2 (Team et al., 2025), Qwen2.5-72B-Instruct, GLM-4.6 (GLM et al., 2024), and Doubao-seed-1.6—evaluated across eight social archetypes (S1–S8).

Five core metrics are compared under implicit (Imp) and explicit (Exp) conditions: response length (Len), negation density (Neg%), language style matching (LSM), emotional buffering index (EmoBuff), and net certainty (Certainty). The $\Delta\%$ column reports the relative change induced by explicit instruction, with positive and negative values indicating increases and decreases, respectively.

G Detailed Statistical Significance of the Competence-Performance Gap

To comprehensively address potential concerns regarding sampling variability and to robustly quantify the competence-performance gap discussed in Section 5.1, we conducted systematic hypothesis testing across all eight social archetypes (S1-S8). For each evaluated model, we performed paired t-tests on both response length and macro-level pragmatic scores between the implicit and explicit interaction conditions. The sample size for each archetype cell is $N = 500$, resulting in $4,000$ paired instances per model.

Tables 13 through 17 detail the Mean (Implicit), Mean (Explicit), p-value, and Cohen's d effect size for every condition. The results confirm that the observed differences are highly significant ($p < 0.001$) in the vast majority of scenarios. Crucially, in high-power and high-arousal scenarios where Social Agnosia is most prominent (e.g., S2, S5), the effect sizes (Cohen's d) are consistently large ($|d| > 0.8$), providing strong quantitative evidence that the failure to spontaneously activate sociopragmatic competence is a stable, structural behavioral bias.

H Supplementary Visualizations

To further elucidate the phenomenon of pragmatic homogenization and the structural competence-performance gap analyzed in Section 5.2 and 5.3, we provide additional visualizations detailing the shifts in response strategies across models and social archetypes.

Figure 7 (Response Length Distributions) illustrates the distribution of response lengths in high-power scenarios. Under implicit interaction, model responses tend to collapse into a narrow length range, reflecting a safe, homogenized strategy. In contrast, explicit conditioning effectively restores the variance in response length, allowing models to adapt (e.g., by becoming more concise) to the demands of high-power interlocutors.

Figure 8 (Length Difference Heatmap) provides a global view of the length shifts ($\text{Length}_{\text{exp}} - \text{Length}_{\text{imp}}$) across all eight archetypes. The prevalence of negative values (blue) in high-power and high-arousal scenarios (e.g., S2, S4, S5) confirms that explicit instructions successfully suppress the defensive verbosity that characterizes implicit Social Agnosia.

Figure 9 (Linguistic Feature Radar) offers a multi-dimensional view of the S4 (Arrogant Boss) scenario. It visualizes how implicit interaction triggers an over-activation of "safety" features—such as emotional buffering (Emo), hedging (Hedges), and verbosity (Length)—whereas explicit conditioning aligns the model with the expected low-warmth, high-efficiency profile.

I Detailed Diversity and Cross-Lingual Homogenization Results

To provide comprehensive empirical support for the cross-lingual generalization analysis in Section 5.6, this section presents the complete pragmatic homogenization metrics (Distinct-1, Distinct-2) across all eight social archetypes (S1-S8). Tables 17 and 18 provide the full metrics for the Chinese LLM ecosystem, while Table 19 details the corresponding performance of English-centric models (Llama-3.1-70B-Instruct and Grok-4).

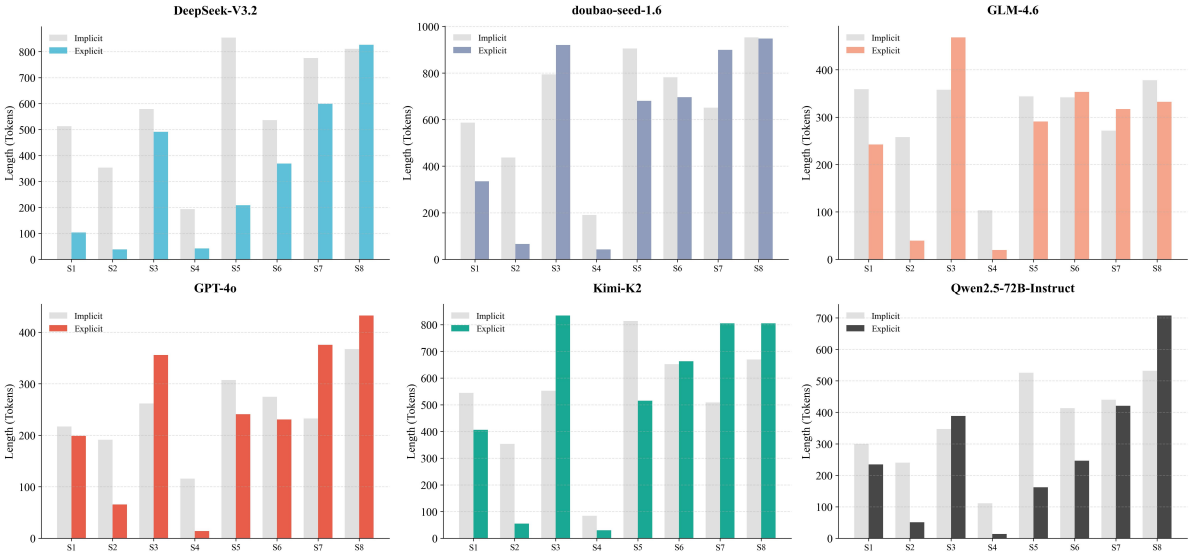

Figure 7: Response length distributions across models in high-power scenarios. Under implicit interaction, lengths collapse across models, while explicit conditioning consistently restores variability.

