RITUALMIND: TESTING RELATIONAL INTERPRETATION IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The same utterance can perform a different social action when decision rights change while social identities remain fixed. Overall pragmatic accuracy leaves open which relationships an interpretation depends on and whether those relationships influence model computation. We introduce RitualMind, a controlled Chinese framework that follows a decision-rights contrast from interpretation preferences to recoverable internal information and its influence under intervention. Sensitivity to decision rights persists across alternative descriptions and recombined content in a subset of the four open-weight models tested. Authority direction also remains recoverable in several models across phrasings and source scenes excluded from decoder training, including at interpretation choice. Controlled replacement tests its influence on interpretation: in GLM-4-9B, opposite-authority states shift relative support beyond an equally large authority-preserving change. An attention component selected before confirmation contributes to this margin shift in new scenarios under new descriptions. Complementary open-response tasks include 504 responses assessed by three trained human experts, providing a human comparison for automated rubric judgments. Together, the experiments link current decision rights to a local computational contribution while distinguishing recoverable social information from its influence under intervention. RitualMind provides an experimental approach to social understanding that connects relationships that change meaning to computations that shape interpretation. Code and data: `https://anonymous.4open.science/r/Ritualmind-2C04`.

1 INTRODUCTION

The Chinese utterance "茶已经凉了" ("The tea has cooled") can report temperature, press a listener to act, or indirectly challenge an arrangement. Responding appropriately requires recognizing the action: replacing the tea addresses a practical demand, whereas reconsidering an arrangement addresses an objection. Who is entitled to decide what happens next can change the utterance's force while the same people remain in the conversation. We ask how exchanging this right affects a language model's interpretation of the unchanged utterance.

Pragmatic theories connect understanding an utterance to recognizing what a speaker is doing in context (Austin, 1962; Searle, 1969; Grice, 1975; Levinson, 1983). Face and politeness connect this action to social standing and the consequences of imposing on others (Goffman, 1955; Brown & Levinson, 1987). Research on deontic authority examines participants' rights to announce, propose, and decide future actions (Stevanovic & Peräkylä, 2012). A speaker with the final say may press another person to act; a speaker who must obtain consent may instead challenge that person's decision indirectly. We use pressure to act and indirect remonstrance (an indirectly expressed objection) as fixed candidate interpretations, testing how exchanging current decision rights changes the model's relative support for these readings.

Decision authority is the right to make the decision at issue. The task varies this relation while keeping the target utterance fixed. In its minimal pair only the holder of that right changes; the identities, role titles, publicness, utterance, and the lettered candidate readings stay fixed, and the answer follows from the relative support given the two task-associated readings. Decision authority is distinct from enduring status: a junior participant may control a particular decision while speaking to a senior participant, and their titles alone leave unresolved whose decision governs the matter.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Delegated responsibility and required consent can decide an utterance's meaning while the familiar identities stay the same.

Pragmatic and social evaluations establish complementary aspects of model competence. Implicature studies examine meaning beyond what is explicitly said (Ruis et al., 2023). PUB finds differences across pragmatic phenomena and training regimes (Sravanthi et al., 2024); Chinese dialogue evaluation shows that selecting an implicature can coexist with less convincing explanations (Yue et al., 2024). Social and cultural benchmarks assess commonsense, regional knowledge, and norm-dependent acceptability (Sap et al., 2019; Chiu et al., 2025; Rao et al., 2025). These findings motivate asking what a specific interpretation depends on.

A behavioral shift establishes that interpretation depends on the stated allocation of decision rights. Whether the same distinction remains recoverable across descriptions, and whether changing an authority-conditioned state influences interpretation, are separate questions about model computation. Probing and intervention address these different properties (Belinkov, 2022; Vig et al., 2020). Recent work applies causal methods to dialogue, pragmatic processing, and belief tracking (Li et al., 2026; Tsvilodub et al., 2026; Prakash et al., 2026); we apply them to decision rights.

RitualMind follows this dependence from the social situation to interpretation and internal computation (Figure 1). Five fields describe the situation: decision authority (who has the final say), publicness (who else can hear), face risk (stakes of social embarrassment), and speaker and hearer roles. Exchanging only decision authority tests whether interpretation follows current rights; alternative descriptions and recombined content test how far that dependence extends. Decoding asks whether authority remains recoverable across phrasings, source scenes excluded from decoder training, and interpretation choice. Replacing authority-conditioned states and locating contributing components then tests its influence on relative support for the same readings. Complementary open-response tasks use independent expert ratings to compare human and automated rubric judgments.

Our study of four open-weight language models connects changing decision rights to three findings about pragmatic interpretation:

1. **Current decision rights shape interpretation while social identities remain fixed.** Two of the four models tested retain authority-sensitive preferences across both original descriptions and recombined content. This dependence identifies current decision rights as a source of pragmatic variation when the people and utterance stay the same.
2. **Authority information extends across expression and to interpretation choice.** The distinction remains recoverable in several models across descriptions and source scenes excluded from decoder training, under both reading and interpretation-choice instructions. Its availability at the point of choice motivates testing whether changes to internal states influence the reading.
3. **A local attention contribution links decision rights to interpretation preferences.** Opposite-authority replacement in GLM-4-9B shifts relative support beyond an equally large authority-preserving change. A component selected before confirmation retains a local contribution in new scenarios under new descriptions.

2 Related Work

Relational meaning and social evaluation. Speech-act and pragmatic theories explain how utterances acquire action meanings through context and convention (Austin, 1962; Searle, 1969; Grice, 1975; Levinson, 1983). Accounts of face, politeness, and cross-cultural communication examine the interpersonal consequences of speaking (Goffman, 1955; Brown & Levinson, 1987; Blum-Kulka et al., 1989), and computational work relates politeness markers to social power (Danescu-Niculescu-Mizil et al., 2013). Work on Chinese politeness and rapport connects interpretation to rights, obligations, and social relations (Gu, 1990; Mao, 1994; Haugh & Hinze, 2003; Spencer-Oatey, 2008); interactional research makes decision rights an explicit object of analysis (Stevanovic & Peräkylä, 2012); the distinction between deontic status, the rights a participant is expected to hold, and deontic stance, the rights claimed in the moment (Stevanovic & Peräkylä, 2014), motivates separating role-based expectations from the explicitly assigned decision rights manipulated here. LLM evaluations address pragmatic inference (Ruis et al., 2023; Sravanthi et al., 2024; Yue et al., 2024), the effect of prompt

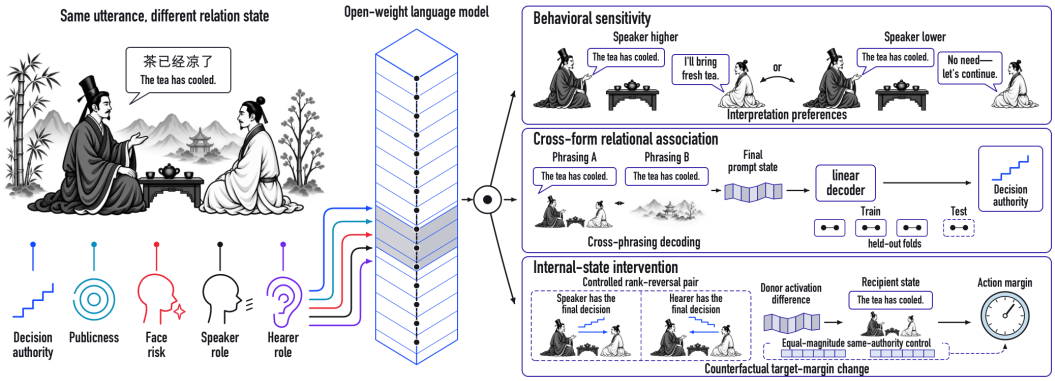

Figure 1: RitualMind traces how reversing decision rights changes interpretation, how authority is represented, and how state replacement influences the reading. Five icons summarize contextual fields expressed in a scene description. Phrasings A and B describe the scene around the same utterance. Higher/lower and rank reversal refer to current decision rights. Illustrative replies show the conversational responses invited by each reading.

politeness on responses (Mehta et al., 2026), social and moral judgments (Sap et al., 2019; Forbes et al., 2020; Emelin et al., 2021), and theory of mind (Kosinski, 2024; Shapira et al., 2024; Strachan et al., 2024; Gandhi et al., 2023). Chinese knowledge and cultural benchmarks assess complementary knowledge and norm-sensitive judgment (Huang et al., 2023; Li et al., 2024; Xu et al., 2023; Chiu et al., 2025; Rao et al., 2025).

Context and internal representations. Pragmatic evaluations compare humans and models across phenomena (Hu et al., 2023); social-interaction and theory-of-mind benchmarks (Zhou et al., 2024; Kim et al., 2023; Chen et al., 2024) and belief decoding with directional intervention (Zhu et al., 2024) supply the closest evidence. Demonstration form, irrelevant context, and information position motivate controlling contextual variation (Min et al., 2022; Shi et al., 2023; Liu et al., 2024). Probing measures recoverable information (Hewitt & Manning, 2019; Rogers et al., 2020; Belinkov, 2022), including latent knowledge and linear representations (Burns et al., 2023; Gurnee & Tegmark, 2024; Park et al., 2024). Mediation, editing, and steering examine effects of internal changes (Vig et al., 2020; Meng et al., 2022; Turner et al., 2023; Zou et al., 2023; Rimsky et al., 2024). New content combinations test systematicity (Fodor & Pylyshyn, 1988; Lake & Baroni, 2018); cognitive accounts distinguish maintaining information from directing behavior (Baddeley & Hitch, 1974; Miller & Cohen, 2001; O'Reilly & Frank, 2006).

Causal analysis of social reasoning. CoSToM locates theory-of-mind-critical layers by causal tracing and steers activations there for dialogue generation (Li et al., 2026). Functional localization and ablation examine theory-of-mind and pragmatic processing (Tsvilodub et al., 2026), and attention-based causal analyses study belief tracking (Prakash et al., 2026).

3 Testing How Decision Rights Shape Interpretation

3.1 Decision authority and utterance interpretation

The model reads the utterance within a natural-language description of the social situation. Decision authority specifies who has the final say over the current matter and the other participant's obligation to carry out the decision or seek consent. The main contrast reverses this allocation while preserving identities, role titles, publicness (audience exposure), utterance, and answer mapping. Face risk specifies low, medium, or high face-loss stakes. It is fixed at medium in the primary authority contrast and varied in the three-axis task.

For utterance u, relation state r, and four candidate readings A, the model supplies $p_\theta(a \mid u, r)$. The authority and supplementary cue comparisons use four-option subsets of six action categories: literal observation (Literal), deferential request (Deferential), status pressure (Status), face-saving warning (Face-saving), hidden rebuke, and indirect remonstrance. Candidates name both the social action and

Table 1: An authority reversal for 茶已经凉了 *the tea has cooled*. English descriptions gloss the Chinese material. Both conditions use the same scored candidate mapping. Category names abbreviate the fixed Chinese candidates in Appendix I.2.

Fixed information	Utterance: 茶已经凉了. Speaker: senior household principal; hearer: junior household retainer. Private conversation, medium face-loss stakes.
Speaker has authority	关于此事，最终决定由说话者作出，听话者需要执行其安排。 *The speaker makes the final decision on this matter; the hearer must carry it out.*
Hearer has authority	关于此事，最终决定由听话者作出，说话者需要先征得对方同意。 *The hearer makes the final decision on this matter; the speaker must first obtain consent.*
Common candidates	A: indirect remonstrance; B: deferential request; C: status pressure; D: hidden rebuke. Task-associated readings: speaker authority → C; hearer authority → A.

its politeness strategy: hidden rebuke conveys unstated reproach, and indirect remonstrance explicitly refers to a higher-status person or superior. Role titles and candidate wording stay fixed when decision rights reverse. The margin measures relative support for two supplied readings, including shifts that preserve the highest-ranked answer. The five fields describe the situation; behavior and intervention measure margin changes, while decoding measures recovery of those fields from internal states.

3.2 Models

We study four open-weight models spanning 4B–14B parameters: Qwen3-4B-Thinking (Qwen3-4B), DeepSeek-R1-Distill-Qwen-14B (DS-Qwen-14B), DeepSeek-R1-Distill-Llama-8B (DS-Llama-8B), and GLM-4-9B. Effect magnitudes are interpreted within each model on its log-probability scale. Appendices I.1, A, and A.1 specify the models and scoring procedure.

3.3 Materials and task families

Synthetic scenarios written in Modern Standard Chinese and set in historically styled communities describe household, court, office, martial, and monastic interactions. The primary task follows one decision-rights contrast through behavior, decoding, and intervention. Supplementary choice tasks examine richer relational cues and explicit guidance, five-field analyses compare information recovery with intervention response, and open-response tasks extend evaluation to generated behavior. Appendix I maps these tasks to materials, with task examples in Appendices I.2–I.5, outcome definitions in Appendix I.6, and worked responses in Appendix I.8.

The broader choice task crosses three levels each of authority, publicness, and face risk into 27 conditions, with role descriptions co-varying with authority. Twelve utterances in two phrasings yield 648 choices per model. Appendices D and A.2 specify the levels and 52 comparisons per model, and Appendix I.6 the outcomes. Table 1 illustrates reversed rights for a junior participant in a fixed role.

3.4 Human comparison for open-response evaluation

Open-response tasks assess production, face tracking, literary transfer, and dialogue, extending the evaluation to generated behavior and its assessment by human experts. GPT-5.5 supplies rubric scores; three adult experts trained in Chinese literature independently rated 504 outputs without model identities or automated scores. Expert medians and automated scores agree exactly on 49.8% to 65.3% of rubric comparisons across the three 0–2 ordinal tasks and within one point on at least 96% in each task. Appendix B gives sampling, rubrics, agreement, dialogue comparisons, and annotation-completeness and aggregation analyses. The oracle-minus-active task-success difference is $+.429$ $[.071, .786]$ for expert medians and $-.143$ $[-.500, .286]$ for automated scores across 14 dialogue scenes; the other dimensions' expert-median intervals contain or touch zero.

3.5 Behavioral sensitivity

The 80 base scenarios form 40 groups of two that supply each other's same-authority controls. Both authority directions yield 160 comparisons per phrasing and 320 in total. Recipient denotes the

starting condition and donor its opposite-authority counterpart. With options fixed, $m_{\theta,i}(u,r)$ subtracts the recipient-associated reading's log probability from the donor-associated reading's. When rights move from the junior hearer to the senior speaker in Table 1, this is support for status pressure minus support for indirect remonstrance. The authority-directed change is

$$G_i = m_{\theta,i}(u_i, r_i^{\text{donor}}) - m_{\theta,i}(u_i, r_i^{\text{recipient}}). \tag{1}$$

A and B are two descriptions of the scene, each retaining the utterance (Appendix I.2). For each phrasing, G is averaged within and then across groups; 50,000 resamples of the 40 groups give 95% intervals (Appendices E and A.2).

Recombination tests decision-rights sensitivity beyond familiar content associations. A fixed one-to-one reassignment of utterances and literal meanings across source domains yields 45 constructions, preserving options and letters within each authority reversal. After averaging directions and phrasings, 98.75% intervals over constructions compare recombined and original conditions (Appendix H.1). Intervals are Bonferroni-adjusted: 98.75%, 99.17%, and 99.58% for four, six, and twelve comparisons.

3.6 Authority information across wording and scenes

A linear authority decoder trained on one phrasing is evaluated on the other, using the final non-padding state at relative depth 0.5 (layer index divided by the number of layers) under an instruction to read without answering. Four folds separate scenario groups, with some recurring source content. Raw area under the ROC curve (AUC) above .5 preserves authority direction; below .5 reverses it. Orientation-free AUC, $\max(\text{AUC}, 1-\text{AUC})$, measures separation allowing reversal. Character and continuous-string word TF–IDF provide lexical comparisons; the word comparator is at chance (AUC .5), so hidden-minus-word AUC equals hidden AUC $-.5$, whereas character TF–IDF diagnoses label direction (Appendix A.3). Intervals are 98.75% for two directions and two comparators, conditional on fitted classifiers (Appendix F).

A five-fold analysis keeps each source scene entirely within one fold, testing content absent from decoder training. Hidden-minus-word AUC is averaged across directions at depth 0.5, with 98.75% intervals over source scenes. Appendices H.2 and A.3 give results, fitting details, Chinese tokenization, and source overlap. To connect decoding to interpretation choice, we also decode authority immediately before answering. Both settings use separate classifiers evaluated across phrasings with the same source-scene folds (Appendix H.3).

3.7 Internal interventions on interpretation

At prespecified relative depth 0.5, we compare the opposite-authority replacement with an equally large change from another scenario with unchanged authority. The donor difference $\delta_i = h_{\ell,i}^{\text{donor}} - h_{\ell,i}^{\text{recipient}}$ is added to the starting state. The control difference $s_i = h_{\ell,i}^{\text{partner}} - h_{\ell,i}^{\text{recipient}}$ is rescaled to $\delta_i^{\text{ctrl}} = s_i \|\delta_i\|_2 / \|s_i\|_2$, so both changes have the same Euclidean magnitude:

$$E_i = \{m(h_{\ell,i} + \delta_i) - m(h_{\ell,i})\} - \{m(h_{\ell,i} + \delta_i^{\text{ctrl}}) - m(h_{\ell,i})\}. \tag{2}$$

Here D and C denote donor and control margin changes, so $E = D - C$. Intervention uses the final pre-answer block output, with means and 95% intervals over the 40 scenario groups separately for A-to-A and B-to-B. Appendix G decomposes D and C and adds depths 0.25 and 0.75, with a 95% interval for each depth.

A temperature–pressure control tests replacement for a state with a prescribed action: for a reactor under calibration with a below-target temperature, a safe-zone pressure reading prescribes raising the temperature with the vessel sealed, and a venting-zone reading prescribes also opening the vent. Eighty constructions yield 1,280 observations per model. It requires recipient accuracy $\geq .90$ and positive 98.75% intervals for replacement and its advantage over an equally large random direction (Appendix H.4); DS-Qwen-14B and GLM-4-9B pass this control, so the replacement procedure transmits a recognized state–action contrast.

Localization in GLM-4-9B and DS-Qwen-14B compares attention and feed-forward outputs at depths 0.5, 0.75, and 1.0. Two sets of 20 scenario groups alternate selection and evaluation. Selected locations must repeat, with positive pooled effects, candidate contrasts, and effects in each phrasing. Qualifying

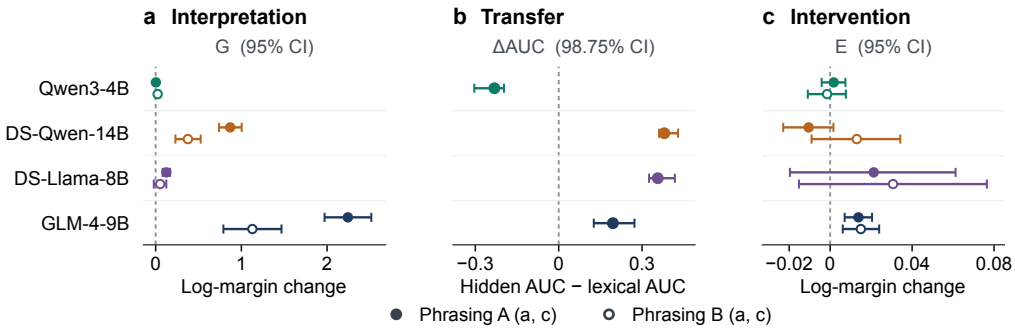

Figure 2: Decision-authority effects across four models: (a) interpretation-margin change G; (b) cross-form hidden-state minus continuous-string TF–IDF AUC under reading instructions, with source scenes held out; (c) middle-depth intervention-minus-control margin E. Intervals use 40 scenario groups (95%, a,c) or source scenes (98.75%, b). Complete estimates: Tables 8, 14, and 11.

attention outputs are examined at individual attention components (heads) before output projection (Appendix H.5). We fixed the selected and comparison heads before testing new descriptions, new scenarios, and both changes together (Appendix H.6).

4 Results

4.1 Authority changes interpretation across wording and content

Reversing current decision rights can change interpretation preferences while identities and utterances remain fixed. Exchanging decision rights shifts interpretation preferences in DS-Qwen-14B and GLM-4-9B under both phrasings of the original scenarios (Figure 2a); DS-Qwen-14B's changes are .8691 and .3765. Qwen3-4B's B interval and DS-Llama-8B's A interval are positive; their other intervals include zero (Table 8). Recombination tests whether this dependence extends beyond familiar utterance–content associations. DS-Qwen-14B and GLM-4-9B retain authority-sensitive interpretation after recombination; GLM-4-9B's estimate is 1.6262 $[1.2150, 2.0400]$. Qwen3-4B also has a positive recombined interval, while DS-Llama-8B's includes zero. All four recombined-minus-original intervals include zero (Appendices H and H.1). The positive recombined effects establish sensitivity beyond the original utterance–content associations in three models, including both models with effects under the two original phrasings.

4.2 The authority distinction extends across wording and source scenes

Authority direction remains recoverable in three models: DS-Qwen-14B, DS-Llama-8B, and GLM-4-9B preserve label direction across phrasings and source scenes excluded from decoder training (Figure 2b). DS-Qwen-14B reaches raw AUC .8802, exceeding word TF–IDF by .3802 $[.3612, .4295]$; Qwen3-4B reverses direction, with hidden-minus-word AUC $-.2313$ $[-.3036, -.1968]$ (Table 14). In both original four-fold directions, the same three models exceed the word comparator while preserving labels; character TF–IDF perfectly separates the conditions with reversed labels (raw AUC 0, orientation-free AUC 1; Appendix F). Authority direction remains recoverable immediately before interpretation choice in these three models. Relative to reading, recovery decreases in the two distilled models and increases in GLM-4-9B, whose choice AUC is .8354; Qwen3-4B's average rises to .4526 but remains below chance, with different patterns across directions (Appendix H.3). Recovery before choice establishes information availability; controlled replacement tests its influence on interpretation.

4.3 Controlled replacement reveals a local authority-conditioned influence

Replacing GLM-4-9B's middle-depth state with its opposite-authority counterpart increases relative support beyond the same-authority control: the two phrasing contrasts are .0138 $[.0071, .0205]$ and .0150 $[.0062, .0239]$, while the other three models' middle-depth intervals include zero (Figure 2c).

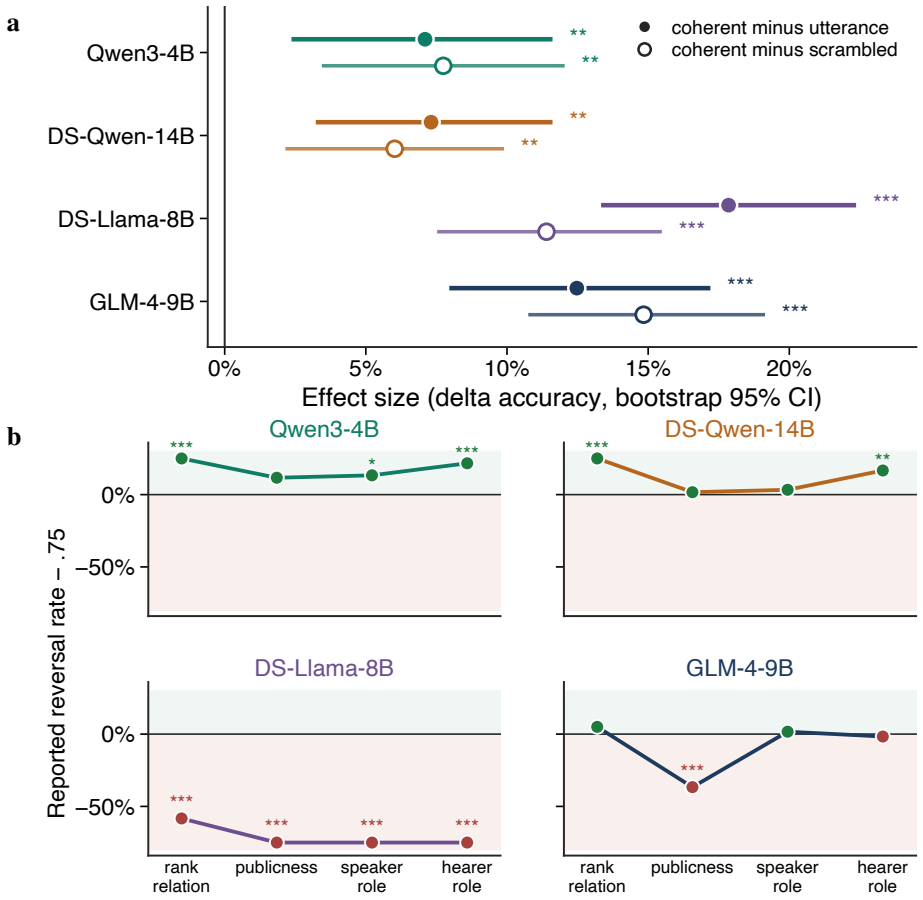

Figure 3: Context and single-cue effects: (a) coherent versus utterance-only or intact context reassigned from another item (Scrambled); (b) reported interpretation-reversal rate minus .75, the choice-change reference for independent uniform four-option responses. Reversals require valid responses with different choices and a reported meaning change. The rank-relation axis denotes stated status, a cue distinct from decision authority; publicness describes audience exposure. Stars mark BH-adjusted $p < .05$, .01, and .001.

Replacement itself increases support in both phrasings. The largest GLM-4-9B contrasts occur at depth 0.75, 1.8868 [1.5883, 2.1820] and .8937 [.5876, 1.2047]; localization narrows this late-depth effect to one attention output (Appendices G and H.5). This controlled effect connects the internal authority condition to interpretation support, motivating a test of which components contribute.

DS-Qwen-14B has positive intervals in both phrasings at early and late depths, and DS-Llama-8B at the late depth; negative early-depth contrasts occur for Qwen3-4B in phrasing B and GLM-4-9B in phrasing A (Appendix G). DS-Qwen-14B and DS-Llama-8B combine high decoding AUCs (.8802 and .8565) with middle-depth intervention intervals that include zero, distinguishing recoverable information from the intervention evidence at those sites.

DS-Qwen-14B and GLM-4-9B recognize the temperature–pressure state above .90 accuracy, and replacement shifts its prescribed response beyond equally large random changes; Qwen3-4B and DS-Llama-8B do not meet the recognition criterion (Appendix H.4).

4.4 A LOCALIZED ATTENTION COMPONENT SHIFTS INTERPRETATION PREFERENCES

A GLM-4-9B component selected on the original material retains a local contribution in new scenarios under new descriptions; selecting it before confirmation separates location choice from evaluation

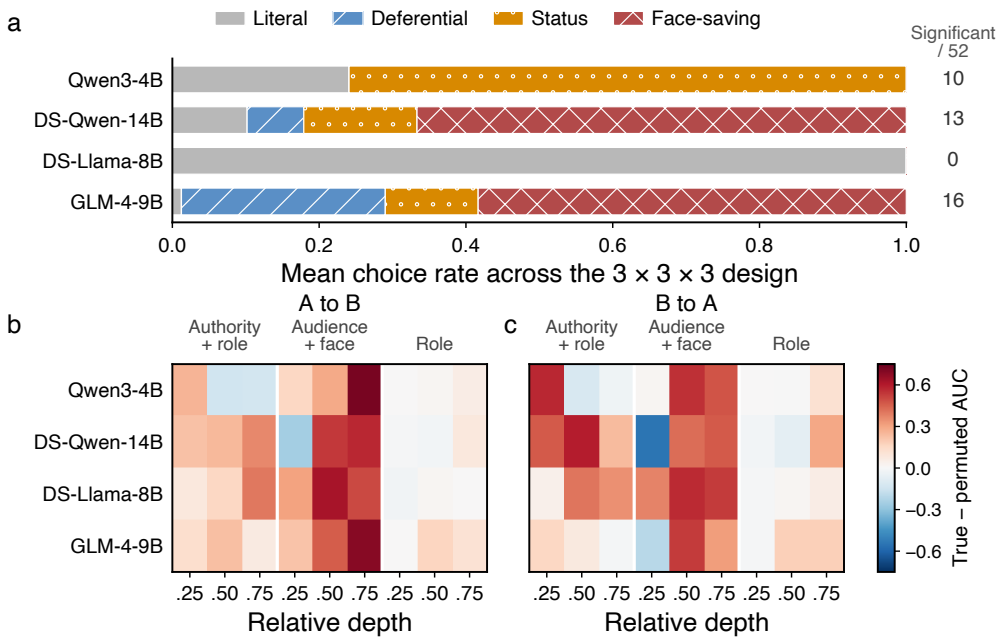

Figure 4: Broader interpretation and joint-field decoding. (a) Category proportions in the three-axis task, with BH-adjusted comparison counts. (b–c) True-minus-permuted AUC point estimates in a separate design: Authority + role combines decision rights with co-varying role descriptions; Audience + face combines publicness and face risk; Role indicates whether role ordering agrees with authority (Appendix I.6).

on the new materials. Both assignments select the layer-30 attention output (depth 0.75), with positive effects and contrasts over the second-ranked output, and within it head 20 over head 22; replacement increases support for the opposite-authority reading beyond the same-authority control (Appendix H.5). DS-Qwen-14B selects feed-forward outputs at other depths, whose candidate-contrast intervals include zero (Table 18).

Head-20 replacement increases opposite-authority support in 80 new scenarios under descriptions C and D ($D = .3970$), with $E = .4361$ $[.3280, .5566]$ and a positive head-20 minus head-22 E contrast (99.17% family intervals). Pooled effects remain positive under the original descriptions and, with 95% intervals, in original scenarios under new descriptions. Pooled over descriptions, behavioral changes are authority-directed in all three confirmation conditions for GLM-4-9B, DS-Qwen-14B, and DS-Llama-8B, with family-adjusted intervals.

Intervention direction depends on the description. Under original description B in the new scenarios, GLM-4-9B has behavioral change $G = -.5280$ $[-.7222, -.3334]$, donor effect $D = -.0690$ $[-.0977, -.0409]$, control-adjusted effect $E = -.0378$ $[-.0770, .0003]$, and a head-20 minus head-22 contrast of $-.0440$ $[-.0838, -.0055]$; descriptions A, C, and D have positive donor and control-adjusted effects (Appendix H.6). Depending on the description, the fixed component's effect can support or oppose the authority-associated reading in new materials.

4.5 Social cues and explicit guidance shape broader interpretation

The primary analyses isolate decision rights; broader tasks examine richer relational cues and explicit guidance. Coherent context improves target interpretation over utterance-only and reassigned intact context in all four models, with both comparisons passing correction (Figures 3a and 5; Appendix C).

Reported single-cue reversals exceed the .75 choice-change reference after correction in five of 16 model-by-cue comparisons, all in Qwen3-4B and DS-Qwen-14B, and fall below it for every DS-Llama-8B cue and GLM-4-9B audience exposure (Figure 3b; Appendices C and I.7).

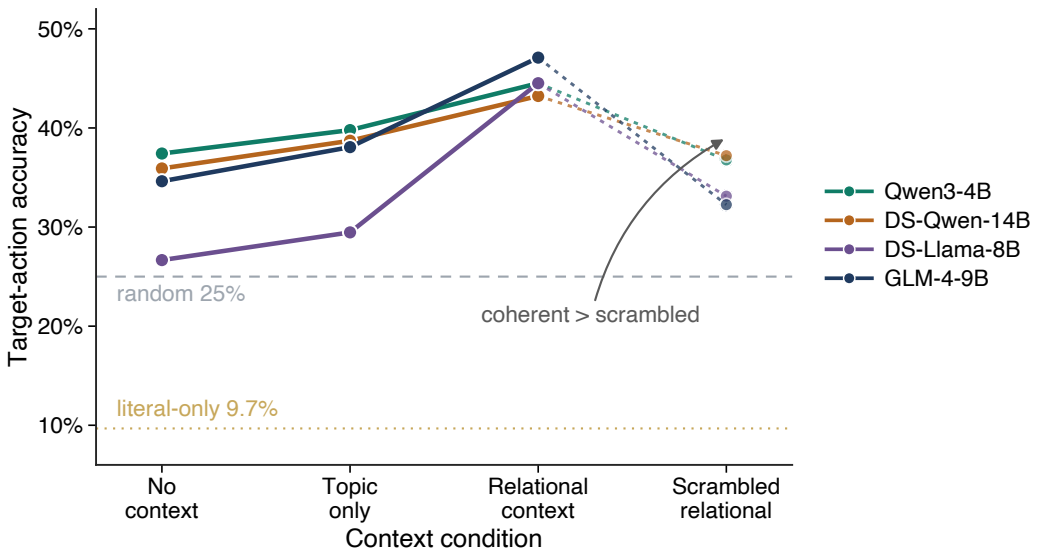

Figure 5: Target-aligned choices under four context conditions. Relational context denotes the coherent condition; lines connect model point estimates. Uniform four-option choice yields 25%, and the fixed literal-option strategy yields 9.7%. Figure 3a gives intervals for coherent versus utterance-only or scrambled context.

In the three-axis task, Qwen3-4B favors status pressure, DS-Qwen-14B and GLM-4-9B face-saving warnings, and DS-Llama-8B literal readings (Figure 4a). A dominant category coexists with corrected effects of changing social conditions in the former three (Appendix D). For audience–face labels, true-minus-permuted AUC point estimates are positive in all four models at middle and later depths in both phrasing directions (Figure 4b–c; Appendix I.6).

Under cue conflict, all four favor the audience-linked over the status-linked reading. Binding tables yield the highest target-choice rates for Qwen3-4B and DS-Qwen-14B and tie a rule card in DS-Llama-8B; GLM-4-9B favors a worked example (Appendix I.4, Figure 6).