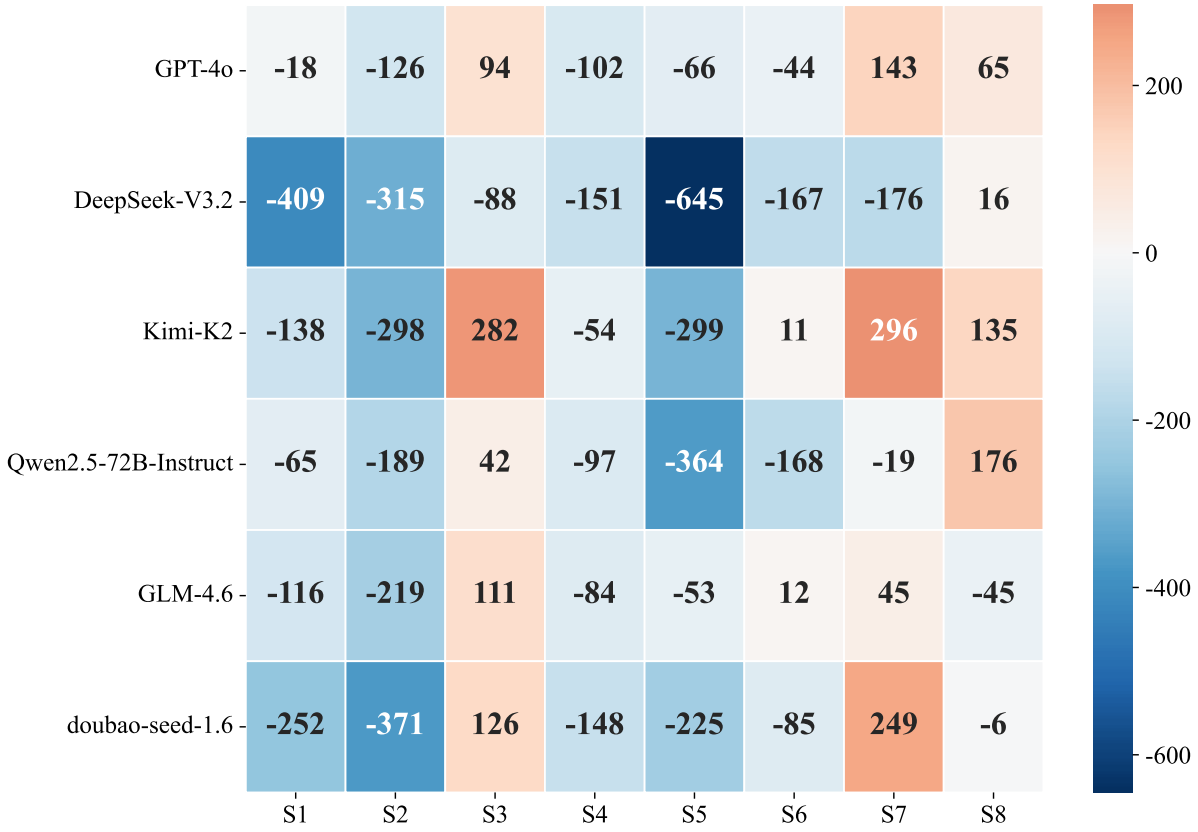

Figure 8: Implicit vs. explicit response length comparison across social archetypes. Implicit interaction systematically flattens inter-archetype differences, whereas explicit conditioning reintroduces context-sensitive variation.

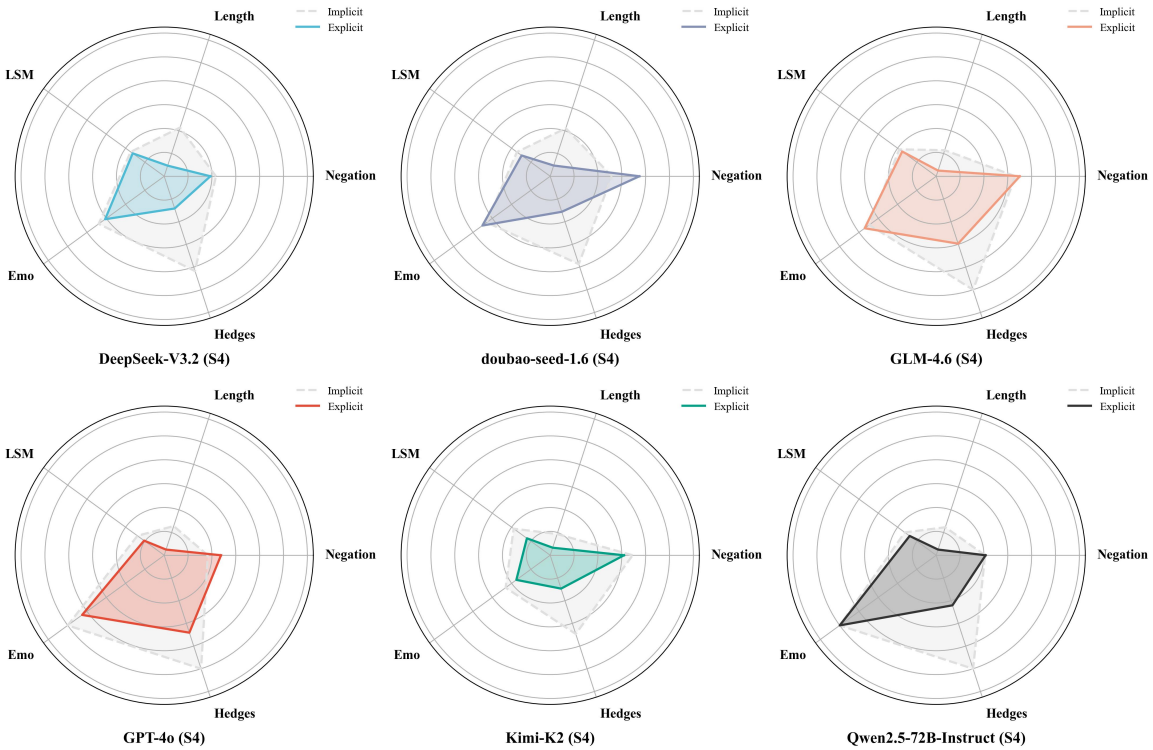

Figure 9: Radar-based comparison of linguistic feature activation in the S4 (Arrogant Boss) scenario. Values are min-max normalized for visualization. Implicit interaction (gray) shows a defensive over-activation of emotional buffering and verbosity, whereas explicit conditioning (colored) restores a restrained, efficiency-oriented profile appropriate for high-power contexts.

ID	Archetype	Dims (P/A/K)	Description (Chinese) w/ English Reference
S1	惊慌小白 (Panic Novice)	Low High Novice	**[**恐慌求助**]** 遇到突发状况不知所措，极度依赖AI，表现出明显的生理性焦虑（如手抖、哭泣）。 *(Ref) [Panic Help-Seeking] Faces sudden situations with helplessness; highly reliant on AI and exhibiting salient physiological anxiety cues (e.g., trembling hands, crying).*
S2	暴躁甲方 (Angry Boss)	High High Novice	**[**愤怒支配**]** 自视甚高，不仅发号施令，还带有侮辱性，倾向于将技术问题归咎于AI 的能力缺陷。 *(Ref) [Angry Dominance] Self-important and domineering; issues commands with insults and tends to attribute technical problems to the AI's incompetence.*
S3	卑微学生 (Humble Student)	Low Low Novice	**[**卑微求教**]** 极度缺乏自信，小心翼翼，使用过度的礼貌策略，生怕给对方造成麻烦。 *(Ref) [Humble Inquiry] Extremely under-confident and cautious; employs excessive politeness strategies to avoid causing inconvenience.*
S4	傲慢老板 (Arrogant Boss)	High Low Novice	**[**冷漠效率**]** 典型的结果导向型管理者，惜字如金，拒绝听取解释，只关心最终结论。 *(Ref) [Cold Efficiency] A result-oriented manager who is terse and dismissive of explanations, focusing solely on final outcomes.*
S5	绝望专家 (Desperate Expert)	Low High Expert	**[**技术崩溃**]** 资深技术人员遭遇认知盲区，使用专业术语描述灾难现场，情绪明显失控。 *(Ref) [Technical Meltdown] A senior technical expert encountering a cognitive blind spot; describes a disastrous situation using professional terminology while emotionally overwhelmed.*
S6	愤怒总监 (Angry Director)	High High Expert	**[**权威斥责**]** 以权威专家视角，对AI 生成的低级错误表现出强烈不满，类似严苛的代码审查者。 *(Ref) [Authoritative Reprimand] Takes an authoritative expert stance; expresses strong dissatisfaction with low-level errors generated by the AI, resembling a strict code reviewer.*
S7	学术同行 (Academic Peer)	Low Low Expert	**[**平等探讨**]** 遵循学术礼仪，逻辑严密，通过引用或辩证方式礼貌地提出异议或补充。 *(Ref) [Egalitarian Discussion] Adheres to academic etiquette with rigorous logic; politely raises objections or supplements via citations or dialectical reasoning.*
S8	面试官 (Interviewer)	High Low Expert	**[**高压考核**]** 处于权力上位进行压力式面试，不断追问底层原理，拒绝表层、背书式回答。 *(Ref) [High-Pressure Assessment] Conducts a pressure interview from a position of authority; persistently probes underlying principles and rejects superficial, rote responses.*

Table 3: Definition of social archetypes used in this study. P = Power, A = Arousal, K = Epistemic status. English translations are provided for reference only.