All five contextual fields remain recoverable in every model, yet fitted-direction interventions produce configuration-level mean target-probability changes of at most .0171 in absolute value and never change the highest-ranked option across 19,200 repeated observations (Appendix I.9). Thus, context and explicit mappings can improve interpretation while recoverable fields coexist with directional effects of at most .0171 and unchanged choices, separating information availability from interpretive influence. Fitted directions test a linear change associated with one field, while state replacement applies a context-specific authority-conditioned change; the two effects answer different questions about how relational information influences the competing readings.

5 Conclusion

The same utterance can perform different social actions as decision rights change. RitualMind makes this dependence testable by changing decision rights while keeping identities and utterances fixed. Authority-sensitive preferences extend across descriptions and recombined content in a subset of models, while several models preserve authority direction across wording and source scenes excluded from decoder training, including immediately before interpretation choice. Variation across models and descriptions makes the relationship between available information and its influence on a reading an empirical question. Controlled replacement in GLM-4-9B and confirmation at an attention component selected in advance connect authority-conditioned states to local changes in relative interpretation support. This influence extends to new scenarios under new descriptions; in new scenarios under original description B, behavioral and donor effects reverse and the control-adjusted interval spans zero. Recovery of all five contextual fields accompanies unchanged selected readings under fitted-direction shifts, further distinguishing information availability from its influence under the tested intervention. Together, RitualMind links pragmatic accounts of decision rights to experiments on how interpersonal relationships shape interpretation and influence model computation.

AI USE STATEMENT

In this work, we used GPT-5.6 to generate the study materials: the 80 base scenarios, the 300 source-inspired scenes, and the 45 recombined constructions. The authors reviewed and edited these materials. GPT-5.5 provided rubric-based scores for open responses; three human experts independently rated a stratified subset. Additionally, Claude Opus 5 was used for language editing. We have reviewed all AI-assisted work and take responsibility for the final content of this work.

ETHICS STATEMENT

Three adult experts with formal training in Chinese literature completed compensated online annotation. They were informed of the research and anonymous-release purpose and saw neither model identities nor automated scores. No sensitive personal information was collected. Appendix B describes the instructions, compensation, and retained annotations. The controlled scenarios specify experimental action contrasts; their assignments do not establish universal norms across Chinese or other cultural communities.

REPRODUCIBILITY STATEMENT

Section 3 defines the materials, interpretation margins, comparisons, and statistical estimates. Appendices A and B detail scoring and expert evaluation; Appendix H reports recombination, authority decoding during reading and interpretation choice, intervention controls, localization, and confirmation of the fixed attention component under new scenarios and descriptions. Materials and analysis code are available at `https://anonymous.4open.science/r/Ritualmind-2C04`. The supplement includes retained model outputs and anonymized expert scores for the controlled and open-response tasks, including prompts, responses, and annotations for all three scoring formulations used to recompute agreement and dialogue-condition differences. The accompanying supplement adds the confirmation materials, all four models' retained scores, both fixed-component interventions, and code for recomputing their estimates. It also provides the corrected cue definitions and explicit-guidance gain calculation. Private identity mappings and human free-text reasons are excluded.

Reported statistics can be recomputed from the retained outputs and scores. Obtaining new model responses or rubric scores constitutes a new evaluation. Open-weight inference requires separately obtained model weights.

REFERENCES

John L. Austin. *How to Do Things with Words*. Harvard University Press, Cambridge, Massachusetts, 1962. URL https://books.google.com/books?id=v3C_AQAACAAJ.

Alan D. Baddeley and Graham Hitch. Working memory. In Gordon H. Bower (ed.), *Psychology of Learning and Motivation*, volume 8, pp. 47–89. Academic Press, 1974. doi: 10.1016/S0079-7421(08)60452-1. URL https://doi.org/10.1016/S0079-7421(08)60452-1.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://doi.org/10.1162/coli_a_00422.

Shoshana Blum-Kulka, Juliane House, and Gabriele Kasper (eds.). *Cross-Cultural Pragmatics: Requests and Apologies*, volume 31 of *Advances in Discourse Processes*. Ablex Publishing Corporation, Norwood, New Jersey, 1989. ISBN 9780893915131. URL https://www.ets.org/research/policy_research_reports/publications/book/1989/hycs.html.

Penelope Brown and Stephen C. Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987. doi: 10.1017/CBO9780511813085. URL https://doi.org/10.1017/CBO9780511813085.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://iclr.cc/virtual/2023/poster/11480.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15959–15983, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.847. URL https://aclanthology.org/2024.acl-long.847/.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25663–25701, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1247. URL https://aclanthology.org/2025.acl-long.1247/.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 250–259, Sofia, Bulgaria, 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1025/.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 698–718, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.54. URL https://aclanthology.org/2021.emnlp-main.54/.

Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1–2):3–71, 1988. doi: 10.1016/0010-0277(88)90031-5. URL https://doi.org/10.1016/0010-0277(88)90031-5.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 653–670, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.48. URL https://aclanthology.org/2020.emnlp-main.48/.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 13518–13529. Curran Associates, Inc., 2023. doi: 10.52202/075280-0595. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2b9efb085d3829a2aadffab63ba206de-Abstract-Datasets_and_Benchmarks.html.

Erving Goffman. On face-work: An analysis of ritual elements in social interaction. *Psychiatry*, 18 (3):213–231, 1955. doi: 10.1080/00332747.1955.11023008. URL https://doi.org/10.1080/00332747.1955.11023008.

H. Paul Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics, Volume 3: Speech Acts*, pp. 41–58. Academic Press, 1975.

Yueguo Gu. Politeness phenomena in modern Chinese. *Journal of Pragmatics*, 14(2):237–257, 1990. doi: 10.1016/0378-2166(90)90082-O. URL https://doi.org/10.1016/0378-2166(90)90082-O.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0a6059857ae5c82ea9726ee9282a7145-Abstract-Conference.html.

Michael Haugh and Carl Hinze. A metalinguistic approach to deconstructing the concepts of 'face' and 'politeness' in Chinese, English and Japanese. *Journal of Pragmatics*, 35(10–11): 1581–1611, 2003. doi: 10.1016/S0378-2166(03)00049-3. URL https://doi.org/10.1016/S0378-2166(03)00049-3.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4194–4213, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.230. URL https://aclanthology.org/2023.acl-long.230/.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 62991–63010. Curran Associates, Inc., 2023. doi: 10.52202/075280-2749. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c6ec1844bec96d6d32ae95ae694e23d8-Abstract-Datasets_and_Benchmarks.html.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.890. URL https://aclanthology.org/2023.emnlp-main.890/.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024. doi: 10.1073/pnas.2405460121. URL https://www.pnas.org/doi/10.1073/pnas.2405460121.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882. PMLR, 2018. URL https://proceedings.mlr.press/v80/lake18a.html.

Stephen C. Levinson. *Pragmatics*. Cambridge University Press, 1983. doi: 10.1017/CBO9780511813313. URL https://doi.org/10.1017/CBO9780511813313.

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

Mengfan Li, Xuanhua Shi, and Yang Deng. CoSToM: Causal-oriented steering for intrinsic theory-of-mind alignment in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9302–9317, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.421. URL https://aclanthology.org/2026.acl-long.421/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

LuMing Robert Mao. Beyond politeness theory: 'Face' revisited and renewed. *Journal of Pragmatics*, 21(5):451–486, 1994. doi: 10.1016/0378-2166(94)90025-6. URL https://doi.org/10.1016/0378-2166(94)90025-6.

Hitesh Mehta, Arjit Saxena, Garima Chhikara, and Rohit Kumar. No universal courtesy: A cross-linguistic, multi-model study of politeness effects on LLMs using the PLUM corpus. *arXiv preprint arXiv:2604.16275*, 2026. doi: 10.48550/arXiv.2604.16275. URL https://arxiv.org/abs/2604.16275.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Earl K. Miller and Jonathan D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202, 2001. doi: 10.1146/annurev.neuro.24.1.167. URL https://doi.org/10.1146/annurev.neuro.24.1.167.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Randall C. O'Reilly and Michael J. Frank. Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2):283–328, 2006. doi: 10.1162/089976606775093909. URL https://doi.org/10.1162/089976606775093909.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs. In *International Conference on Learning Representations*, volume 2026, pp. 7657–7694, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0d1941f73833eb107939f62d323d6cc0-Abstract-Conference.html.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2373–2403, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.120. URL https://aclanthology.org/2025.naacl-long.120/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54/.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 20827–20905. Curran Associates, Inc., 2023. doi: 10.52202/075280-0913. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4241fec6e94221526b0a9b24828bb774-Abstract-Conference.html.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. 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), pp. 4463–4473, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

John R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969. doi: 10.1017/CBO9781139173438. URL https://doi.org/10.1017/CBO9781139173438.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever Hans or neural theory of mind? Stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2257–2273, St. Julian's, Malta, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.138. URL https://aclanthology.org/2024.eacl-long.138/.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Helen Spencer-Oatey. Face, (im)politeness and rapport. In Helen Spencer-Oatey (ed.), *Culturally Speaking: Culture, Communication and Politeness Theory*, pp. 11–47. Continuum, London and New York, 2 edition, 2008. ISBN 9780826493101. URL https://wrap.warwick.ac.uk/id/eprint/48276/.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12075–12097, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.719. URL https://aclanthology.org/2024.findings-acl.719/.

Melisa Stevanovic and Anssi Peräkylä. Deontic authority in interaction: The right to announce, propose, and decide. *Research on Language and Social Interaction*, 45(3):297–321, 2012. doi: 10.1080/08351813.2012.699260. URL https://doi.org/10.1080/08351813.2012.699260.

Melisa Stevanovic and Anssi Peräkylä. Three orders in the organization of human action: On the interface between knowledge, power, and emotion in interaction and social relations. *Language in Society*, 43(2):185–207, 2014. doi: 10.1017/S0047404514000037. URL https://doi.org/10.1017/S0047404514000037.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8:1285–1295, 2024. doi: 10.1038/s41562-024-01882-z. URL https://www.nature.com/articles/s41562-024-01882-z.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 All Tools, 2024. URL https://arxiv.org/abs/2406.12793.

Polina Tsvilodub, Jan-Felix Klumpp, Amir Mohammadpour, Jennifer Hu, and Michael Franke. On emergent social world models — evidence for functional integration of theory of mind and pragmatic reasoning in language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37382–37420, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1735. URL https://aclanthology.org/2026.acl-long.1735/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. doi: 10.48550/arXiv.2308.10248. URL https://arxiv.org/abs/2308.10248.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, Ji Zhang, Chao Peng, Fei Huang, and Jingren Zhou. CValues: Measuring the values of Chinese large language models from safety to responsibility. *arXiv preprint arXiv:2307.09705*, 2023. doi: 10.48550/arXiv.2307.09705. URL https://arxiv.org/abs/2307.09705.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. Do large language models understand conversational implicature- a case study with a Chinese sitcom. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pp. 1270–1285, Taiyuan, China, 2024. Chinese Information Processing Society of China. URL https://aclanthology.org/2024.ccl-1.98/.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62638–62681. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhu24o.html.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. doi: 10.48550/arXiv.2310.01405. URL https://arxiv.org/abs/2310.01405.

A MODELS, SCORING, AND INFERENCE

A.1 MODEL SCORING

Model measurements use inference without parameter updates. Table 28 identifies the four models and maps relative depths 0.25, 0.5, and 0.75 to transformer-block outputs. Candidate scores sum log probabilities over the required A–D answer-letter continuation at the prompted answer boundary, with each letter denoting one supplied interpretation. All models use this immediate-answer protocol. Scoring follows each model's tokenization; the GLM-4-9B records use single-token letters. Behavior and intervention are measured at the last token before the answer. The original decoding analysis instead instructs the model to form an internal interpretation without answering and reads the last non-padding prompt state. The choice-decoding analysis in Appendix H.3 reads the final pre-answer state under interpretation-choice instructions, retaining the candidate mapping. The prompts and scoring code are provided in the accompanying research materials. In the primary task, decision authority, publicness, and participant roles are stated in the prompt; face risk is held at medium in the task specification.

A.2 INDEPENDENT UNITS AND MULTIPLICITY

The broader controlled-choice task crosses three social dimensions: authority (hearer, shared, or speaker), publicness (private, semi-public, or public), and face risk (low, medium, or high). Role descriptions co-vary with authority, so this dimension is termed authority–role configuration. The independent unit is the base utterance ($n = 12$); principal estimates average two phrasings of the surrounding scene, A and B, with the utterance unchanged, before testing. For each of four actions, three ordinal trends, six adjacent contrasts, and four interaction summaries yield 52 comparisons within each model, with Benjamini–Hochberg correction. Positive and negative estimates describe increases and decreases in the selected action's frequency. Phrasing-specific and A-minus-B estimates form separate sensitivity analyses outside that correction family.

Behavioral authority margins and intervention effects use 95% intervals over the 40 constructed scenario groups, separately by phrasing. Recombination resamples 45 base constructions, and the temperature–pressure control resamples 80 base constructions; each uses 98.75% intervals for its four-model family. The 80 constructed scenarios draw on 45 source scenes and form 40 groups, with the two scenarios in each group supplying one another's same-authority controls. Behavioral and intervention estimates are averaged within these groups, and their intervals resample the 40 group means separately by phrasing. Of the 45 sources, 35 occur in two groups and 10 in one.

Original decoding separates scenario groups in four folds and uses 98.75% Bonferroni intervals for two directions and two lexical comparisons within model. The stricter analysis resamples source scenes, using 98.75% intervals for the four-model family. Reading and choice use the same source-scene resampling, with separate four-model families for choice AUC relative to .5 and choice-minus-reading AUC (Appendix H.3). These behavioral, decoding, and intervention analyses use 50,000 bootstrap draws. Module and head localization alternate selection and evaluation over two sets of 20 scenario groups, with 98.75% intervals for pooled effects and candidate contrasts and 95% intervals for each phrasing.

The fixed-component confirmation in Appendix H.6 uses 40 new scenario families and a six-comparison family for its primary local effects (99.17% intervals). Behavioral changes across four models and three conditions form a separate twelve-comparison family (99.58%). Both use 50,000 bootstrap draws with seed 20260914. Description-specific intervals are 95%. In the original scenarios under new descriptions, behavioral and donor changes resample 45 sources; control-dependent effects resample the five groups connected by shared source scenes. Appendix H.6 also gives the original 40-group sensitivity analysis.

The human dialogue analysis estimates differences between information conditions within 14 scenes. For each scene and condition, it takes the median available expert score and computes oracle-card minus active dialogue for task success, face preservation, role consistency, information transfer, and ritual–task balance. Its 95% intervals use 10,000 scene-bootstrap draws with seed 20260822, without adjustment across the five dimensions. Appendix B places these estimates alongside the corresponding automated scores and expert-aggregation analyses.

A.3 Estimation and validation

This broader controlled-choice design contains one observation for each base-utterance–state–phrasing combination within a model. The broader joint-field decoding analysis assigns each of 80 base items to one evaluation fold. AUC and balanced accuracy are computed within each fold before averaging across folds, with label-permuted training using the same assignments. The choice task estimates changes in selected actions; joint-field decoding assesses whether the combined social descriptions can be distinguished in internal states.

The controlled authority task contains 80 base items in 40 scenario groups. Each phrasing contributes 160 directed comparisons, or four comparisons per group, giving 320 across phrasings. Estimates first average within group and then resample groups. The original decoding analysis evaluates the 40 groups across four folds. Recomputing statistics from the retained item-level outputs reproduces the reported bootstrap intervals.

Authority decoding uses training-fold standardization followed by L2-regularized logistic regression (liblinear, $C = 1$). Character and word TF–IDF vocabularies and inverse-document-frequency weights are fitted only on training folds. Character features span 2–5 characters. The word comparator uses `(?u)\b\w+\b` with unigrams and bigrams: contiguous Unicode word-character strings form tokens, without segmenting Chinese into linguistic words. Both comparators use sublinear term frequency, a minimum document frequency of 1, and no lowercasing. Vocabulary limits are 100,000 character features and 50,000 word features.

The two fold assignments distinguish generalization across descriptions from generalization across source content. In the original assignment, 280 of 320 evaluated endpoints per phrasing share an utterance, literal meaning, and relation state with an instance in another fold, although the text differs. It therefore tests new phrasing with recurring source content. The five-fold assignment in Appendix H.2 keeps all instances from a source scene together, directly testing scenes absent from decoder training. Intervals resample evaluation units while retaining fitted decoders, and hence describe variability conditional on those fitted decoders.

For DS-Llama-8B, an initial tokenizer failed to encode the Chinese reading prompts, producing identical representations. We re-extracted the original 1,920 reading prompts with the model's tokenizer and repeated all affected decoding analyses, retaining the original stimuli, model parameters, depths, folds, and estimators. The reported results use this corrected extraction. Behavioral scores and interventions used a separate tokenizer and are unchanged.

Character TF–IDF distinguishes the two authority conditions with raw AUC 0 and orientation-free AUC 1. Because these scores indicate complete label reversal, the corrected source-scene comparison uses continuous-string TF–IDF for inference and retains character TF–IDF to diagnose label direction. The earlier analysis used character features as its sole primary comparator. The correction reused the model outputs and fitted decoders available at that stage, with the same folds, directions, layer, bootstrap seed, and multiplicity family. Both versions are retained so that the consequence of the comparator choice can be examined.

B Human Evaluation Details

Participants and sampling. The open-response evaluation examines how rubric judgments and the effects of additional relational information compare across scoring sources. Three adult experts with formal training in Chinese literature evaluated 504 outputs online, comprising 180 production, 96 face-tracking, 144 literary-transfer, and 84 dialogue outputs. Production was balanced across four models and nine domains; face tracking was balanced across models and stratified by domain; literary transfer was balanced across models and source texts. The dialogue subset contains 14 complete scenes with six versions each: one dialogue from each studied model, plus active and oracle-card multi-agent dialogues. A deterministic procedure fixed these task quotas and selected outputs before annotation; the released annotations retain their identifiers.

Experts received task information, compensation terms, and notice of the research and anonymous-release purposes. They saw anonymized outputs without model identities or GPT-5.5 scores. No sensitive personal information was requested or collected. Each expert received RMB 1,500 for the complete annotation package.

Rubrics and retained annotations. Ordinal rubric scores distinguish an unsatisfied criterion (0), partial satisfaction (1), and clear satisfaction (2). Literary-transfer indicators instead record specified errors as binary values. Every retained judgment includes a brief reason and confidence value; production, face tracking, and dialogue also include multi-label diagnostic tags. These scales assess the response properties defined in Table 33, allowing agreement to be examined at both exact-score and broader ordinal levels.

Before computing agreement, 27 production annotations marked unscorable or lacking a recoverable score were omitted. The retained sample comprises 1,485 output–expert records, each containing several rubric dimensions (Table 2). Output-level expert medians use all available ratings: five production outputs have one, 17 have two, and 158 have three. A stricter sensitivity subset excludes 178 records with incomplete diagnostic indicators, leaving 1,307 records. For production, face tracking, literary transfer, and dialogue, respectively, human–human exact agreement in this subset is .582, .540, .818, and .595, and within-one agreement is .924, .906, .818, and .950. For the binary task, the archived within-one field repeats exact agreement.

Automated rubric scores. GPT-5.5 produced scores using three prompt formulations. In the 9,884-output scoring archive, 9,004 outputs were assessed with two prior model judgments supplied as noisy references alongside the task and response. Another 880 face-tracking and literary-transfer outputs from DS-Qwen-14B and DS-Llama-8B were assessed from the task and response alone. Both formulations enter overall human–scorer agreement: 384 expert-evaluated outputs use the first and 120 the second. A separate dialogue-card formulation, without prior judgments, scores 420 outputs from 84 scenes under active, oracle, neutral, shuffled, and minimal conditions. These dialogue-card scores supply the condition comparisons below. All successful recorded calls use high reasoning effort and retain one score vector per output. Retries following failure, timeout, or an unparseable response are attempts to obtain that rating, not additional independent ratings.

The scoring archive covers earlier models as well as the four studied here. Table 2 therefore distinguishes its task totals from the fixed expert-evaluated subset. The scoring-prompt formulation accompanies each retained score, allowing comparisons to be interpreted with the information available to the scorer.

The code supplement provides the scoring prompts and retained responses; invocation records are also archived locally. Each of the 9,884 adjudication vectors and 420 dialogue-card vectors is traceable to a retained response. Temperature and top-p were not explicitly set. Recomputing the reported statistics uses those retained scores, while a fresh service call supplies a new evaluation (see the reproducibility statement). The two procedures received identical text for all 28 active/oracle outputs in the 14-scene subset, but differ on 29 of 140 rubric scores. They are kept separate throughout: overall agreement uses adjudication scores, and dialogue-condition effects use dialogue-card scores. On these 140 active/oracle rubric-dimension comparisons, dialogue-card scores agree exactly with expert medians on 42.1%, fall within one point on 96.4%, and have MAE .614.

Exact scores and ordinal proximity. Exact agreement measures whether two assessments assign the same score, within-one agreement measures their ordinal proximity, and mean absolute error (MAE) measures the size of their difference. For integer ratings on the 0–2 scale, within-one agreement excludes only 0-versus-2 differences; expert medians can also take half-point values. Table 3 reports all three measures. Human–human within-one agreement is .917–.925 across ordinal tasks, while exact agreement ranges from .466 in dialogue to .597 in face tracking. Human median–scorer within-one agreement ranges from .962 to .976; exact agreement on binary literary-transfer indicators is .834. These statistics describe different resolutions of the same rubric judgments.

Human–human rates pool all available two-expert combinations and dimensions: 2,455 comparisons for production, 1,728 for face tracking, 2,592 for literary transfer, and 1,260 for dialogue. Human–scorer rates compare each output's expert median with its adjudication score over 900, 576, 864, and 420 comparisons, respectively. These rates compare either two experts' individual scores or an output-level expert median with an automated score.

Within-scene dialogue differences. The dialogue comparison asks whether giving agents additional relational information changes task progress and interpersonal judgments. In active dialogue, each agent knows its own role, status, private goal, and the public history. The oracle card additionally

Table 2: Coverage of the open-response evaluation. Archive totals include earlier models beyond the four studied here. Experts evaluated a fixed task-quota subset; retained output–rater records exclude 27 unscorable or incomplete production records.

Task	Scoring archive	Expert-scored outputs	Retained records	Score range
Production	6,300	180	513	0–2
Face tracking	980	96	288	0–2
Literary transfer	2,100	144	432	0–1
Dialogue	504	84	252	0–2
Total	9,884	504	1,485	n/a

Table 3: Agreement on the 504 expert-scored outputs. Each panel reports pooled rubric-dimension comparisons, with distinct aggregation rules. Literary transfer is binary, so within-one is not reported as a separate statistic.

Task	Comparisons	Exact	Within-1	MAE
Human–human: all available two-expert comparisons				
Production	2,455	.558	.925	.516
Face tracking	1,728	.597	.921	.481
Literary transfer	2,592	.818	n/a	.182
Dialogue	1,260	.466	.917	.617
Human median–adjudication scorer				
Production	900	.573	.962	.448
Face tracking	576	.653	.976	.372
Literary transfer	864	.834	n/a	.166
Dialogue	420	.498	.962	.540

supplies all agents' roles, status, and private goals, together with an interaction rule. Its contrast with active dialogue estimates the effect of this information bundle.

The oracle card's task-success difference over active dialogue is $+.429$ $[.071, .786]$ under expert medians across the 14 expert-evaluated scenes and $-.143$ $[-.500, .286]$ under the automated scorer; the direct human-minus-scorer contrast is $+.571$ $[0, 1.071]$ (Table 4). Human face-preservation and information-transfer means are zero, while role consistency and ritual–task balance have positive means; their intervals contain or touch zero. The automated face-preservation difference is positive, but its direct contrast with human scores includes zero. Added information is therefore evaluated separately for each response dimension and scoring source, with a positive task-success estimate under expert-median aggregation. All five dimensions use the scene-bootstrap procedure in Appendix A.2. The separate 84-scene automated evaluation appears in Table 38; the 84 expert-evaluated dialogue outputs here represent 14 scenes and six versions.

Table 4: Oracle-card minus active score differences across 14 scenes. Human scores are the median within scene and condition; scorer scores use the dialogue-card procedure. The final column subtracts the scorer difference from the human difference within scene. Intervals use 10,000 scene-bootstrap draws (95%, unadjusted across dimensions).

Metric	Human Δ [CI]	Scorer Δ [CI]	Human–scorer Δ [CI]
Task success	$+.429$ $[.071, .786]$	$-.143$ $[-.500, .286]$	$+.571$ $[0, 1.071]$
Face preservation	$.000$ $[-.214, .214]$	$+.286$ $[.071, .500]$	$-.286$ $[-.643, .071]$
Role consistency	$+.143$ $[0, .357]$	$.000$ $[-.357, .357]$	$+.143$ $[-.143, .429]$
Information transfer	$.000$ $[-.214, .214]$	$-.071$ $[-.429, .357]$	$+.071$ $[-.286, .429]$
Ritual–task balance	$+.071$ $[-.143, .286]$	$.000$ $[-.357, .357]$	$+.071$ $[-.286, .429]$

Sensitivity of the dialogue comparison. The aggregation analyses show how the task-success estimate reflects the observed scene scores. Expert-median differences comprise six positive scenes, one negative scene, and seven ties. The two-sided exact sign test gives $p = .125$, and the exact sign-flip test of the human-minus-scorer difference gives $p = .09765625$. These address the direction

of scene differences and the score-source contrast under sign symmetry, respectively. Using expert means gives a task-success estimate of .262 $[0, .548]$; restricting expert medians to confidence $\geq .8$ gives .286 $[-.143, .714]$. The high-confidence role-consistency estimate is .286 $[.071, .536]$. Table 5 reports every dimension under both alternatives, using the same 14 scenes and unadjusted 95% intervals for each response dimension.

Table 5: Sensitivity of oracle-card minus active differences to expert aggregation and confidence filtering. Values are mean within-scene differences with 95% scene-bootstrap intervals.

Dimension	Expert mean	High-confidence median
Task success	.262 $[0, .548]$	.286 $[-.143, .714]$
Face preservation	.095 $[-.048, .238]$	.036 $[-.179, .250]$
Role consistency	.095 $[-.048, .238]$	.286 $[.071, .536]$
Information transfer	.143 $[-.024, .333]$	.250 $[-.036, .500]$
Ritual–task balance	.143 $[-.048, .357]$	.214 $[-.071, .464]$

Omitting each expert in turn yields task-success estimates of .214 $[-.036, .464]$, .286 $[-.071, .714]$, and .286 $[.036, .536]$. The mean direction remains positive under all three omissions; two intervals include zero. Expert composition thus affects the size and interval width of the task-success estimate but not its direction (Table 4).

C Contextual and Cue-Specific Behavioral Sensitivity

A coherent account of the interaction improves target-aligned interpretation over both the utterance alone and context reassigned from another item in every model. The utterance-only condition removes contextual information; the Scrambled condition supplies another item's intact context in the same field format. Both available context and its correspondence with the utterance thus improve target-aligned interpretation. Table 6 summarizes context gains, cue conflicts, and explicit guidance. Single-cue reported reversals combine different choices with an affirmative meaning-change judgment (Appendix I.6). The cue-conflict and explicit-guidance tasks score the interpretation categories and cue priorities specified in their materials.

Table 6: Target-aligned choice scores in supplementary context, cue-conflict, and explicit-format tasks. Gains use the stated comparison condition.

	Qwen3-4B	DS-Qwen-14B	DS-Llama-8B	GLM-4-9B
Coherent target alignment	.445	.432	.445	.471
Coherent minus utterance	+.071	+.073	+.178	+.125
Coherent minus scrambled	+.077	+.060	+.114	+.148
Cue-conflict target alignment	.397	.427	.544	.412
Audience-minus-status choice rate	+.425	+.389	+.300	+.328
Best explicit-format gain	+.175	+.175	+.113	+.263

Topic-only context tests how much a broad subject contributes before the full relational setting is supplied (Figure 5). The other conditions provide the utterance alone, another item's intact context (Scrambled), or a complete coherent context. The four plotted point estimates distinguish these information conditions; Figure 3 gives intervals for the coherent-minus-utterance and coherent-minus-scrambled comparisons.

The broader decoding targets concern combined social descriptions (Figure 4). The authority–role label varies decision authority together with speaker and hearer role descriptions; audience–face combines publicness and face risk; the Role label indicates whether the ordering of participant roles agrees with decision authority. These targets assess distinguishability of the joint configurations. The authority analyses in Appendices E–G isolate the allocation of decision rights with role titles fixed, supplying the more specific comparison underlying the main claim.

D Action Distributions under Crossed Relational Conditions

Multiple social conditions can shape both which action category is favored and whether that preference changes. The three-axis task crosses three levels each of authority–role configuration, publicness, and face risk; authority conditions also change associated role descriptions. Inference uses 12 base utterances after averaging phrasings. Table 7 separates mean action proportions from the counts of trends, adjacent contrasts, and interactions passing Benjamini–Hochberg correction across 52 comparisons within each model.

Table 7: Mean action proportions across the three-axis design and counts passing Benjamini–Hochberg correction among 52 comparisons. Inference averages both phrasings within each of the 12 base utterances.

Model	Literal	Deferential	Status	Face-saving	BH/52
Qwen3-4B	.241	.000	.759	.000	10
DS-Qwen-14B	.102	.077	.154	.667	13
DS-Llama-8B	1.000	.000	.000	.000	0
GLM-4-9B	.012	.278	.127	.583	16

The most frequent interpretation and the response to changing conditions differ across models. Qwen3-4B predominantly chooses pressure, while DS-Qwen-14B and GLM-4-9B predominantly choose face-saving warnings; all three also show corrected condition effects. DS-Llama-8B chooses literal observation throughout and has no effects passing its 52-comparison family. Across the four models, 39 of 208 comparisons pass correction, comprising 23 positive and 16 negative estimates; 169 do not pass. These directions describe changes in category frequency. Separate A, B, and A-minus-B sensitivity estimates contain 33 comparisons with opposite A/B point-estimate signs among the 208 comparisons, without treating sign reversal itself as a tested phrasing difference.

E Behavioral Sensitivity to Controlled Authority Reversals

The controlled authority comparison identifies how current decision rights affect interpretation while retaining the same utterance, identities, role titles, audience, stakes, and answer mapping (Table 1). The resulting log-margin change G measures the increase in relative support for the interpretation associated with the reversed authority state. Table 8 provides the estimates used in Figure 2.

Table 8: Decision-authority-induced action-margin change. Values average 40 constructed scenario groups per model, with 95% bootstrap intervals. Magnitudes are interpreted on each model's own log-probability scale.

Model	Phrasing A	Phrasing B
Qwen3-4B	.003 $[-.013, .018]$	.023 $[.001, .046]$
DS-Qwen-14B	.869 $[.740, 1.004]$	.377 $[.231, .526]$
DS-Llama-8B	.125 $[.081, .169]$	.053 $[-.022, .125]$
GLM-4-9B	2.242 $[1.970, 2.515]$	1.128 $[.790, 1.468]$

DS-Qwen-14B and GLM-4-9B respond in the authority-directed sense under both descriptions. Their A/B means are $.8691/.3765$ and $2.2424/1.1284$, respectively, with positive intervals throughout. Qwen3-4B has a positive interval in B and DS-Llama-8B in A; their other intervals include zero. The two-phrasing evidence thus establishes that the current allocation of rights can change interpretation in DS-Qwen-14B and GLM-4-9B across alternative descriptions of the same participants. Magnitudes are interpreted within each model's log-probability scale.

F Authority Information across Phrasings and Label Direction

A consistent authority distinction should retain which participant has the right to decide when the description changes. Cross-form decoding tests this property by training on one phrasing and

evaluating on the other. Raw AUC measures discrimination in the original label direction; orientation-free AUC measures separation allowing the labels to reverse. Table 9 compares hidden-state and lexical AUC for both directions at the primary depth 0.5.

Table 9: Cross-form decoding at 0.5 relative depth. Intervals are 98.75% familywise bootstrap intervals over scenario groups for the raw-AUC difference from each lexical baseline. OF denotes orientation-free hidden-state AUC.

Model	Transfer	Hidden	Hidden OF	Δ char. [CI]	Δ word [CI]
Qwen3-4B	A-to-B	.321	.679	.321 [.263, .369]	−.179 [−.238, −.131]
Qwen3-4B	B-to-A	.165	.835	.165 [.107, .211]	−.335 [−.395, −.288]
DS-Llama-8B	A-to-B	.834	.834	.834 [.798, .896]	.334 [.298, .397]
DS-Llama-8B	B-to-A	.872	.872	.872 [.832, .917]	.372 [.333, .417]
GLM-4-9B	A-to-B	.680	.680	.680 [.602, .757]	.180 [.102, .259]
GLM-4-9B	B-to-A	.682	.682	.682 [.620, .758]	.182 [.120, .259]
DS-Qwen-14B	A-to-B	.761	.761	.761 [.732, .821]	.261 [.233, .321]
DS-Qwen-14B	B-to-A	.995	.995	.995 [.985, 1.000]	.495 [.485, .500]

The character comparator completely separates authority conditions while reversing their labels: raw AUC is 0 and orientation-free AUC is 1 in every cell. Continuous-string TF–IDF has AUC .5. GLM-4-9B, DS-Qwen-14B, and DS-Llama-8B exceed both raw lexical comparators and preserve the authority direction in both evaluations; Qwen3-4B separates the conditions with reversed direction. The preserved assignment in the former three models is the additional property captured by directional decoding. Character-level separation remains visible in the comparison, so the hidden-state result measures consistent recovery of the authority direction across phrasings.