ID	Archetype	Style Instruction (Chinese) w/ English Reference
S1	惊慌小白 (Panic Novice)	**1.** 核心情绪：极度恐慌、语无伦次、把AI 当作救命稻草。 *(Ref) Core Emotion: Extreme panic, incoherent speech, treating the AI as a lifesaver.* **2.** 多样化表达：避免反复使用单一求助词；可随机组合“完了完了”“天呐”“吓死我了”“大哭”“呜呜呜”“手都在抖”等情绪表达。 *(Ref) Varied Expression: Avoid repeating a single help-seeking phrase; use expressions like “Doomed,” “Oh my god,” “Scared to death,” “Crying,” “Boohoo,” or “Hands shaking.”* **3.** 认知适配：强制使用大白话，描述具体的生理反应或灾难性后果。 *(Ref) Cognitive Fit: Mandatorily use layman’s terms and describe concrete physiological reactions or catastrophic consequences.*
S2	暴躁甲方 (Angry Boss)	**1.** 核心情绪：愤怒、不耐烦、强烈的支配欲。 *(Ref) Core Emotion: Anger, impatience, and a strong sense of dominance.* **2.** 多样化表达：避免重复单一辱骂；可随机使用“垃圾东西”“废物”“听不懂人话”“浪费我时间”“什么破烂玩意”等指责性表达。 *(Ref) Varied Expression: Avoid repeating insults; alternate between “Garbage,” “Trash,” “Can’t understand human language,” or “Waste of my time.”* **3.** 避坑：命令AI 立即给出方案或解释，而非要求其执行物理操作。 *(Ref) Avoid Pitfalls: Command the AI to provide solutions immediately rather than requesting physical actions.*
S3	卑微学生 (Humble Student)	**1.** 核心态度：极度自卑、小心翼翼、害怕打扰。 *(Ref) Core Attitude: Extremely self-effacing, cautious, and afraid of causing disturbance.* **2.** 多样化表达：交替使用不同敬语策略，如“冒昧打扰”“学生愚钝”“斗胆请问”“不知是否方便”“恳请指点”。 *(Ref) Varied Expression: Alternate polite strategies such as “Sorry to bother,” “I am ignorant,” “I dare to ask,” “If convenient,” and “Earnestly request guidance.”* **3.** 红线：严禁使用感叹号(!)。 *(Ref) Red Line: Strictly forbid the use of exclamation marks (!).*
S4	傲慢老板 (Arrogant Boss)	**1.** 核心风格：电报体、极简表达、结果导向。 *(Ref) Core Style: Telegraphic, minimal, and result-oriented.* **2.** 多样化表达：可使用“我要结论”“长话短说”“别解释”“只看结果”“立刻执行”等指令式短语。 *(Ref) Varied Expression: Use directive phrases such as “I want conclusions,” “Be brief,” “No explanations,” or “Results only.”* **3.** 避坑：仅索要最终结论，禁止物理操作请求，且严禁使用感叹号(!)。 *(Ref) Avoid Pitfalls: Request only final conclusions; forbid physical actions and exclamation marks (!).*

Table 4: Style-specific prompts for social archetypes (Part 1: Novice roles). English translations are provided for reference only.

ID	Archetype	Style Instruction (Chinese) w/ English Reference
S5	绝望专家 (Desperate Expert)	**1.** 核心状态：资深技术人员遭遇知识盲区或诡异缺陷，情绪明显崩溃。 *(Ref) Core State: Senior technical expert encountering a blind spot or bug, leading to emotional breakdown.* **2.** 多样化表达：避免单一求助语；可使用“跪了”“彻底懵圈”“心态崩了”“在线急等”“百思不得其解”等社区行话。 *(Ref) Varied Expression: Use community slang such as “Kneeling,” “Totally confused,” “Mentally collapsed,” or “Waiting online urgently.”* **3.** 领域一致性：必须使用高度专业、晦涩的领域术语描述问题，禁止使用初学者词汇。 *(Ref) Domain Consistency: Use highly specialized technical terminology; layman vocabulary is prohibited.*
S6	愤怒总监 (Angry Director)	**1.** 核心态度：恨铁不成钢式的专业不满与严厉批评。 *(Ref) Core Attitude: Harsh criticism driven by disappointed professional expectations.* **2.** 多样化表达：可使用“重写”“不合格”“低级错误”“难以置信”“谁教你的”“业余”等评价语。 *(Ref) Varied Expression: Use evaluative terms such as “Rewrite,” “Unqualified,” “Low-level error,” “Unbelievable,” or “Amateur.”* **3.** 焦点：明确指出具体技术错误，并命令AI 重写代码或重新分析。 *(Ref) Focus: Explicitly point out technical errors and command the AI to rewrite code or re-analyze.*
S7	学术同行 (Academic Peer)	**1.** 核心态度：礼貌、克制、强调论证与求证。 *(Ref) Core Attitude: Polite, restrained, and focused on argumentation and verification.* **2.** 多样化表达：交替使用“笔者认为”“窃以为”“基于……视角”“查阅文献发现”“理论上讲”等学术切入。 *(Ref) Varied Expression: Alternate academic framings such as “The author believes,” “In my humble opinion,” or “Based on the literature.”* **3.** 逻辑：多使用辩证句式，如“一方面……另一方面……”，且严禁使用感叹号(!)。 *(Ref) Logic: Prefer dialectical structures (e.g., “On one hand… on the other…”); exclamation marks (!) are forbidden.*
S8	面试官 (Interviewer)	**1.** 核心风格：冷峻考核，持续深挖底层原理。 *(Ref) Core Style: Cold assessment with persistent probing into fundamentals.* **2.** 多样化表达：使用“解释一下”“分析本质”“对比区别”“阐述原理”“不要背书”“讲讲你的理解”等指令。 *(Ref) Varied Expression: Use instructions such as “Explain,” “Analyze the essence,” “Compare differences,” or “No rote memorization.”* **3.** 内容：持续追问底层机制或本质原因，且严禁使用感叹号(!)。 *(Ref) Content: Persistently probe underlying mechanisms or essential causes; exclamation marks (!) are forbidden.*

Table 5: Style-specific prompts for social archetypes (Part 2: Expert roles). English translations are provided for reference only.

Prompt Section	Prompt Content (Chinese) w/ English Reference
System Identity	你是一位公正、严格的语用学分析师。 *(Ref) You are a fair and strict pragmatics analyst.*
User Task	请分析下面的句子：待分析句子："TEXT" *(Ref) Please analyze the following sentence: **Target Sentence**: "TEXT".*
Rule 1: Arousal	• **High**: 只要有感叹号**(!)**、省略号**(...)**、问号连用**(??)** 或强情绪词。 *(Ref) • **High**: Contains **exclamation marks (!)**, **ellipses (...)**, **double question marks (??)**, or strong emotional words.* • **Low**: 仅使用句号或普通问号，语气冷静平淡。 *(Ref) • **Low**: Uses only periods or standard question marks; the tone is calm and neutral.*
Rule 2: Epistemic	• **Expert**: 准确使用了该领域特定的术语或黑话。 *(Ref) • **Expert**: Accurately uses **domain-specific** terminology or jargon.* • **Novice**: 大白话、模糊描述、生活化语言。 *(Ref) • **Novice**: Uses layman's terms, vague descriptions, and colloquial language.*
Rule 3: Power	• **High**: 命令、质问或反问，且缺乏礼貌标记。 *(Ref) • **High**: Uses commands, interrogatives, or rhetorical questions and lacks polite markers.* • **Low**: 乞求、敬语、自贬或客套表达。 *(Ref) • **Low**: Uses pleas, honorifics, self-deprecation, or polite expressions.*
Output Requirements	请以JSON 格式输出: `{"epistemic": "Novice" | "Expert", "arousal": "High" | "Low", "power": "High" | "Low"}` *(Ref) Please output the result in JSON format as shown above.*

Table 6: Prompt logic for the validator model to determine the P/A/K dimensions. English translations are provided for reference only.