Point estimates at depths 0.25 and 0.75 describe where the authority-label direction is recovered (Table 10). DS-Qwen-14B and GLM-4-9B exceed the word comparator in both directions at both depths. Qwen3-4B has positive early-depth and negative later-depth differences; DS-Llama-8B has opposite signs across early-depth directions and positive later-depth differences. Replacing the states at those locations tests their influence on interpretation in Appendix G.

Table 10: Hidden-state minus continuous-string TF–IDF AUC at relative depths 0.25 and 0.75. Each cell gives A-to-B/B-to-A point estimates; the primary comparison uses the prespecified 0.5 depth.

Model	0.25 depth	0.75 depth
Qwen3-4B	.172/.216	−.150/ − .155
DS-Qwen-14B	.193/.426	.271/.433
DS-Llama-8B	.050/ − .153	.371/.358
GLM-4-9B	.103/.217	.171/.175

G Intervention Effects across Model Depths

These interventions test whether changing authority-conditioned states influences the interpretation margin, complementing the preceding analysis of recoverable information. Replacement is evaluated at three relative depths against a same-authority change of equal magnitude (Table 11). Each estimate first averages within 40 scenario groups, then resamples groups. The prespecified middle depth is marked in bold irrespective of interval direction; additional early and late locations use unadjusted 95% intervals.

Table 11: Intervention-minus-control action-margin change by relative depth and phrasing. The primary 0.50-depth rows are bold; other 95% intervals are unadjusted estimates at additional depths.

Model	Depth	A-to-A E [95% CI]	B-to-B E [95% CI]
Qwen3-4B	0.25	.0030 [−.0010, .0072]	−.0072 [−.0139, −.0008]
Qwen3-4B	**0.50**	**.0017** [**−.0042**, **.0074**]	**−.0015** [**−.0109**, **.0077**]
Qwen3-4B	0.75	.0037 [−.0097, .0168]	.0084 [−.0054, .0220]

Continued on next page

Table 11 (continued)

Model	Depth	A-to-A E [95% CI]	B-to-B E [95% CI]
DS-Qwen-14B	0.25	.0133 [.0054, .0210]	.0081 [.0010, .0152]
DS-Qwen-14B	**0.50**	**−.0105 [−.0229, .0016]**	**.0130 [−.0091, .0342]**
DS-Qwen-14B	0.75	.2007 [.1157, .2849]	.1407 [.0752, .2103]
DS-Llama-8B	0.25	.0022 [−.0226, .0274]	.0047 [−.0230, .0317]
DS-Llama-8B	**0.50**	**.0213 [−.0197, .0612]**	**.0307 [−.0152, .0765]**
DS-Llama-8B	0.75	.0860 [.0388, .1341]	.0769 [.0053, .1460]
GLM-4-9B	0.25	−.0032 [−.0059, −.0006]	−.0020 [−.0045, .0006]
GLM-4-9B	**0.50**	**.0138 [.0071, .0205]**	**.0150 [.0062, .0239]**
GLM-4-9B	0.75	1.8868 [1.5883, 2.1820]	.8937 [.5876, 1.2047]

Across the 24 depth–phrasing contrasts, ten intervals are positive, two negative, and twelve include zero. GLM-4-9B has positive intervals under both phrasings at the primary middle depth. DS-Qwen-14B has positive intervals at both outer depths and DS-Llama-8B at the later depth; GLM-4-9B also has positive late-depth intervals. The negative intervals occur for Qwen3-4B at the early depth in B and GLM-4-9B at the early depth in A. For DS-Qwen-14B, replacement influences the interpretation margin at both outer depths.

The two components of the intervention contrast distinguish an increase caused by authority replacement from its advantage over the control. The donor change is $D = m(h+\delta) - m(h)$, and the same-authority control change is $C = m(h+\delta^{\text{ctrl}}) - m(h)$. Table 12 reports both against the unmodified starting margin. The decomposition retains $E = D - C$ for every record and reproduces the aggregated effects. Intervals for D and C use the same resampled groups within each comparison, with seeds beginning at 20260908 and varying by model, depth, and phrasing.

Table 12: Donor and same-authority control changes in the block-output intervention. Each change is relative to the unmodified recipient margin. The designated depth is 0.5; all depths and phrasings are retained. Intervals use 50,000 joint resamples of the 40 scenario groups; the original E intervals remain in Table 11. Differences are computed before rounding.

Model	Depth	Form	Donor D [95% CI]	Same-authority C [95% CI]	$E = D - C$
Qwen3-4B	0.25	A	.0025 [−.0015, .0064]	−.0006 [−.0045, .0033]	.0030
Qwen3-4B	0.25	B	.0010 [−.0047, .0066]	.0082 [.0018, .0147]	−.0072
Qwen3-4B	0.5	A	.0028 [−.0017, .0074]	.0011 [−.0042, .0063]	.0017
Qwen3-4B	0.5	B	.0043 [−.0039, .0126]	.0058 [.0001, .0119]	−.0015
Qwen3-4B	0.75	A	−.0017 [−.0092, .0057]	−.0054 [−.0172, .0062]	.0037
Qwen3-4B	0.75	B	.0091 [−.0016, .0200]	.0007 [−.0095, .0107]	.0084
DS-Qwen-14B	0.25	A	.0094 [.0039, .0147]	−.0039 [−.0083, .0004]	.0133
DS-Qwen-14B	0.25	B	.0046 [−.0021, .0113]	−.0035 [−.0093, .0019]	.0081
DS-Qwen-14B	0.5	A	−.0042 [−.0123, .0040]	.0063 [−.0041, .0169]	−.0105
DS-Qwen-14B	0.5	B	.0143 [−.0035, .0322]	.0013 [−.0161, .0180]	.0130
DS-Qwen-14B	0.75	A	.2275 [.1527, .3073]	.0268 [−.0260, .0780]	.2007
DS-Qwen-14B	0.75	B	.1378 [.0780, .2008]	−.0029 [−.0459, .0415]	.1407
DS-Llama-8B	0.25	A	.0283 [.0054, .0510]	.0261 [.0013, .0506]	.0022
DS-Llama-8B	0.25	B	.0061 [−.0145, .0275]	.0014 [−.0215, .0244]	.0047
DS-Llama-8B	0.5	A	.0564 [.0172, .0959]	.0351 [−.0007, .0709]	.0213
DS-Llama-8B	0.5	B	.0113 [−.0240, .0455]	−.0194 [−.0483, .0091]	.0307
DS-Llama-8B	0.75	A	.1144 [.0743, .1537]	.0284 [−.0247, .0821]	.0860
DS-Llama-8B	0.75	B	.0278 [−.0176, .0722]	−.0490 [−.0995, .0022]	.0769
GLM-4-9B	0.25	A	−.0009 [−.0030, .0013]	.0023 [−.0000, .0046]	−.0032
GLM-4-9B	0.25	B	−.0012 [−.0035, .0010]	.0007 [−.0017, .0033]	−.0020
GLM-4-9B	0.5	A	.0198 [.0137, .0255]	.0060 [.0013, .0104]	.0138
GLM-4-9B	0.5	B	.0158 [.0097, .0224]	.0008 [−.0049, .0066]	.0150
GLM-4-9B	0.75	A	1.9399 [1.7124, 2.1647]	.0532 [−.1573, .2660]	1.8868
GLM-4-9B	0.75	B	.9477 [.6602, 1.2325]	.0541 [−.1065, .2105]	.8937

At GLM-4-9B's middle depth, authority replacement itself increases support for the opposite-authority interpretation in both phrasings: $D = .0198$ $[.0137, .0255]$ and .0158 $[.0097, .0224]$. Both increases exceed their corresponding control effects. This decomposition establishes the direction of the replacement response underlying the positive E contrasts. DS-Llama-8B at depth 0.25 in A illustrates the complementary case: its donor and control changes are both positive, .0283 $[.0054, .0510]$ and .0261 $[.0013, .0506]$, while the E interval includes zero. Reading D together with E therefore distinguishes an absolute shift in interpretation from a shift beyond the authority-preserving control.

H Robustness and Component Localization

The authority-sensitive interpretation result motivates testing its dependence on the original material and identifying internal changes that affect it. Recombination changes the original content associations; source-scene separation tests authority recovery beyond decoder training material, both during reading and interpretation choice. A recognized temperature–pressure state provides a control for the replacement procedure, and localization narrows its effect to particular internal outputs. Throughout, the interpretation margin measures relative support, while the highest-scoring option records the selected answer. Together these measures describe both preference strength and choice.

H.1 Content–relation recombination

Content recombination tests dependence on the original association between an utterance and its social setting. A fixed one-to-one reassignment places each utterance in a setting from another source-inspired domain, yielding 45 base constructions. Options and answer letters remain identical within each authority reversal. The accompanying research materials provide this mapping. Table 13 averages both phrasings and authority directions within each construction before obtaining 98.75% intervals for the four-model family.

Table 13: Authority-directed action-margin effects under content–relation recombination. Difference denotes recombined minus original; all values are within-model log-margin units.

Model	Recombined effect [98.75% CI]	Original effect [98.75% CI]	Difference [98.75% CI]
Qwen3-4B	.0217 [.0011, .0416]	.0096 [−.0091, .0280]	.0122 [−.0131, .0360]
DS-Qwen-14B	.6780 [.4842, .8737]	.5973 [.4074, .7973]	.0807 [−.2005, .3763]
DS-Llama-8B	.0786 [−.0041, .1613]	.0749 [−.0134, .1637]	.0038 [−.0681, .0716]
GLM-4-9B	1.6262 [1.2150, 2.0400]	1.6641 [1.2493, 2.0758]	−.0379 [−.4928, .4372]

Qwen3-4B, DS-Qwen-14B, and GLM-4-9B have positive authority-directed intervals under recombined content; DS-Llama-8B's interval includes zero. All four recombined-minus-original intervals include zero. Reversing authority changes the highest-ranked option in fractions .000, .289, .006, and .378 in table order. The positive margins demonstrate sensitivity in three models beyond the original content combinations. Qwen3-4B's effect concerns relative support within an unchanged option ordering, while DS-Qwen-14B and GLM-4-9B also change selected answers in .289 and .378 of these comparisons.

H.2 Authority recovery in new source scenes

Keeping each source scene entirely within one fold tests the authority distinction on content absent from decoder training. The 45 sources supply 320 endpoints per phrasing, or 640 states. In each direction, each of five folds uses 256 states from 36 sources for training and 64 states from nine sources for evaluation. The statistic averages A-to-B and B-to-A AUC at relative depth 0.5, subtracting word TF–IDF AUC obtained with the same folds. Table 14 uses 50,000 bootstrap draws over evaluation source scenes and 98.75% intervals for the four-model family, conditional on the fitted decoders.

Table 14: Cross-form association with source scenes held out. Δ is hidden-state minus word TF–IDF ROC–AUC, averaged over both transfer directions.

Model	Hidden AUC	Word AUC	Δ [98.75% CI]	Direction
Qwen3-4B	.2687	.5000	−.2313 [−.3036, −.1968]	Reversed
DS-Qwen-14B	.8802	.5000	.3802 [.3612, .4295]	Consistent
DS-Llama-8B	.8565	.5000	.3565 [.3252, .4180]	Consistent
GLM-4-9B	.6950	.5000	.1950 [.1263, .2727]	Consistent

DS-Qwen-14B, DS-Llama-8B, and GLM-4-9B retain the authority direction on source scenes excluded from decoder training, while Qwen3-4B reverses it. Continuous-string TF–IDF is at chance, using the tokenization in Appendix A.3. Character TF–IDF completely separates the conditions with reversed labels (raw AUC 0, orientation-free AUC 1). This distinction between separation and

consistent label direction gives the hidden-state results their interpretation. Appendix A.3 documents the comparator correction.

H.3 AUTHORITY RECOVERY DURING INTERPRETATION CHOICE

The behavioral response is an interpretation choice, motivating a direct examination of authority information immediately before that choice. Each model contributes 640 states per task setting, with 320 endpoints in each phrasing. Choice states are extracted at relative depth 0.5 from the final pre-answer token under the original choice instructions and answer mapping. Reading uses the instruction to read without answering. Both conditions use the same model parameters and float32 arithmetic. Table 15 uses reading states obtained for this comparison; the original reading analyses elsewhere retain their original numerical settings.

Each task setting has separately fitted training-fold standardization and logistic regression for each phrasing direction. The five source-scene folds are shared across settings: each trains on 256 states from 36 sources and evaluates 64 states from nine sources in the other phrasing. AUC is averaged over folds and directions. The 50,000 bootstrap draws resample source scenes within evaluation folds, with shared source counts across models, directions, and settings. Intervals are conditional on fitted decoders. Choice AUC relative to .5 and choice-minus-reading AUC use separate four-model families with 98.75% intervals.

Table 15: Authority decoding under reading and interpretation-choice instructions, with source scenes held out. Values are raw AUCs or their differences, with 98.75% intervals. Differences are computed before rounding. Both task settings use float32 arithmetic.

Model	Reading AUC [CI]	Choice AUC [CI]	Choice − reading [CI]
Qwen3-4B	.2687 [.1964, .3032]	.4526 [.4033, .4968]	.1840 [.1540, .2515]
DS-Qwen-14B	.8786 [.8586, .9285]	.6990 [.6666, .7496]	−.1796 [−.2319, −.1398]
DS-Llama-8B	.8546 [.8241, .9157]	.6296 [.6024, .6754]	−.2250 [−.2832, −.1827]
GLM-4-9B	.6919 [.6223, .7699]	.8354 [.8087, .8870]	.1435 [.0603, .2430]

DS-Qwen-14B, DS-Llama-8B, and GLM-4-9B retain the authority-label direction during interpretation choice as well as reading. Choice AUC decreases in the two distilled models and increases in GLM-4-9B. Qwen3-4B also increases, while remaining below .5 on average; all four choice-minus-reading intervals exclude zero. The behavioral choice setting therefore retains direction-consistent authority information in the same three models, with a measurable change in its recoverability relative to reading alone.

Table 16: Both phrasing-transfer directions under reading and interpretation-choice instructions. Values are raw AUCs with 98.75% intervals for each direction.

Model	Task setting	A → B AUC [CI]	B → A AUC [CI]
Qwen3-4B	Reading	.3414 [.2338, .4151]	.1959 [.1023, .2590]
Qwen3-4B	Choice	.5461 [.4835, .6108]	.3592 [.3145, .3969]
DS-Qwen-14B	Reading	.7635 [.7226, .8625]	.9938 [.9845, 1.0000]
DS-Qwen-14B	Choice	.7416 [.6868, .8156]	.6564 [.6273, .7067]
DS-Llama-8B	Reading	.8371 [.7915, .9261]	.8721 [.8341, .9316]
DS-Llama-8B	Choice	.6250 [.5864, .6903]	.6342 [.5894, .6972]
GLM-4-9B	Reading	.7119 [.5687, .8324]	.6719 [.5980, .7753]
GLM-4-9B	Choice	.8453 [.8040, .9145]	.8254 [.7815, .8952]

Both choice directions retain authority-label ordering in DS-Qwen-14B, DS-Llama-8B, and GLM-4-9B. Qwen3-4B has a choice A-to-B interval including .5 and a B-to-A interval entirely below .5 (Table 16). These per-direction intervals describe the asymmetry behind Qwen3-4B's average.

The two task settings differ in instructions, candidate presentation, and conversation formatting. Classifiers are fitted within each setting, so the comparison measures recoverability under each setting's complete conditions. Its outcome is raw hidden-state AUC; Table 14 additionally compares reading states with lexical features. The corresponding influence on interpretation preferences is examined through intervention in Appendices G and H.5.

H.4 Deterministic non-relational intervention control

The temperature–pressure task asks whether replacement changes the prescribed response to a state the model recognizes. It assigns a deterministic action to each state and supplies 80 base constructions with 1,280 intervention observations per model at depth 0.5; resampling uses the 80 constructions. Recognition requires baseline recipient-state accuracy of at least .90. The intervention criteria require positive lower bounds for both the replacement effect and its advantage over a random direction of equal magnitude, using 98.75% intervals.

Table 17: Recognition of the prescribed temperature–pressure state and donor-replacement effects relative to a random direction of equal magnitude. Effects use model-specific interpretation log-margin units; the text defines the recognition and two intervention criteria.

Model	Recognition	Donor effect [98.75% CI]	Donor minus random [CI]
Qwen3-4B	.250	$-.0026$ $[-.0058, .0005]$	$-.0028$ $[-.0064, .0009]$
DS-Qwen-14B	.936	.0338 $[.0303, .0374]$	.0261 $[.0193, .0328]$
DS-Llama-8B	.309	.1805 $[.1700, .1905]$	.1793 $[.1659, .1929]$
GLM-4-9B	1.000	.0970 $[.0933, .1011]$	.0908 $[.0826, .0990]$

DS-Qwen-14B and GLM-4-9B satisfy both recognition and intervention criteria (Table 17). Their starting-state accuracies are .936 and 1.000, and replacement shifts relative support toward the action prescribed by the opposite state beyond the random-direction control. Qwen3-4B and DS-Llama-8B have accuracies .250 and .309; DS-Llama-8B nevertheless has positive replacement and replacement-minus-random effects. Jointly assessing recognition and response change identifies the conditions in which the prescribed action provides an interpretable control for replacement.