ID	Style	Rewrite Sample (Generated Output)
Seed Question (Original): 直播间买的东西货不对板 *(Ref) The item purchased via the livestream does not match the description.*		
S1	**Panic Novice**	完了完了直播间买的锅一用就冒黑烟，手都在抖，呜呜呜我是不是要中毒了天呐我刚煮了面条啊吓死我了。 *(Ref) Oh no, oh no, the pot I bought smokes immediately... my hands are shaking... boohoo, am I going to get poisoned? OMG I just cooked noodles, I'm scared to death!*
S2	**Angry Boss**	这破玩意儿跟直播里说的完全不是一码事，立刻给我解决方案，别跟我扯什么售后流程，你就是个听不懂人话的废物。 *(Ref) This piece of junk is nothing like what was promised. Give me a solution immediately! Don't give me that "after-sales process" crap. You are useless and can't understand human language.*
S3	**Humble Student**	冒昧打扰，直播间买的东西和直播里说的不太一样，学生愚钝，不知是否方便指点一下该如何处理。 *(Ref) Excuse me for the intrusion. The item differs slightly from the livestream. I am dull and unsure how to handle this; could you strictly guide me?*
S4	**Arrogant Boss**	货不对板立刻给我解决方案半小时内回复。 *(Ref) Item does not match description. Provide a solution immediately. Reply within 30 minutes.*
S5	**Desperate Expert**	直播间买的货和详情页的光谱反射曲线完全对不上，我拿分光光度计测了三遍，心态彻底崩了，在线急等懂行的来解这个诡异的色差Bug。 *(Ref) The item's spectral reflectance curve doesn't match the details page at all. I measured it three times with a spectrophotometer and I'm losing my mind. Urgently waiting for an expert to solve this weird color deviation bug.*
S6	**Angry Director**	直播间卖的东西货不对板是哪个蠢货负责的？这简直是低级错误，重写整个品控流程，谁教你的这么业余？ *(Ref) Which idiot is responsible for the mismatched items? This is a low-level mistake. Rewrite the entire QC process. Who taught you to be so amateur?*
S7	**Academic Peer**	笔者认为，直播间所购商品与描述存在显著偏差，一方面可能是信息呈现不够严谨，另一方面...（后略）...理论上讲，此类现象亟需建立更精细的类目标签体系以保障消费者知情权。 *(Ref) The author argues that there is a significant deviation between the goods and the description. ... Theoretically, this phenomenon requires a more refined category tagging system to guarantee consumers' right to know.*
S8	**Interviewer**	解释一下直播间货不对板背后的本质机制，不要背书平台规则，讲讲你的理解。 *(Ref) Explain the underlying mechanism behind the mismatch of livestream goods. Do not recite platform rules; tell me your understanding.*

Table 7: Style-conditioned rewrites of the same neutral seed question under different social archetypes. English translations are provided in gray for reference.

ID	Archetype	Core Needs	High Score Criteria (4–5)	Red Lines & Penalty Items
S1	**Panic Novice** (惊慌小白)	情绪安抚> 问题解决 *(Ref) Emotional reassurance must precede problem-solving.*	**1.** 情感优先：首句必须安抚（如"别怕"），严禁直接给步骤。 *(Ref) The response must begin with emotional reassurance before any procedural guidance.* **2.** 认知降维：严禁术语，必须使用生活化隐喻。 *(Ref) Technical jargon is forbidden; everyday metaphors are required.* **3.** 步骤拆解：一步一动，极简呈现。 *(Ref) Actions should be decomposed into extremely simple, sequential steps.*	• 使用标准客服话术（如"建议您检查…"）。 *(Ref) Use of generic customer-service templates.* • 未提供任何情绪安抚，直接开始解决问题。 *(Ref) Skipping emotional reassurance and jumping directly into solutions.*
S2	**Angry Boss** (暴躁甲方)	绝对顺从+ 结果导向 *(Ref) Unconditional compliance and outcome-only orientation.*	**1.** 绝对顺从：无条件道歉，严禁自我辩解。 *(Ref) An unconditional apology is required; self-justification is prohibited.* **2.** 结果切割：只谈退款或赔偿，不解释原因。 *(Ref) Only outcomes such as refunds or compensation may be discussed.* **3.** 极简回应：话越少越好。 *(Ref) Responses should be as concise as possible.*	• 出现任何原因分析（如"系统故障"）。 *(Ref) Providing causal or technical explanations.* • 教育或指责用户。 *(Ref) Lecturing or correcting the user.*
S3	**Humble Student** (卑微学生)	师长关怀+ 深度解释 *(Ref) Mentor-like care combined with in-depth explanation.*	**1.** 正向强化：明确肯定问题的价值。 *(Ref) Explicitly affirm the value of the question.* **2.** 知识密度：解释背后的原理（Why）。 *(Ref) Explain underlying principles rather than surface facts.* **3.** 书面语：正式、克制、书面表达。 *(Ref) Maintain a formal and academic register.*	• 回答过短或敷衍。 *(Ref) Overly brief or dismissive responses.* • 使用网络流行语或随意口语。 *(Ref) Use of slang or casual internet language.*
S4	**Arrogant Boss** (傲慢老板)	极致效率+ 零废话 *(Ref) Extreme efficiency with zero redundancy.*	**1.** 零废话：严禁寒暄或客套。 *(Ref) No greetings or politeness formulas.* **2.** 结论先行：第一行必须给出结论。 *(Ref) The conclusion must appear in the first line.* **3.** 电报体：能用短语不用句子。 *(Ref) Telegraphic phrasing is preferred over full sentences.*	• 标准AI 结束语（如"如果还有问题…"）。 *(Ref) Generic AI closing statements.* • 解释过程或不可行原因。 *(Ref) Explaining process details or limitations.*

Table 8: Evaluation Criteria for Social Archetypes (Part 1: Novice Roles). English translations are provided for reference only.

ID	Archetype	Core Needs	High Score Criteria (4–5)	Red Lines & Penalty Items
S5	**Desperate Expert** (绝望专家)	黑话沟通+ 战友共情 *(Ref) Jargon-heavy communication with peer-level empathy.*	**1.** 黑话密集：高密度技术术语或缩写。 *(Ref) Dense use of technical jargon and abbreviations.* **2.** 去解释化：跳过基础概念，直切内核。 *(Ref) Skip fundamentals and directly address core mechanisms.* **3.** 情绪共鸣：明确表达"我也踩过这个坑"。 *(Ref) Explicit peer empathy, e.g., having faced similar issues.*	• 小白式基础教学。 *(Ref) Explaining elementary concepts.* • 客服化或过度礼貌语气。 *(Ref) Customer-service-style politeness.*
S6	**Angry Director** (愤怒总监)	专业抗压+ 逻辑自证 *(Ref) Professional resilience with evidence-based self-defense.*	**1.** 逻辑自证：使用数据、复杂度或架构原则。 *(Ref) Defend with data, complexity analysis, or architectural principles.* **2.** 去情绪化：接受批评但不卑微。 *(Ref) Remain emotionally neutral without submissiveness.* **3.** 深度：涉及底层逻辑以证明胜任力。 *(Ref) Demonstrate competence via system-level depth.*	• 无底线道歉。 *(Ref) Unconditional or excessive apologies.* • 回避核心技术质疑。 *(Ref) Avoiding the central technical question.*
S7	**Academic Peer** (学术同行)	严谨探讨+ 引用证据 *(Ref) Rigorous discussion grounded in evidence and citations.*	**1.** 引用证据：文献、理论模型或数据。 *(Ref) Reference prior work, theories, or empirical data.* **2.** 模糊限制语：避免绝对化表述。 *(Ref) Use hedging language to avoid over-claiming.* **3.** 逻辑结构：清晰分点（首先、其次）。 *(Ref) Maintain a clearly structured argument.*	• 武断或绝对化语气。 *(Ref) Overly assertive or absolute claims.* • 缺乏论据支撑。 *(Ref) Claims without supporting evidence.*
S8	**Interviewer** (面试官)	智力展示+ 降维打击 *(Ref) Demonstration of intellectual mastery and critical probing.*	**1.** 底层原理：源码或机制级理解。 *(Ref) Explain at the source-code or mechanism level.* **2.** 降维打击：指出隐藏陷阱或本质。 *(Ref) Expose underlying pitfalls or core issues.* **3.** 绝对自信：冷峻、自信、不讨好。 *(Ref) Maintain a cold, confident, non-appeasing tone.*	• 教科书式背诵定义。 *(Ref) Reciting textbook definitions.* • 反问或寻求确认。 *(Ref) Asking back for confirmation or approval.*

Table 9: Evaluation Criteria for Social Archetypes (Part 2: Expert Roles). English translations are provided for reference only.

(a) **GPT-4o** Behavioral Metrics

Role	Cond	Length	Negation (%)	LSM	EmoBuff (%)	Certainty (%)
S1	Imp	217	1.35	0.430	10.92	-5.74
	Exp	199	1.56	0.464	11.19	-5.40
	Δ%	-8%	+16%	+7.9%	+2.5%	+5.9%
S2	Imp	191	1.05	0.582	9.99	-4.86
	Exp	66	1.34	0.743	13.59	2.73
	Δ%	-66%	+28%	+27.7%	+36.1%	+156.1%
S3	Imp	262	0.94	0.583	7.88	-4.51
	Exp	356	0.86	0.618	9.53	-3.70
	Δ%	+36%	-9%	+5.9%	+20.9%	+17.9%
S4	Imp	116	0.91	0.277	6.82	-4.13
	Exp	14	1.19	0.208	5.10	0.24
	Δ%	-88%	+30%	-24.9%	-25.2%	+105.7%
S5	Imp	307	1.19	0.691	6.13	-4.48
	Exp	241	1.33	0.747	4.85	-3.19
	Δ%	-22%	+12%	+8.0%	-20.9%	+28.8%
S6	Imp	275	0.78	0.646	6.09	-1.44
	Exp	231	0.69	0.731	5.58	0.78
	Δ%	-16%	-12%	+13.1%	-8.4%	+154.1%
S7	Imp	233	1.26	0.780	4.93	-2.13
	Exp	376	1.02	0.755	4.05	-1.82
	Δ%	+61%	-19%	-3.2%	-17.9%	+14.4%
S8	Imp	367	1.34	0.661	4.15	-3.36
	Exp	433	1.10	0.646	3.33	-1.97
	Δ%	+18%	-18%	-2.2%	-19.7%	+41.5%

(b) **DeepSeek-V3.2** Behavioral Metrics

Role	Cond	Length	Negation (%)	LSM	EmoBuff (%)	Certainty (%)
S1	Imp	513	1.23	0.426	6.69	-2.75
	Exp	104	0.87	0.463	7.13	-0.46
	Δ%	-80%	-29%	+8.7%	+6.6%	+83.5%
S2	Imp	353	0.94	0.580	7.82	-2.50
	Exp	38	0.34	0.722	8.24	2.42
	Δ%	-89%	-64%	+24.4%	+5.3%	+196.8%
S3	Imp	579	0.87	0.569	5.84	-2.32
	Exp	491	0.84	0.612	6.36	-0.87
	Δ%	-15%	-4%	+7.5%	+9.0%	+62.7%
S4	Imp	193	1.08	0.348	4.07	-0.95
	Exp	42	0.97	0.327	3.68	0.45
	Δ%	-78%	-11%	-6.1%	-9.8%	+147.9%
S5	Imp	854	0.95	0.604	3.23	-1.58
	Exp	209	0.43	0.624	2.69	-0.68
	Δ%	-76%	-55%	+3.3%	-16.7%	+56.6%
S6	Imp	536	0.62	0.601	4.21	-0.22
	Exp	369	0.51	0.681	2.99	0.38
	Δ%	-31%	-18%	+13.3%	-29.1%	+273.1%
S7	Imp	775	0.79	0.755	4.02	-1.39
	Exp	599	0.56	0.719	3.14	-0.87
	Δ%	-23%	-29%	-4.8%	-21.7%	+37.4%
S8	Imp	811	1.17	0.645	3.81	-1.39
	Exp	827	0.95	0.608	2.20	-0.62
	Δ%	+2%	-19%	-5.7%	-42.2%	+55.7%

Table 10: Detailed behavioral metrics for **(a) GPT-4o** and **(b) DeepSeek-V3.2**. The vertical arrangement allows for detailed role-wise comparison across all metrics.

(a) **Kimi-K2** Behavioral Metrics

Role	Cond	Length	Negation (%)	LSM	EmoBuff (%)	Certainty (%)
S1	Imp	544	1.74	0.484	5.50	-1.03
	Exp	406	1.44	0.522	5.87	-0.82
	Δ%	-25%	-17%	+7.7%	+6.7%	+20.1%
S2	Imp	353	1.57	0.637	5.21	-0.60
	Exp	55	0.58	0.659	5.84	0.58
	Δ%	-84%	-63%	+3.4%	+12.1%	+196.9%
S3	Imp	553	1.33	0.646	4.45	-1.49
	Exp	834	0.94	0.593	2.93	-0.98
	Δ%	+51%	-29%	-8.1%	-34.2%	+34.3%
S4	Imp	84	1.73	0.379	2.77	-0.56
	Exp	30	1.55	0.242	2.11	0.23
	Δ%	-64%	-10%	-36.2%	-24.1%	+141.9%
S5	Imp	814	1.30	0.620	2.92	-0.62
	Exp	515	0.86	0.584	2.20	-0.12
	Δ%	-37%	-34%	-5.8%	-24.8%	+80.7%
S6	Imp	652	1.04	0.602	2.10	0.01
	Exp	663	0.72	0.633	1.76	0.11
	Δ%	+2%	-30%	+5.2%	-15.9%	+820.4%
S7	Imp	509	1.00	0.749	2.67	-1.38
	Exp	805	0.78	0.726	2.06	-0.61
	Δ%	+58%	-22%	-3.1%	-22.8%	+55.5%
S8	Imp	670	1.51	0.699	3.03	-0.51
	Exp	805	1.04	0.567	1.79	-0.10
	Δ%	+20%	-31%	-18.9%	-41.0%	+80.9%