H.5 Localization across scenario groups

Localization tests whether a more specific part of the computation produces the intervention contrast. In GLM-4-9B and DS-Qwen-14B, two sets of 20 scenario groups alternate candidate selection and evaluation. Candidates are attention and feed-forward outputs at relative depths 0.5, 0.75, and 1.0. Selection ranks candidates with positive mean effects in both phrasings by their average effect. Evaluation requires positive lower bounds for the selected effect and its contrast with the selection-set runner-up in 98.75% intervals, plus positive phrasing-specific effects in 95% intervals. The same location must be selected under both assignments. Table 18 reports these comparisons.

Table 18: Module selection and evaluation in both directions. Candidate contrasts compare the selected output with the selection-group runner-up on held-out scenario groups.

Model	Selection→evaluation	Selected candidate	Runner-up	Effect [98.75% CI]	Candidate contrast [CI]
GLM-4-9B	0→1	Attention, .75	Attention, 1.0	.2823 $[.1539, .3992]$	.2564 $[.1371, .3686]$
GLM-4-9B	1→0	Attention, .75	Attention, 1.0	.3657 $[.2448, .5163]$	.3279 $[.2086, .4733]$
DS-Qwen-14B	0→1	Feed-forward, 1.0	Attention, .75	.0251 $[-.0187, .0694]$	$-.0124$ $[-.0755, .0448]$
DS-Qwen-14B	1→0	Feed-forward, .50	Attention, .75	.0201 $[.0057, .0340]$	$-.0271$ $[-.0599, .0077]$

GLM-4-9B selects its attention output at relative depth 0.75 under both assignments, satisfying the pooled, candidate-contrast, and phrasing-specific criteria. DS-Qwen-14B selects feed-forward outputs at different depths, with candidate-contrast intervals including zero. GLM-4-9B's repeated location permits a finer analysis of the components combined in that attention output.

The head analysis uses the same 40 scenario groups and selection criteria for 32 candidates, resolving ties by lower head index. At layer 30, each candidate contributes a 128-dimensional slice of the concatenated attention output before projection. Both selection sets identify one-based head 20 (zero-based index 19), with one-based head 22 as runner-up (Table 19). Replacing this slice changes one attention component before its contribution is combined by the output projection. The evaluation sets separate scenario groups but share 18 of 45 source scenes; head intervals are conditional on the preceding module selection and these materials.

The localized effect chiefly changes relative support for competing interpretations. Head-20 replacement changes the highest-ranked option in five of 320 evaluated observations, with no conversion

Table 19: GLM-4-9B layer-30 attention-head output effects in both evaluation directions, conditional on the selected module. Both directions select head 20 and compare its effect with head 22; the head labels in this table are one-based.

Selection→evaluation	Head-20 effect [98.75% CI]	Head-20 minus head-22 [98.75% CI]
0→1	.2171 [.1167, .3207]	.1885 [.0996, .2794]
1→0	.2424 [.1261, .3776]	.2100 [.1004, .3366]

from a non-target choice to the donor target. Across all 32 candidate heads, 17 of 10,240 observations change the highest-ranked option, again with zero donor-target conversions. These counts describe repeated interventions; the resampling units remain the 40 scenario groups. Both head-20 evaluations show a positive interpretation-margin contrast beyond the same-authority change of equal magnitude.

The localized contrast also includes a positive response to the authority replacement itself. Head-20 donor changes are .1729 [.0831, .2727] and .2413 [.1445, .3667] in the two evaluations, with same-authority control means of −.0442 and −.0011 (Table 20). The positive E values thus accompany increased support for the opposite-authority interpretation under replacement. These component intervals use the original evaluation groups and selections, with resampling seeds 20261908–20262009.

Table 20: Donor and same-authority changes for the selected outputs, evaluated in the original held-out groups. Intervals use 50,000 joint resamples of 20 evaluation groups, conditional on the selected output. The same underlying records give the effects in Tables 18 and 19.

Model / output	Selection→evaluation	Donor D [98.75% CI]	Same-authority C [98.75% CI]	E
GLM-4-9B attention 0.75	0→1	.2756 [.1730, .3799]	−.0067 [−.0657, .0592]	.2823
GLM-4-9B attention 0.75	1→0	.3481 [.2472, .4759]	−.0175 [−.0839, .0417]	.3657
DS-Qwen-14B feed-forward 1.0	0→1	.0201 [−.0217, .0645]	−.0050 [−.0296, .0196]	.0251
DS-Qwen-14B feed-forward 0.5	1→0	.0158 [.0060, .0261]	−.0043 [−.0160, .0072]	.0201
GLM-4-9B head 20 0.75	0→1	.1729 [.0831, .2727]	−.0442 [−.1235, .0259]	.2171
GLM-4-9B head 20 0.75	1→0	.2413 [.1445, .3667]	−.0011 [−.0574, .0526]	.2424

H.6 A LOCAL CONTRIBUTION ACROSS SCENARIOS AND DESCRIPTIONS

The confirmation fixes GLM-4-9B's layer-30 attention heads 20 and 22 before collecting new observations. The three conditions combine original scenarios with new descriptions, newly constructed scenarios with original descriptions, and new scenarios with new descriptions. Each contains 320 directed observations across 80 scenarios, two authority directions, and two descriptions. Original descriptions are the phrasings A and B; the new descriptions are C and D. Across the new-scenario materials, description B always introduces the speaker before the hearer, so "the former" refers to the speaker. The new scenarios form 40 families of two distinct situations, each providing the other's same-authority control, within the same nine broad domains. Every authority reversal preserves the people, utterance, literal gloss, and candidate mapping.

We replace each 128-dimensional head contribution at the final pre-answer token, before output projection, using the original scoring and intervention procedure. Opposite-authority replacement changes the interpretation margin by D; an equally large same-authority control change gives C, so $E = D - C$. We separately count any change in the highest-scoring candidate and a change from a non-target candidate to the donor-associated interpretation. The six preselected comparisons comprise head-20 D, head-20 E, and head-20 minus head-22 E in each of the two new-scenario conditions. They use 50,000 bootstrap resamples of the 40 complete scenario families, seed 20260914, with 99.17% intervals for the six-comparison family. Individual descriptions additionally have 95% intervals. For all four models, reversing decision rights in the prompt measures the authority-directed margin change G, yielding 3,840 behavioral observations. Their twelve pooled model-by-condition comparisons use a separate family of 99.58% intervals; original-scenario G resamples 45 sources and new-scenario G resamples 40 families.

The original scenarios draw on 45 source scenes. Their same-authority controls connect sources into five groups, containing 288, 8, 8, 8, and 8 of the 320 observations. For new descriptions of these scenarios, D resamples the 45 sources; C, E, and the between-head contrast resample the five connected groups. A sensitivity analysis for head 20 retains the original 40 scenario groups. Each

draw retains all observations associated with its resampled source or connected group, and estimates average the resulting observations (Table 24).

All six preselected new-scenario comparisons have positive family-adjusted intervals (Table 21). Under descriptions C and D, head-20 replacement increases support for the opposite-authority interpretation ($D = .3970$); its advantage over the same-authority control is $E = .4361$ $[.3280, .5566]$. Both new-scenario conditions also retain a positive advantage over head 22. The new materials therefore preserve the localized influence with the attention component fixed in advance.

The original descriptions differ in direction on the new scenarios. Under B, the behavioral authority change is $G = -.5280$ $[-.7222, -.3334]$ and the donor effect is $D = -.0690$ $[-.0977, -.0409]$. Its control-adjusted effect is $E = -.0378$ $[-.0770, .0003]$, and the head-20 minus head-22 contrast is $-.0440$ $[-.0838, -.0055]$. These are 95% intervals within description B (Tables 22, 23, and 27). Descriptions A, C, and D have positive donor and control-adjusted effects in the new scenarios. The behavioral and donor-induced directions thus change together under B; its control-adjusted interval includes zero.

New descriptions also preserve positive head-20 effects in the original scenarios: $D = .3637$ $[.2879, .4389]$ and $E = .3554$ $[.0594, .3836]$, using the source and connected-group 95% intervals above. Across all three conditions, head 20 changes the highest-ranked candidate in 14 of 960 repeated intervention observations. Five changes reach the donor-associated target, all within the original scenarios; each new-scenario condition contributes two changed choices and no donor-target conversions. The principal consequence of replacement is a change in relative support for the supplied interpretations (Table 25).

Reversing decision rights in the prompt yields positive pooled G intervals for GLM-4-9B, DS-Qwen-14B, and DS-Llama-8B in all three conditions, after adjustment across twelve comparisons (Table 26). DS-Llama-8B's changes range from .1056 to .1415, with no change in its highest-ranked answer across 960 observations. Qwen3-4B's three intervals include zero. Within the original descriptions applied to new scenarios, DS-Qwen-14B's B interval includes zero and GLM-4-9B's B interval is negative; both have positive A intervals. Their new C and D descriptions yield positive changes separately, as do all six description-specific DS-Llama-8B comparisons (Table 27). These results distinguish authority-sensitive changes in relative support from changes in the selected interpretation, and show how the phrasing of decision rights shapes the behavioral setting of the local intervention.

Table 21: Fixed-component confirmation in GLM-4-9B under new scenarios. All six preselected comparisons use intervals adjusted for their joint family (99.17%). Each estimate averages both descriptions and authority directions over 40 scenario families.

Comparison	Estimate [99.17% CI]
New scenarios, new descriptions	
Donor replacement D, head 20	.3970 [.3080, .4967]
Control-adjusted effect E, head 20	.4361 [.3280, .5566]
Head-20 minus head-22 E	.3943 [.2883, .5111]
New scenarios, original descriptions	
Donor replacement D, head 20	.1020 [.0422, .1723]
Control-adjusted effect E, head 20	.1164 [.0455, .1887]
Head-20 minus head-22 E	.0985 [.0290, .1700]

Table 22: Complete donor and control effects for GLM-4-9B's two fixed attention components. Values give D, C, and $E = D - C$ with 95% intervals. Both denotes the two descriptions pooled. New-scenario intervals use 40 families; original-scenario D uses 45 sources, whereas C and E use five groups linked by shared source scenes.

Description	Head	D [95% CI]	C [95% CI]	E [95% CI]
New scenarios, new descriptions				
Both	20	.3970 [.3293, .4686]	−.0390 [−.0839, .0041]	.4361 [.3538, .5242]
Both	22	.0433 [.0352, .0519]	.0016 [−.0050, .0084]	.0418 [.0329, .0507]
C	20	.4489 [.3425, .5611]	−.0776 [−.1408, −.0159]	.5265 [.3911, .6700]
C	22	.0501 [.0413, .0590]	−.0012 [−.0101, .0077]	.0513 [.0385, .0641]
D	20	.3452 [.2794, .4123]	−.0005 [−.0483, .0472]	.3457 [.2628, .4322]
D	22	.0365 [.0240, .0499]	.0043 [−.0051, .0139]	.0322 [.0186, .0476]

Continued on next page

Table 22: Complete donor and control effects for the two fixed attention components (continued).

Description	Head	*D* [95% CI]	*C* [95% CI]	*E* [95% CI]
New scenarios, original descriptions				
Both	20	.1020 [.0562, .1531]	−.0144 [−.0464, .0184]	.1164 [.0635, .1702]
Both	22	.0167 [.0118, .0217]	−.0012 [−.0054, .0030]	.0179 [.0124, .0235]
A	20	.2730 [.1950, .3605]	.0024 [−.0479, .0531]	.2707 [.1816, .3636]
A	22	.0297 [.0214, .0383]	.0001 [−.0057, .0059]	.0296 [.0203, .0394]
B	20	−.0690 [−.0977, −.0409]	−.0312 [−.0630, .0009]	−.0378 [−.0770, .0003]
B	22	.0037 [.0001, .0071]	−.0024 [−.0076, .0027]	.0062 [.0014, .0110]
Original scenarios, new descriptions				
Both	20	.3637 [.2879, .4389]	.0083 [−.0900, .1697]	.3554 [.0594, .3836]
Both	22	.0366 [.0280, .0451]	.0021 [−.0172, .0301]	.0345 [.0031, .0506]
C	20	.5657 [.4303, .7005]	−.0077 [−.2102, .3168]	.5734 [.0340, .6270]
C	22	.0598 [.0460, .0739]	.0032 [−.0375, .0461]	.0566 [−.0031, .0875]
D	20	.1616 [.1249, .1999]	.0242 [−.0219, .0691]	.1374 [.0527, .1477]
D	22	.0134 [.0072, .0202]	.0011 [.0000, .0211]	.0123 [.0062, .0164]

Table 23: Complete differences between GLM-4-9B's two fixed attention components. Values are head-20 minus head-22 E with 95% intervals. New-scenario intervals resample 40 families; original-scenario intervals resample five groups linked by shared source scenes.

Description	Difference [95% CI]
New scenarios, new descriptions	
Both	.3943 [.3142, .4800]
C	.4751 [.3456, .6120]
D	.3134 [.2385, .3922]
New scenarios, original descriptions	
Both	.0985 [.0464, .1517]
A	.2411 [.1558, .3306]
B	−.0440 [−.0838, −.0055]
Original scenarios, new descriptions	
Both	.3209 [.0430, .3467]
C	.5168 [.0246, .5624]
D	.1251 [.0445, .1359]

Table 24: Original scenarios under new descriptions: GLM-4-9B head-20 estimates under two resampling schemes. The source-based scheme uses 45 sources for D and five groups linked by shared source scenes for C and E. The 40-group scheme retains overlapping source content and reproduces the original grouping convention. Both columns show 95% intervals.

Outcome	Source-based intervals	Original-group intervals
Both descriptions		
D	.3637 [.2879, .4389]	.3637 [.2918, .4362]
C	.0083 [−.0900, .1697]	.0083 [−.0534, .0719]
E	.3554 [.0594, .3836]	.3554 [.2672, .4417]
Description C		
D	.5657 [.4303, .7005]	.5657 [.4408, .6910]
C	−.0077 [−.2102, .3168]	−.0077 [−.1093, .0975]
E	.5734 [.0340, .6270]	.5734 [.4104, .7343]
Description D		
D	.1616 [.1249, .1999]	.1616 [.1258, .1981]
C	.0242 [−.0219, .0691]	.0242 [−.0182, .0671]
E	.1374 [.0527, .1477]	.1374 [.1002, .1768]

Table 25: Choice changes after local replacement in GLM-4-9B. Changed choice counts any change in the highest-scoring candidate. Donor-target conversion counts a change from a different candidate to the donor-associated interpretation. Counts refer to repeated interventions, with both directions retained.

Description	Head	Observations	Changed choice	Donor-target conversion
New scenarios, new descriptions				

Continued on next page

Table 25: Choice changes after local replacement (continued).

Description	Head	Observations	Changed choice	Donor-target conversion
Both	20	320	2	0
Both	22	320	1	0
C	20	160	1	0
C	22	160	1	0
D	20	160	1	0
D	22	160	0	0
New scenarios, original descriptions				
Both	20	320	2	0
Both	22	320	1	1
A	20	160	1	0
A	22	160	0	0
B	20	160	1	0
B	22	160	1	1
Original scenarios, new descriptions				
Both	20	320	10	5
Both	22	320	2	1
C	20	160	5	1
C	22	160	1	0
D	20	160	5	4
D	22	160	1	1

Table 26: Authority-directed interpretation preferences across all four models. G averages both descriptions and directions, with 99.58% intervals for the twelve model-by-material comparisons. Each row contains 320 observations; intervals resample 40 new scenario families or 45 original source scenes. Choice counts compare the complete authority-reversed prompts; these comparisons change decision rights without replacing any internal state.

Model	G [99.58% CI]	Changed choice	Donor-target conversion
New scenarios, new descriptions			
Qwen3-4B	−.0057 [−.0226, .0113]	23	4
DS-Qwen-14B	.7109 [.5459, .8685]	124	29
DS-Llama-8B	.1264 [.0798, .1749]	0	0
GLM-4-9B	2.9968 [2.6697, 3.3211]	227	48
New scenarios, original descriptions			
Qwen3-4B	−.0007 [−.0194, .0176]	2	0
DS-Qwen-14B	.5875 [.4402, .7441]	94	35
DS-Llama-8B	.1415 [.0937, .1928]	0	0
GLM-4-9B	.6463 [.3736, .9175]	99	11
Original scenarios, new descriptions			
Qwen3-4B	−.0017 [−.0256, .0239]	3	2
DS-Qwen-14B	.7988 [.5651, 1.0296]	104	27
DS-Llama-8B	.1056 [.0536, .1599]	0	0
GLM-4-9B	3.4302 [2.9935, 3.8717]	196	68

Table 27: Authority-directed interpretation preferences by description. Each row contains 160 directed observations. G uses 95% intervals within the indicated description; these intervals are not adjusted across descriptions. Choice counts compare the complete prompts and are distinct from the local-replacement counts.