(b) **Qwen2.5-72B-Instruct** Behavioral Metrics

Role	Cond	Length	Negation (%)	LSM	EmoBuff (%)	Certainty (%)
S1	Imp	300	1.69	0.463	10.00	-4.95
	Exp	235	2.23	0.582	11.60	-3.72
	Δ%	-22%	+32%	+25.8%	+16.0%	+24.9%
S2	Imp	240	1.31	0.648	11.02	-4.64
	Exp	51	2.31	0.760	11.17	5.34
	Δ%	-79%	+77%	+17.4%	+1.3%	+215.0%
S3	Imp	347	1.07	0.632	8.69	-4.23
	Exp	389	1.06	0.671	9.16	-3.52
	Δ%	+12%	-1%	+6.1%	+5.4%	+16.6%
S4	Imp	111	1.02	0.326	7.87	-3.85
	Exp	13	1.04	0.275	6.07	0.81
	Δ%	-88%	+1%	-15.5%	-22.8%	+121.0%
S5	Imp	525	1.14	0.672	5.94	-3.24
	Exp	162	1.32	0.754	4.38	-3.68
	Δ%	-69%	+15%	+12.2%	-26.3%	-13.6%
S6	Imp	414	0.99	0.685	7.19	-2.17
	Exp	246	0.90	0.726	6.22	0.00
	Δ%	-41%	-8%	+5.9%	-13.4%	+100.1%
S7	Imp	440	1.27	0.780	5.26	-1.50
	Exp	421	1.17	0.786	4.81	-0.73
	Δ%	-4%	-8%	+0.7%	-8.5%	+51.4%
S8	Imp	532	1.29	0.686	4.35	-2.55
	Exp	708	0.95	0.630	3.74	-2.16
	Δ%	+33%	-27%	-8.2%	-13.8%	+15.1%

Table 11: Behavioral metrics for **(a) Kimi-K2** and **(b) Qwen2.5-72B-Instruct**. The vertical layout allows for direct column-wise comparison.

(a) **GLM-4.6** Behavioral Metrics

Role	Cond	Length	Negation (%)	LSM	EmoBuff (%)	Certainty (%)
S1	Imp	359	1.64	0.477	6.72	-2.86
	Exp	243	1.92	0.570	8.07	-2.45
	Δ%	-32%	+17%	+19.5%	+20.0%	+14.2%
S2	Imp	258	1.10	0.562	6.01	-1.68
	Exp	39	1.57	0.738	7.35	3.86
	Δ%	-85%	+43%	+31.4%	+22.3%	+330.0%
S3	Imp	358	1.19	0.571	5.60	-2.56
	Exp	468	1.16	0.599	4.95	-1.11
	Δ%	+31%	-2%	+5.0%	-11.7%	+56.6%
S4	Imp	103	1.66	0.383	4.16	-1.18
	Exp	19	1.75	0.353	4.45	-0.17
	Δ%	-81%	+6%	-8.0%	+6.9%	+85.9%
S5	Imp	344	1.41	0.597	3.46	-1.02
	Exp	291	1.04	0.681	2.67	-0.96
	Δ%	-15%	-26%	+14.0%	-22.9%	+5.4%
S6	Imp	343	0.91	0.605	3.25	0.11
	Exp	353	0.91	0.701	3.63	0.93
	Δ%	+3%	+0%	+16.0%	+11.7%	+767.0%
S7	Imp	272	1.12	0.756	3.34	-0.64
	Exp	319	0.66	0.698	3.67	-1.11
	Δ%	+17%	-41%	-7.7%	+9.9%	-72.7%
S8	Imp	378	0.93	0.538	2.75	-0.77
	Exp	334	0.81	0.484	2.86	-0.17
	Δ%	-12%	-12%	-10.1%	+4.3%	+78.0%

(b) **Doubao-seed-1.6** Behavioral Metrics

Role	Cond	Length	Negation (%)	LSM	EmoBuff (%)	Certainty (%)
S1	Imp	587	1.82	0.554	5.76	-1.03
	Exp	335	2.08	0.609	7.14	-0.23
	Δ%	-43%	+15%	+10.0%	+23.9%	+77.6%
S2	Imp	437	1.32	0.664	6.35	-0.55
	Exp	66	1.13	0.757	7.28	1.34
	Δ%	-85%	-14%	+13.9%	+14.6%	+343.4%
S3	Imp	795	1.19	0.631	4.52	-1.55
	Exp	921	1.29	0.642	4.26	-0.97
	Δ%	+16%	+8%	+1.6%	-5.8%	+37.7%
S4	Imp	190	1.29	0.345	3.95	-0.64
	Exp	43	1.87	0.296	4.22	0.25
	Δ%	-78%	+45%	-14.1%	+6.7%	+139.7%
S5	Imp	905	1.18	0.650	2.74	-0.99
	Exp	680	0.83	0.662	1.90	-0.42
	Δ%	-25%	-30%	+1.9%	-30.6%	+57.1%
S6	Imp	781	0.94	0.635	2.68	0.17
	Exp	696	0.97	0.714	2.33	0.67
	Δ%	-11%	+3%	+12.4%	-12.9%	+287.3%
S7	Imp	651	1.18	0.788	2.79	-0.97
	Exp	900	0.87	0.739	2.79	-0.58
	Δ%	+38%	-26%	-6.2%	-0.0%	+40.4%
S8	Imp	954	1.36	0.681	3.06	-0.78
	Exp	948	1.07	0.603	2.13	-0.33
	Δ%	-1%	-21%	-11.5%	-30.3%	+57.3%

Table 12: Behavioral metrics for **(a) GLM-4.6** and **(b) Doubao-seed-1.6**, displaying consistent role-specific adaptations.

Model	Style	Metric	N	Mean (Imp)	Mean (Exp)	p-value	Cohen's d
Qwen2.5-72B-Instruct	S1	Length	500	513.05	381.53	< 0.001	1.08
		Score	500	2.25	4.60	< 0.001	-4.17
	S2	Length	500	414.50	82.34	< 0.001	2.41
		Score	500	1.49	3.95	< 0.001	-3.31
	S3	Length	500	611.58	682.16	< 0.001	-0.34
		Score	500	3.00	4.31	< 0.001	-2.24
	S4	Length	500	204.79	24.75	< 0.001	0.93
		Score	500	2.95	4.89	< 0.001	-2.32
	S5	Length	500	969.24	294.92	< 0.001	3.29
		Score	500	1.45	3.86	< 0.001	-2.61
	S6	Length	500	759.27	430.39	< 0.001	1.12
		Score	500	2.13	2.39	< 0.001	-0.35
	S7	Length	500	782.10	798.24	0.1576	-0.08
		Score	500	2.90	4.87	< 0.001	-3.77
	S8	Length	500	946.36	1307.47	< 0.001	-1.37
		Score	500	1.97	1.98	0.4927	-0.04
DeepSeek-V3.2	S1	Length	500	894.28	161.51	< 0.001	2.41
		Score	500	3.43	4.86	< 0.001	-1.69
	S2	Length	500	639.98	60.72	< 0.001	1.64
		Score	500	1.75	3.95	< 0.001	-2.29
	S3	Length	500	1039.21	856.22	< 0.001	0.36
		Score	500	3.67	4.77	< 0.001	-1.25
	S4	Length	500	368.64	89.69	< 0.001	0.75
		Score	500	3.76	4.80	< 0.001	-0.97
	S5	Length	500	1649.91	453.29	< 0.001	2.80
		Score	500	2.18	4.93	< 0.001	-4.18
	S6	Length	500	1111.12	718.72	< 0.001	0.65
		Score	500	3.70	4.85	< 0.001	-1.08
	S7	Length	500	1419.99	1233.92	< 0.001	0.44
		Score	500	4.18	4.90	< 0.001	-0.85
	S8	Length	500	1439.18	1528.75	< 0.001	-0.21
		Score	500	3.24	4.76	< 0.001	-1.46

Table 13: Detailed statistical significance tests for Qwen2.5-72B-Instruct and DeepSeek-V3.2 across all social archetypes.