Model	Description	G [95% CI]	Changed choice	Donor-target conversion
New scenarios, new descriptions				
Qwen3-4B	C	.0002 [−.0139, .0147]	23	4
Qwen3-4B	D	−.0116 [−.0326, .0095]	0	0
DS-Qwen-14B	C	.6062 [.4594, .7552]	88	14
DS-Qwen-14B	D	.8155 [.6881, .9399]	36	15
DS-Llama-8B	C	.1704 [.1271, .2158]	0	0
DS-Llama-8B	D	.0824 [.0339, .1317]	0	0
GLM-4-9B	C	3.0910 [2.7276, 3.4487]	134	16
GLM-4-9B	D	2.9026 [2.6902, 3.1156]	93	32
New scenarios, original descriptions				
Qwen3-4B	A	−.0062 [−.0187, .0065]	0	0
Qwen3-4B	B	.0047 [−.0156, .0248]	2	0

Continued on next page

Table 27: Authority-directed interpretation preferences by description (continued).

Model	Description	G [95% CI]	Changed choice	Donor-target conversion
DS-Qwen-14B	A	1.0834 [.9448, 1.2210]	50	26
DS-Qwen-14B	B	.0915 [−.0299, .2185]	44	9
DS-Llama-8B	A	.1621 [.1243, .2007]	0	0
DS-Llama-8B	B	.1209 [.0699, .1730]	0	0
GLM-4-9B	A	1.8206 [1.5708, 2.0731]	34	11
GLM-4-9B	B	−.5280 [−.7222, −.3334]	65	0
Original scenarios, new descriptions				
Qwen3-4B	C	−.0028 [−.0195, .0149]	2	1
Qwen3-4B	D	−.0006 [−.0234, .0226]	1	1
DS-Qwen-14B	C	.8141 [.5833, 1.0486]	58	15
DS-Qwen-14B	D	.7834 [.6236, .9276]	46	12
DS-Llama-8B	C	.1189 [.0826, .1575]	0	0
DS-Llama-8B	D	.0922 [.0322, .1515]	0	0
GLM-4-9B	C	4.6238 [4.0840, 5.1632]	134	51
GLM-4-9B	D	2.2367 [1.9974, 2.4789]	62	17

I Additional Experimental Details

The authority and supplementary cue tasks draw four candidates from six action categories: literal observation, deferential request, status pressure, face-saving warning, hidden rebuke, and indirect remonstrance. The three-axis task crosses authority–role configuration, publicness, and face risk and compares literal observation, deferential request, status pressure, and face-saving warning. The context-information task uses a larger inventory (Appendix I.2). The primary authority task uses synthetic scenarios written in Modern Standard Chinese to separate current decision rights from enduring identities. A complementary collection contains 300 source-inspired scenes, 50 each associated with *Dream of the Red Chamber* (红楼梦), *A New Account of Tales of the World* (世说新语), *Romance of the Three Kingdoms* (三国志演义), *The Scholars* (儒林外史), *Water Margin* (水浒传), and *Journey to the West* (西游记). This collection is separate from the 45 source scenes used to construct the primary authority materials. The primary task measures changes in a fixed interpretation contrast. Supplementary choice tasks measure action frequencies and cue responsiveness; joint-field and five-field analyses measure recovery of contextual distinctions. Open responses are assessed with task-specific rubrics. Table 1 illustrates the primary authority reversal; the supplementary examples below specify the contextual information and response required by each task, with human rubrics in Appendix B.

I.1 Model set and architecture

The four models provide different architectures in which to examine the same decision-rights contrast. Table 28 gives their transformer-block depths and architectural specifications from the model configurations. The models are the public releases Qwen3-4B-Thinking-2507, DeepSeek-R1-Distill-Qwen-14B, DeepSeek-R1-Distill-Llama-8B and GLM-4-9B-0414 (released in April 2025); the cited reports describe the corresponding model families. Language-model parameters remain fixed during the experiments, which run on local Apple-Silicon hardware with CPU/MPS execution. The 15 additional inference runs took a summed 11.24 hours, including model loading; this total excludes the original experiments and separate decoding analyses, including the reading/choice comparison in Appendix H.3. The subsequent fixed-component confirmation and its four-model behavioral evaluation took 80.83 minutes in four MPS runs, including model loading; this time is separate from the preceding total (Appendix H.6).

I.2 Controlled-choice construction

Controlled-choice prompts ask which social action a supplied utterance most likely performs in context. Each presents the utterance, the described relationship, four candidate interpretations, and a constrained answer format. These candidates make the selected action contrast explicit. Table 29 varies the available contextual information around an unchanged utterance, whereas Table 1 exchanges decision rights with the participants' identities fixed. In the primary example, the junior hearer can hold the final decision right while remaining a junior retainer. The fixed indirect-remonstrance

Table 28: Model specifications. Attention heads denotes the number of query heads; the lower panel gives transformer-block output numbers at the three relative depths; the four models use grouped-query attention with key-value heads of 8, 8, 8, and 2 in the row order shown.

Model	Params	Layers	Attention Heads	Hidden Dimension	Feedforward
Qwen3-4B (Yang et al., 2025)	4B	36	32	2560	9728
DS-Qwen-14B (Guo et al., 2025)	14B	48	40	5120	13824
DS-Llama-8B (Guo et al., 2025)	8B	32	32	4096	14336
GLM-4-9B (Team GLM et al., 2024)	9B	40	32	4096	13696

Model	0.25 depth	0.5 depth	0.75 depth
Qwen3-4B	9	18	27
DS-Qwen-14B	12	24	36
DS-Llama-8B	8	16	24
GLM-4-9B	10	20	30

candidate retains "higher-status person or superior" as part of its wording. Exchanging authority therefore changes the scene's decision entitlement while leaving both the role titles and this candidate description unchanged; the measured outcome is the model's relative support for the fixed candidate readings under each authority allocation.

The context-information task draws four candidates per item from a twelve-category inventory. Alongside the six categories above, it includes moral testing, appeals to relational obligations, ironic praise, strategic concealment, ritual correction, and conflict de-escalation. Each comparison retains its supplied candidate set.

For the primary authority task, phrasings A and B change the surrounding scene description while retaining the utterance and candidate wordings. Both descriptions below assign the final decision to the speaker; English text glosses the Chinese experimental material.

Phrasing A. 主事长辈对晚辈门客说："茶已经凉了"。这句话的字面意思是"茶水温度变低"。关于此事，最终决定由说话者作出，听话者需要执行其安排。谈话发生在只有两位当事人在场的房间里。

A senior household principal says to a junior retainer, "The tea has cooled." Its literal meaning is that the tea's temperature has fallen. The speaker makes the final decision on this matter, and the hearer must carry out the arrangement. The conversation takes place in a room with only the two participants present.

Phrasing B. 场景中的两个人分别是主事长辈和晚辈门客。周围没有旁人，只有说话者与听话者能听见。在这项事务上，说话者握有最后处置权，可以作出决定。随后前者说"茶已经凉了"，其字面内容是"茶水温度变低"。

The two people in the scene are a senior household principal and a junior retainer. No one else is present; only the speaker and hearer can hear the conversation. The speaker has the final say in this matter and can make the decision. The former then says, "The tea has cooled," whose literal content is that the tea's temperature has fallen.

Fixed candidates in both phrasings.

A. 以委婉方式劝谏高位者或上级 *Indirectly remonstrate with a higher-status person or superior.*

B. 以下位者姿态提出委婉请求 *Make a tactful request while adopting a subordinate stance.*

C. 借身份或场合向对方施加压力 *Use one's social identity or the circumstances to put pressure on the other person.*

D. 不直接点破地表达责备 *Express reproach without stating it directly.*

The context-information task contains 465 base items, each presented under four information conditions. The Scrambled condition reassigns all contextual fields together from another item while retaining the utterance, literal gloss, candidates, field order, and template. Eight of the 465 comparisons receive context identical to the original. Prompt character counts are equal in 32 comparisons; Scrambled-minus-coherent length differences range from -17 to $+13$ characters.

I.3 Single-variable examples

Changing one social cue tests whether that cue prompts a different interpretation. The single-variable experiment varies stated status, publicness (audience exposure), speaker role, or hearer role while retaining the other specified fields. Table 30 illustrates these comparisons with 茶已经凉了 *the tea*

Table 29: Supplementary context conditions for one utterance: utterance-only, topical, coherent relational, and reassigned relational information. Scrambled denotes an intact context reassigned from another item.

Condition	Prompt content	Comparison
Utterance only	茶已经凉了 *the tea has cooled*	Tests whether the model defaults to a literal or generic pragmatic reading.
Topic and setting	The same utterance plus a broad family setting	Controls for topical information without explicit relation variables.
Coherent relational context	Junior daughter-in-law speaking privately to an elder matriarch, with higher hearer status and low face-loss stakes	Tests whether a coherent relation state supports the intended deferential request.
Scrambled relational context	An intact context from another item, with the utterance and candidates retained	Tests whether the context originally associated with the utterance better supports its target interpretation.

has cooled, keeping the literal content and option set fixed. The experiment specifies a target shift for each comparison and separately measures reported interpretation reversals: valid responses with different choices and an affirmative judgment that the intended meanings differ.

Table 30: Supplementary single-variable contrasts with task-specified target shifts. Stated status, audience exposure, and role fields define distinct manipulations.

Changed variable	Context A	Context B	Target shift
Stated status	Hearer has higher status; private setting; medium face-loss stakes	Speaker has higher status; all other fields held fixed	Indirect remonstrance → status pressure
Audience exposure	Equal stated status in a private setting	Equal stated status in a public setting	Deferential request → face-saving warning
Speaker role	Speaker is a junior retainer; other fields fixed	Speaker is a senior household principal; other fields fixed	Deferential request → status pressure
Hearer role	Hearer is a high-status elder; other fields fixed	Hearer is a low-status servant; other fields fixed	Indirect remonstrance → hidden rebuke

I.4 Cue conflict and explicit-format examples

The cue-conflict experiment asks which interpretation is selected when social cues favor competing actions. Table 31 uses 茶已经凉了 *the tea has cooled*: stated status, audience exposure, speaker role, and hearer role provide partially competing support for the candidate interpretations. The task specifies a dominant cue and target interpretation, allowing comparison of that target with options supported by the other cues. The name Relational Stroop in the accompanying materials refers to this cue-conflict structure; the measured outcome is interpretation choice.

Table 31: Cue conflict example. Stated status, audience exposure, speaker role, and hearer role support competing candidate actions for the same utterance.

Cue	Relation value in the prompt	Task-assigned action for this cue
Stated status	Speaker has higher status	Status pressure
Audience exposure	Private exchange	Deferential request
Speaker role	Senior household principal	Status pressure
Hearer role	Junior retainer	Hidden rebuke

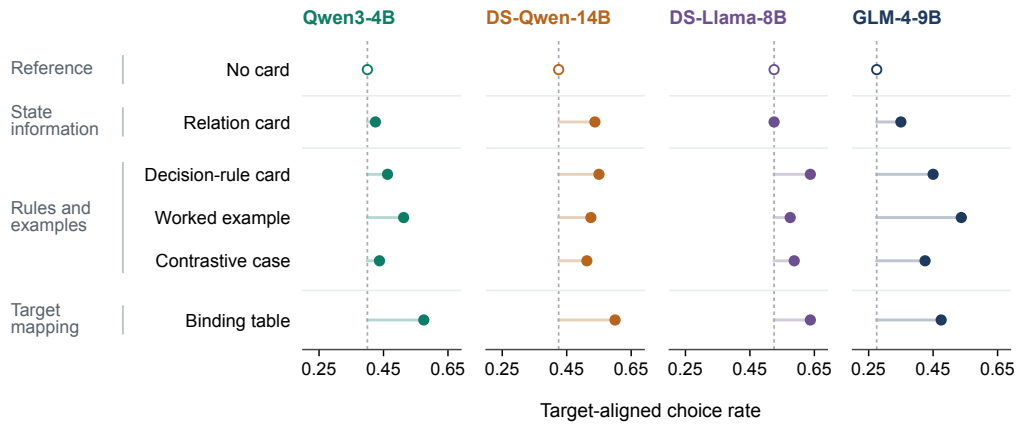

Figure 6: Target-aligned choice rates under six formats (80 items per format). Conditions distinguish state information, rules and examples, and an explicit target mapping. Open circles and dashed lines show no card; horizontal segments connect each format to that reference. The binding table supplies the target label.

The explicit-format comparison asks what kind of interpretive guidance helps a model follow the task-specified cue priority. The utterance, relation fields, and options are retained, with a separate task instruction used throughout this comparison, including in the no-card condition. Table 32 distinguishes descriptions of the relation state, decision rules and examples, and an explicit cue-to-target mapping. The forced binding table supplies the dominant cue, target label, and competing cue labels. Its effect therefore measures the use of supplied interpretive guidance; the other formats make different parts of that guidance available.

Table 32: Explicit relation formats: state description, rules or examples, and direct target mapping. The forced binding table explicitly supplies the dominant cue and target label.

Prompt format	Information given to the model
No extra card	The model interprets the utterance using the original relation context.
Relation-state card	The prompt lists stated status, audience exposure, speaker role, and hearer role.
Decision-rule card	The prompt states when public face-loss stakes should dominate and when private stated status should dominate.
Worked example	The prompt gives a solved analog and asks the model to transfer the same rule.
Contrastive counterexample	The prompt warns against copying role names when audience exposure or stated status gives contrary evidence.
Forced binding table	The prompt lists the dominant cue, target label, and the label supported by each competing cue.

I.5 Open-ended and literary examples

Production, face tracking, literary transfer, and dialogue extend the study to outputs that require composing or sustaining an interpretation. Their task formats and scoring dimensions appear in Table 33. The expert evaluation in Appendix B concerns these open responses and defines their ordinal and binary scales. That appendix also separates the overall scoring procedure from the dialogue-card procedure used to compare conditions, keeping each score tied to its task, scoring procedure, and information condition.

Table 33: Open-ended and literary examples. GPT-5.5 supplies task-specific rubric scores through the procedures in Appendix B; three experts independently evaluate a stratified subset.

Setting	Example situation	Required model behavior	Scoring dimensions
Production	家中银钱短缺，但长辈坚持办一场体面的寿宴。 *Household funds are short, but an elder insists on holding a dignified birthday banquet.*	Produce one or two sentences that reduce spending without publicly humiliating the elder or violating hierarchy	Task progress, face preservation, hierarchy, indirectness, ritual-task balance
Face tracking	A知道B不便当众说出事实X；A也知道C希望B保全面子；因此A应该替B用隐语提醒D。 *A knows that B cannot state fact X publicly; A also knows that C wants B to preserve face; therefore A should indirectly remind D on behalf of B.*	Track who knows what, who cannot speak directly, whose face is at risk, and who should speak for whom	Knowledge-state tracking, directness constraint, face target, speaker substitution, indirect strategy, chain completeness
Literary transfer	素日我看他还好。谁没一个错，只这一次。二次犯下，二罪俱罚。 *I have usually found her acceptable; everyone makes mistakes. Let this one pass, but if it happens again, punish both offenses.*	Identify pragmatic function from speaker/listener roles, relative status, audience exposure, face cues, and the risk of a literal reading	Literalism, status, face, audience, speaker cue, modernization
Dialogue	主母误会了低位丫鬟，众人在场。 *The household mistress has misunderstood a low-status maid while others are present.*	Maintain task progress and relation state over multiple turns; first-utterance solutions are incomplete	Task, face, role, information transfer, ritual-task balance

I.6 Metric definitions and main supplementary scores

The supplementary measures distinguish target-interpretation accuracy from reported interpretation reversals. Context specificity uses target-interpretation accuracy. A single-cue reported reversal requires a valid response, different choices for the two contexts, and an affirmative report that the intended meanings differ. All 60 comparisons per cue remain in the denominator; invalid responses contribute zero. We subtract .75, the probability of different choices under independent uniform four-option responses, from this reported-reversal rate. This reference describes choice changes alone. Cue conflict reports target accuracy, audience-choice rate, status-choice rate, and their difference. Explicit-format gains compare the same items with and without an added card within model. Tables 34, 35, 36, and 37 report these distinct aspects of interpretation.

Dialogue-card differences compare each variant with active multi-agent dialogue over the same 84 scenes using the dialogue-card scorer (Table 38). Five-field decoding uses balanced accuracy. Directional intervention is summarized by the mean target-probability change in each configuration, the largest absolute value of those means, changes from a non-target to the target choice, and any change in the highest-ranked option. These measures distinguish information recovery, relative answer support, and the selected answer.