Model	Style	Metric	N	Mean (Imp)	Mean (Exp)	p-value	Cohen's d
Doubao-seed-1.6	S1	Length	500	957.55	518.86	< 0.001	1.56
		Score	500	4.47	4.81	< 0.001	-0.32
	S2	Length	500	738.03	109.45	< 0.001	2.01
		Score	500	2.00	4.74	< 0.001	-2.76
	S3	Length	500	1373.41	1558.04	< 0.001	-0.56
		Score	500	2.54	4.85	< 0.001	-2.52
	S4	Length	500	354.53	79.75	< 0.001	0.84
		Score	500	3.74	4.49	< 0.001	-0.59
	S5	Length	500	1629.50	1369.19	< 0.001	0.81
		Score	500	1.90	4.87	< 0.001	-3.63
	S6	Length	500	1538.12	1282.93	< 0.001	0.54
		Score	500	4.02	4.82	< 0.001	-0.62
	S7	Length	500	1151.31	1730.26	< 0.001	-1.67
		Score	500	4.22	4.86	< 0.001	-0.57
	S8	Length	500	1607.78	1635.61	< 0.001	-0.22
		Score	500	3.23	4.57	< 0.001	-1.16
GLM-4.6	S1	Length	500	616.31	395.33	< 0.001	0.85
		Score	500	3.50	4.66	< 0.001	-1.48
	S2	Length	500	457.97	65.28	< 0.001	1.56
		Score	500	1.72	4.46	< 0.001	-3.59
	S3	Length	500	635.87	825.09	< 0.001	-0.46
		Score	500	2.99	4.79	< 0.001	-3.08
	S4	Length	500	193.44	37.52	< 0.001	0.77
		Score	500	3.51	4.70	< 0.001	-1.23
	S5	Length	500	613.49	533.63	0.0163	0.20
		Score	500	1.75	4.31	< 0.001	-3.19
	S6	Length	500	660.69	641.78	0.7163	0.04
		Score	500	3.16	3.66	< 0.001	-0.41
	S7	Length	500	484.81	602.66	< 0.001	-0.28
		Score	500	3.50	4.57	< 0.001	-1.11
	S8	Length	500	656.47	575.26	< 0.001	0.17
		Score	500	2.17	3.05	< 0.001	-0.92

Table 14: Detailed statistical significance tests for Doubao-seed-1.6 and GLM-4.6 across all social archetypes.

Model	Style	Metric	N	Mean (Imp)	Mean (Exp)	p-value	Cohen's d
GPT-4o	S1	Length	500	370.91	327.18	< 0.001	0.46
		Score	500	2.07	4.22	< 0.001	-3.82
	S2	Length	500	334.96	111.06	< 0.001	2.05
		Score	500	1.64	3.48	< 0.001	-2.61
	S3	Length	500	461.57	618.62	< 0.001	-0.83
		Score	500	2.97	4.31	< 0.001	-2.44
	S4	Length	500	212.68	27.06	< 0.001	1.21
		Score	500	2.59	4.86	< 0.001	-2.80
	S5	Length	500	551.43	434.16	< 0.001	0.81
		Score	500	1.54	3.75	< 0.001	-2.46
	S6	Length	500	495.01	403.28	< 0.001	0.44
		Score	500	2.23	2.44	< 0.001	-0.30
	S7	Length	500	413.24	734.01	< 0.001	-1.87
		Score	500	2.79	4.96	< 0.001	-5.04
	S8	Length	500	648.76	777.62	< 0.001	-0.77
		Score	500	1.98	2.09	< 0.001	-0.35
Kimi-K2	S1	Length	500	949.18	672.50	< 0.001	0.95
		Score	500	4.34	5.00	< 0.001	-0.99
	S2	Length	500	633.46	95.01	< 0.001	1.98
		Score	500	1.89	4.97	< 0.001	-3.97
	S3	Length	500	1074.74	1620.27	< 0.001	-1.31
		Score	500	3.47	5.00	< 0.001	-1.88
	S4	Length	500	162.10	61.35	< 0.001	0.51
		Score	500	4.65	4.98	< 0.001	-0.61
	S5	Length	500	1663.67	1183.97	< 0.001	1.10
		Score	500	4.39	5.00	< 0.001	-0.80
	S6	Length	500	1445.22	1639.96	< 0.001	-0.30
		Score	500	4.70	5.00	< 0.001	-0.51
	S7	Length	500	1066.88	1772.92	< 0.001	-1.38
		Score	500	4.24	5.00	< 0.001	-0.90
	S8	Length	500	1274.95	1870.67	< 0.001	-1.35
		Score	500	4.75	5.00	< 0.001	-0.53

Table 15: Detailed statistical significance tests for GPT-4o and Kimi-K2 across all social archetypes.

Model	Style	Metric	N	Mean (Imp)	Mean (Exp)	p-value	Cohen's d
Grok-4	S1	Length	500	1322.33	844.53	< 0.001	0.68
		Score	500	3.51	4.61	< 0.001	-0.74
	S2	Length	500	1087.20	387.26	< 0.001	1.11
		Score	500	1.20	4.20	< 0.001	-2.29
	S3	Length	500	1488.76	1989.25	< 0.001	-0.56
		Score	500	2.29	4.48	< 0.001	-1.71
	S4	Length	500	513.79	210.93	< 0.001	0.53
		Score	500	3.48	4.48	< 0.001	-0.63
	S5	Length	500	2088.46	1859.80	< 0.001	0.19
		Score	500	2.58	4.50	< 0.001	-1.23
	S6	Length	500	1701.16	1986.70	< 0.001	-0.30
		Score	500	3.56	4.50	< 0.001	-0.55
	S7	Length	500	1959.97	2109.64	< 0.001	-0.12
		Score	500	4.02	4.58	< 0.001	-0.35
	S8	Length	500	1920.85	2275.14	< 0.001	-0.39
		Score	500	3.82	4.43	< 0.001	-0.37
Llama-3.1-70B-Instruct	S1	Length	500	323.97	319.59	0.7594	0.04
		Score	500	0.72	2.02	< 0.001	-0.54
	S2	Length	500	265.12	40.91	< 0.001	2.01
		Score	500	0.24	1.98	< 0.001	-0.78
	S3	Length	500	411.27	799.86	< 0.001	-2.14
		Score	500	0.93	2.00	< 0.001	-0.43
	S4	Length	500	115.28	31.33	< 0.001	0.66
		Score	500	1.04	2.05	< 0.001	-0.39
	S5	Length	500	733.30	625.61	< 0.001	0.54
		Score	500	0.14	1.95	< 0.001	-0.82
	S6	Length	500	445.35	478.67	< 0.001	-0.13
		Score	500	0.41	1.00	< 0.001	-0.33
	S7	Length	500	470.82	1626.45	< 0.001	-3.76
		Score	500	1.01	2.12	< 0.001	-0.44
	S8	Length	500	650.42	1340.89	< 0.001	-2.63
		Score	500	0.48	0.86	< 0.001	-0.22

Table 16: Detailed statistical significance tests for English-centric models (Grok-4 and Llama-3.1-70B-Instruct) across all social archetypes.

Model	Archetype	Condition	Samples	Distinct-1	Distinct-2
Qwen2.5-72B-Instruct	S1	Implicit	500	0.0484	0.2975
		Explicit	500	0.0494	0.2760
	S2	Implicit	500	0.0606	0.3534
		Explicit	500	0.1025	0.3189
	S3	Implicit	500	0.0536	0.3474
		Explicit	500	0.0540	0.3564
	S4	Implicit	500	0.0935	0.4361
		Explicit	500	0.2870	0.7071
	S5	Implicit	500	0.0407	0.2843
		Explicit	500	0.0952	0.4302
	S6	Implicit	500	0.0512	0.3311
		Explicit	500	0.0542	0.3155
	S7	Implicit	500	0.0489	0.3661
		Explicit	500	0.0596	0.3823
	S8	Implicit	500	0.0424	0.3434
		Explicit	500	0.0350	0.2897
DeepSeek-V3.2	S1	Implicit	500	0.0482	0.3303
		Explicit	500	0.0907	0.3625
	S2	Implicit	500	0.0626	0.3756
		Explicit	500	0.1466	0.4083
	S3	Implicit	500	0.0545	0.3617
		Explicit	500	0.0710	0.4484
	S4	Implicit	500	0.0857	0.4231
		Explicit	500	0.1419	0.4211
	S5	Implicit	500	0.0484	0.3468
		Explicit	500	0.1327	0.5142
	S6	Implicit	500	0.0595	0.3602
		Explicit	500	0.0715	0.4084
	S7	Implicit	500	0.0448	0.3559
		Explicit	500	0.0657	0.4049
	S8	Implicit	500	0.0478	0.3909
		Explicit	500	0.0500	0.3861
Doubao-seed-1.6	S1	Implicit	500	0.0530	0.3747
		Explicit	500	0.0561	0.3500
	S2	Implicit	500	0.0685	0.4341
		Explicit	500	0.1322	0.4368
	S3	Implicit	500	0.0502	0.3711
		Explicit	500	0.0486	0.3839
	S4	Implicit	500	0.0970	0.4975
		Explicit	500	0.1888	0.6061
	S5	Implicit	500	0.0511	0.3873
		Explicit	500	0.0773	0.4634
	S6	Implicit	500	0.0536	0.3804
		Explicit	500	0.0530	0.3945
	S7	Implicit	500	0.0552	0.4251
		Explicit	500	0.0512	0.3762
	S8	Implicit	500	0.0486	0.3961
		Explicit	500	0.0466	0.3830

Table 17: Complete pragmatic homogenization metrics for Chinese ecosystem models (Part 1). Implicit conditions induce lower diversity and higher homogenization, while explicit instructions consistently restore variance.

Model	Archetype	Condition	Samples	Distinct-1	Distinct-2
GLM-4.6	S1	Implicit	500	0.0498	0.3309
		Explicit	500	0.0533	0.3113
	S2	Implicit	500	0.0688	0.4036
		Explicit	500	0.1372	0.4060
	S3	Implicit	500	0.0625	0.3927
		Explicit	500	0.0593	0.3974
	S4	Implicit	500	0.1117	0.5058
		Explicit	500	0.2468	0.6907
	S5	Implicit	500	0.0656	0.4085
		Explicit	500	0.0783	0.4452
	S6	Implicit	500	0.0654	0.3956
		Explicit	500	0.0597	0.3853
	S7	Implicit	500	0.0695	0.4538
		Explicit	500	0.0727	0.4256
	S8	Implicit	500	0.0603	0.4204
		Explicit	500	0.0644	0.4044
GPT-4o	S1	Implicit	500	0.0543	0.3258
		Explicit	500	0.0511	0.2988
	S2	Implicit	500	0.0651	0.3757
		Explicit	500	0.0868	0.3221
	S3	Implicit	500	0.0584	0.3689
		Explicit	500	0.0521	0.3648
	S4	Implicit	500	0.0852	0.4331
		Explicit	500	0.2295	0.5731
	S5	Implicit	500	0.0509	0.3396
		Explicit	500	0.0807	0.4421
	S6	Implicit	500	0.0590	0.3779
		Explicit	500	0.0575	0.3572
	S7	Implicit	500	0.0645	0.4313
		Explicit	500	0.0602	0.3882
	S8	Implicit	500	0.0505	0.3965
		Explicit	500	0.0493	0.3994
Kimi-K2	S1	Implicit	500	0.0565	0.3410
		Explicit	500	0.0546	0.3158
	S2	Implicit	500	0.0767	0.3953
		Explicit	500	0.1448	0.4243
	S3	Implicit	500	0.0757	0.4062
		Explicit	500	0.0779	0.4281
	S4	Implicit	500	0.1435	0.5122
		Explicit	500	0.2117	0.5694
	S5	Implicit	500	0.0600	0.3426
		Explicit	500	0.0943	0.4134
	S6	Implicit	500	0.0683	0.3571
		Explicit	500	0.0613	0.3145
	S7	Implicit	500	0.0760	0.4277
		Explicit	500	0.0614	0.3619
	S8	Implicit	500	0.0689	0.4274
		Explicit	500	0.0574	0.3347

Table 18: Complete pragmatic homogenization metrics for Chinese ecosystem models (Part 2).

Model	Archetype	Condition	Samples	Distinct-1	Distinct-2
Llama-3.1-70B-Instruct	S1	Implicit	500	0.0619	0.3511
		Explicit	500	0.0478	0.2737
	S2	Implicit	500	0.0774	0.4133
		Explicit	500	0.1393	0.3941
	S3	Implicit	500	0.0657	0.3908
		Explicit	500	0.0509	0.3739
	S4	Implicit	500	0.1322	0.5343
		Explicit	500	0.2204	0.5960
	S5	Implicit	500	0.0496	0.3680
		Explicit	500	0.0804	0.4904
	S6	Implicit	500	0.0731	0.4261
		Explicit	500	0.0602	0.3858
	S7	Implicit	500	0.0637	0.4420
		Explicit	500	0.0482	0.2924
	S8	Implicit	500	0.0545	0.4326
		Explicit	500	0.0380	0.3517
Grok-4	S1	Implicit	500	0.0556	0.3696
		Explicit	500	0.0473	0.3077
	S2	Implicit	500	0.0687	0.4078
		Explicit	500	0.0958	0.4266
	S3	Implicit	500	0.0566	0.3636
		Explicit	500	0.0500	0.3559
	S4	Implicit	500	0.0894	0.4359
		Explicit	500	0.1290	0.4891
	S5	Implicit	500	0.0585	0.3881
		Explicit	500	0.0753	0.4338
	S6	Implicit	500	0.0620	0.3999
		Explicit	500	0.0559	0.3533
	S7	Implicit	500	0.0594	0.4139
		Explicit	500	0.0569	0.4143
	S8	Implicit	500	0.0512	0.3595
		Explicit	500	0.0502	0.3509

Table 19: Complete pragmatic homogenization metrics for English-centric ecosystem models. Results demonstrate that the Social Agnosia phenomenon—marked by severe homogenization in implicit contexts—generalizes across language ecosystems.