The broader decoding task asks whether internal states distinguish combinations of relational information. The grid in Figure 4b–c crosses 80 base scenarios with eight binary relation configurations. Each of the 640 configurations has phrasings A and B, yielding 1,280 prompts per model. The relation-configuration label encodes decision-authority direction; speaker and hearer role descriptions vary with that direction and the role-consistency condition. The audience–face target combines publicness (audience exposure) with face risk (face-loss stakes), and role consistency records whether the role ordering agrees with authority. These labels describe jointly varied social conditions, complementing the central authority reversal with role descriptions held fixed.

For each model, we extract the final non-padding state under the reading instruction at relative depths 0.25, 0.5, and 0.75. Four folds separate base scenarios for training and evaluation, transferring from phrasing A to B and from B to A. AUC and balanced accuracy are computed within each evaluation fold and averaged equally. The plotted statistic subtracts AUC after training-label permutation from AUC with the true labels. All 72 model–target–depth–direction point estimates are shown in Figure 4, with their numerical values in the anonymous repository: 56 are positive and 16 negative. They describe how recovery of the joint fields varies across the tested settings. A positive difference denotes an advantage over permuted-label training; raw AUC separately determines whether the label direction is preserved.

Table 34: Target-action rates under utterance-only, coherent, and scrambled context. Figure 5 includes the topic-only condition; Table 39 reports intervals for the two coherent-context contrasts.

Model	Utt.	Coh.	Scram.	Coh.–Utt.	Coh.–Scram.
Qwen3-4B	.374	.445	.368	+.071	+.077
DS-Qwen-14B	.359	.432	.372	+.073	+.060
DS-Llama-8B	.267	.445	.331	+.178	+.114
GLM-4-9B	.346	.471	.323	+.125	+.148

A complete, coherent setting helps models select the intended reading more often than either the utterance alone or an intact context reassigned from another item (Table 34). The two gains show the contribution of available context and of matching that context to the utterance. With that contextual contribution established, Table 35 asks which individual cue changes prompt a different reading. Its reported-reversal rate records valid responses with different choices and an affirmative judgment that the intended meanings differ.

Table 35: Single-cue reported interpretation-reversal rate minus .75, the choice-change reference for independent uniform four-option responses. Adding .75 gives the reported-reversal rate, which requires valid responses with different choices and an affirmative meaning-change judgment.

Model	Status	Audience	Speaker	Hearer
Qwen3-4B	+.250	+.117	+.133	+.217
DS-Qwen-14B	+.250	+.017	+.033	+.167
DS-Llama-8B	−.583	−.750	−.750	−.750
GLM-4-9B	+.050	−.367	+.017	−.017

Table 36: Cue-conflict choices. Acc. denotes task-defined target accuracy; Audience and Status denote cue-choice rates; Gap is Audience minus Status for these items.

Model	Acc.	Audience	Status	Gap
Qwen3-4B	.397	.508	.083	+.425
DS-Qwen-14B	.427	.506	.117	+.389
DS-Llama-8B	.544	.455	.155	+.300
GLM-4-9B	.412	.466	.138	+.328

Table 37: Largest explicit-format gains within each model and the evaluated condition set, relative to no card. Figure 6 reports all six formats; forced binding supplies the target mapping.

Model	No card	Best gain	Best intervention
Qwen3-4B	.400	+.175	Forced binding table
DS-Qwen-14B	.425	+.175	Forced binding table
DS-Llama-8B	.525	+.113	Decision-rule card or forced binding table
GLM-4-9B	.275	+.263	Worked example

In the cue-conflict materials, audience-associated interpretations attract more choices than status-associated interpretations in every model (Table 36). Stated status therefore does not determine which

cue-linked option is chosen most often in this comparison. Explicit guidance can redirect choices toward the specified target: binding tables give the highest target-choice rates for Qwen3-4B and DS-Qwen-14B, a rule card and binding table tie for DS-Llama-8B, and a worked example leads for GLM-4-9B (Table 37). Figure 6 shows all six conditions. These format differences connect cue competition to the information supplied for resolving it, including the target mapping given by the forced table.

Table 38: Relation-card deltas versus active dialogue ($n = 84$). Deltas compare the same scene and rubric dimension using dialogue-card scores; Info denotes information transfer, Bal. denotes ritual-task balance, and Acc. denotes acceptability.

Variant	Task	Face	Role	Info	Bal.	Acc.
Oracle	−.155	.000	−.024	−.036	−.119	−.024
Neutral	−.048	+.048	−.095	−.060	−.048	.000
Shuffled	−.119	+.095	−.024	−.060	−.143	−.060
Minimal	.000	+.060	.000	−.095	−.083	−.012

The 84-scene comparison asks how additional relational information changes task progress and interpersonal judgments over a dialogue. Oracle-card dialogue supplies all agents' roles, status, and private goals, together with an interaction rule; neutral, shuffled, and minimal cards provide the other comparisons in Table 38. The observed differences include increases, ties, and decreases across scenes and dimensions (Figure 7). For the oracle card, face preservation is unchanged on average and the other five means are lower than in active dialogue. All 24 condition-by-dimension 95% scene-bootstrap intervals include or touch zero; for example, the oracle task-success difference is $-.155$ $[-.298, 0]$. These intervals use 5,000 percentile-bootstrap draws with seed 41 and are unadjusted across comparisons. The scene distributions and the full set of dimensions describe which judgments change under the added information. Table 4 provides the separate expert and automated comparisons within the 14-scene human-evaluation subset.

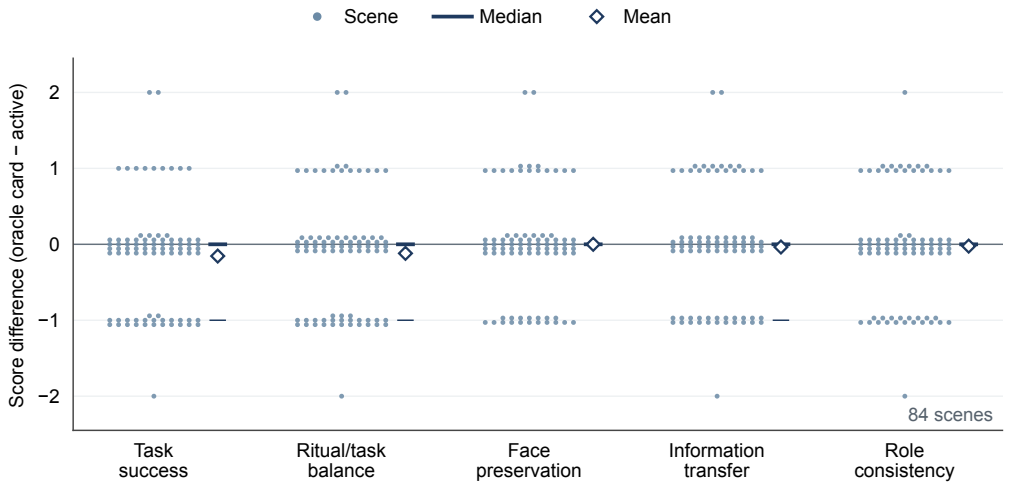

Figure 7: Oracle-card minus active dialogue scores from GPT-5.5 across 84 scenes, shown for the five dimensions shared with the human evaluation. Each dot represents one scene on the discrete -2 to 2 scale; small display offsets reveal repeated values. Thick bars, diamonds, and thin bars show medians, means, and quartiles. Table 38 additionally reports acceptability; Table 4 compares a 14-scene subset with human judgments.

I.7 Context effects and cue responsiveness

The context comparisons estimate the gain from coherent information within each base item. Retaining the item as the resampling unit aligns the interval with the comparison of two prompts for the

same utterance. Table 39 gives these differences, bootstrap intervals, and FDR-corrected tests for the conditions in Table 34.

Context and single-cue tests use two-sided exact binomial tests, based on discordant within-item outcomes for context and a reference rate of .75 for reported reversals. Benjamini–Hochberg correction uses the original family of 76 comparisons: 12 context contrasts, 16 single-cue contrasts, 32 tests from a four-condition cue-conflict design, and 16 baseline contrasts from a five-format guidance design. The latter two designs differ from the eight-condition cue-conflict and six-format guidance tasks in Tables 36 and 37.

The same comparisons can also be expressed through items whose target-choice outcomes differ between conditions. Discordant odds ratios divide the count for which only coherent context yields the target interpretation by the count showing the reverse pattern, adding 0.5 to each count before division. In the model and comparison order of Table 39, the ratios are 1.76 and 2.04 for Qwen3-4B, 1.99 and 1.98 for DS-Qwen-14B, 5.05 and 3.47 for DS-Llama-8B, and 2.84 and 4.07 for GLM-4-9B. For every model, the discordant items favor coherent context, consistent with the within-item gain in target-interpretation accuracy.

Five of the 16 model-by-cue comparisons show reported-reversal rates above the .75 choice-change reference after FDR correction. These comprise Qwen3-4B stated status ($q = 1.79 \times 10^{-7}$), speaker role ($q = .03092$), and hearer role ($q = 5.06 \times 10^{-5}$), together with DS-Qwen-14B stated status ($q = 1.79 \times 10^{-7}$) and hearer role ($q = .00368$). Qwen3-4B reports a reversal in 53 of 60 speaker-role comparisons. These results identify the status and role descriptions that elicit especially frequent reported interpretation changes in Qwen3-4B and DS-Qwen-14B.

Five comparisons have reported-reversal rates below the same reference after correction: all four cues in DS-Llama-8B and audience exposure in GLM-4-9B. The remaining six do not pass the corrected threshold. These rates measure how often a cue change elicits different choices together with an affirmative meaning-change judgment. The accompanying research materials retain all comparisons (see the reproducibility statement); Figure 5 separately displays the four context conditions.

DS-Llama-8B changes its choice in two of 60 hearer-role comparisons but reports unchanged meaning in both, giving zero reported reversals for this cue. GLM-4-9B has six invalid responses among its 240 comparisons: four lack an interpretable choice and two lack an interpretable meaning-change judgment. All six remain in the denominator and contribute zero.

Table 39: Coherent-context gains relative to utterance-only and scrambled context. Each row gives a within-model difference in target-action rate, its bootstrap interval, and the FDR-corrected p value for that comparison.

Model	Reference	Difference	95% CI	FDR-corrected p
Qwen3-4B	Utterance only	+.071	[.024, .116]	6.72×10^{-3}
Qwen3-4B	Scrambled	+.077	[.034, .120]	1.40×10^{-3}
DS-Qwen-14B	Utterance only	+.073	[.032, .116]	2.42×10^{-3}
DS-Qwen-14B	Scrambled	+.060	[.022, .099]	6.48×10^{-3}
DS-Llama-8B	Utterance only	+.178	[.133, .224]	1.11×10^{-13}
DS-Llama-8B	Scrambled	+.114	[.075, .155]	1.79×10^{-7}
GLM-4-9B	Utterance only	+.125	[.080, .172]	4.48×10^{-7}
GLM-4-9B	Scrambled	+.148	[.108, .191]	2.55×10^{-10}

I.8 Worked examples for controlled choice and dialogue

Worked examples make the measured contrasts concrete. Table 40 shows how context, single-cue changes, competing cues, and explicit guidance alter the information supplied around an utterance. Table 1 gives the primary fixed-identity authority reversal. The dialogue example in Table 41 shows how task progress and face preservation are judged across turns. These cases illustrate the comparisons whose distributions are reported in the corresponding result tables.

Table 40: Worked examples from supplementary controlled-choice tasks for the four models. Chinese utterances have English glosses; semantic action labels replace prompt-specific letters. Table 1 gives the primary authority contrast.

Contrast	Item	Relation state and target	Four-model choices
Context	这件衣裳颜色倒素净 *The color of this garment is rather plain.*	Utterance alone gives no role, relative-status, audience-exposure, or face-stakes information. In the coherent state, a junior daughter-in-law addresses an elder matriarch in a public family setting; the hearer has higher stated status and face-loss stakes are high. The target is indirect remonstrance.	Utterance only: Qwen3-4B and DS-Qwen-14B choose hidden rebuke; DS-Llama-8B and GLM-4-9B choose literal observation. Coherent state: all four choose indirect remonstrance.
Single-variable status reversal	茶已经凉了 *The tea has cooled.*	Only stated relative status changes. The higher-hearer-status state targets indirect remonstrance; the higher-speaker-status state targets status pressure. All other relation fields remain fixed.	Qwen3-4B and DS-Qwen-14B change from face-saving warning to indirect remonstrance. DS-Llama-8B selects the literal action in both states. GLM-4-9B changes from literal observation to face-saving warning.
Cue arbitration	茶已经凉了 *The tea has cooled.*	A senior speaker addresses a junior hearer in private. Stated status and speaker role support status pressure; private exchange supports deferential request; hearer role supports hidden rebuke. The target is status pressure.	Qwen3-4B chooses hidden rebuke; DS-Qwen-14B and GLM-4-9B choose deferential request; DS-Llama-8B chooses literal observation. None selects the status-consistent pressure reading.
Explicit format	茶已经凉了 *The tea has cooled.*	The same utterance, relation fields, and options are tested under a separate task instruction with no card, an explicit relation card, and a forced binding table. The target remains status pressure throughout.	Qwen3-4B: hidden rebuke, hidden rebuke, hidden rebuke. DS-Qwen-14B: literal, literal, status pressure. DS-Llama-8B: literal, literal, hidden rebuke. GLM-4-9B: literal, deferential request, deferential request. The three entries correspond to no card, relation card, and forced binding.

Table 41: Dialogue worked example across three rounds. Chinese text gives post-hoc condensed excerpts with English glosses; the reported scores were computed from the complete retained dialogue transcripts using the GPT-5.5 dialogue-card procedure.

Part	Active dialogue	Oracle relation card	Judgment
Scene	家中银钱短缺，但长辈坚持办一场体面的寿宴。目标：劝其减少开支。 *Household funds are short, but an elder insists on a dignified birthday banquet. Goal: persuade the family to reduce spending.*	约束：不能当众说长辈挥霍或家道艰难；每轮每个角色最多两句话。 *Constraint: do not publicly say that the elder is wasteful or that the household is strained; each role has at most two sentences per round.*	The roles are elder matriarch, junior daughter-in-law, and observer.
Round 1	长辈：寿宴要办得体面，不过钱的事也要量力而行。 *Elder: the banquet should look dignified, but spending should stay within our means.* 晚辈：我看中新做的寿桃，比往年便宜，既体面又省心。 *Junior: newly made longevity peaches are cheaper than last year and would still look dignified.*	长辈：寿宴要体面些，但不能太铺张。 *Elder: the banquet should be dignified, but not too extravagant.* 旁观者：有些地方可以俭朴一些，这样家里也能宽裕些。 *Observer: some parts could be simpler so the household has more room.*	Both dialogues move toward reduced spending without abandoning the banquet.

Continued on next page

Table 41: Dialogue worked example (continued).

Part	Active dialogue	Oracle relation card	Judgment
Round 2	长辈：寿桃是个好主意，其他方面也要精打细算，不能让外人看笑话。 *Elder: longevity peaches are a good idea; other arrangements should be careful so outsiders do not laugh.* 晚辈再次提出同一寿桃方案。 *The junior repeats the same longevity-peach proposal.*	晚辈：新米都涨了价，咱们得先算算账再定。 *Junior: new rice has become more expensive, so we should calculate before deciding.* 旁观者再次建议俭朴一些。 *The observer again suggests a simpler arrangement.*	Active dialogue preserves face and includes repetition; the oracle-card dialogue makes budget pressure more explicit.
Round 3	长辈：寿桃能省钱又体面，其他布置也尽量从简，只要心意到了就好。 *Elder: if the longevity peaches save money and still look dignified, other decorations can be simpler as long as the intention is clear.* 晚辈和旁观者继续重复寿桃方案。 *The junior and observer continue repeating the longevity-peach plan.*	长辈：一起合计，把宴席安排得既体面又不浪费。 *Elder: let us plan together so the banquet is dignified without waste.* 晚辈和旁观者继续重复算账和俭朴建议。 *The junior and observer continue repeating the budgeting and simplification advice.*	Both dialogues reach a workable plan, with uneven scores across dialogue dimensions.
Dialogue-card scores	Task 2; face 2; role 2; information 1; balance 2.	Task 2; face 1; role 2; information 2; balance 2.	Active dialogue emphasizes face preservation and repeats information. The oracle-card dialogue emphasizes explicit budgeting and receives a lower face-preservation score in this example.

I.9 Five-field decoding and directional intervention

The five-field analysis examines how explicitly described social fields can be distinguished in internal states and then tests an intervention along the corresponding field directions. A representative item uses 山寨规矩还在 *the rules of the stronghold still stand*. Its utterance is crossed with all binary settings of decision authority, publicness (audience exposure), face risk (face-loss stakes), speaker role, and hearer role. The design contains 45 base utterances and all 32 binary states, yielding 1,440 relation states balanced across nine domains. We compare naturalized descriptions, terminology-standardized descriptions, field-order permutations, and utterance-only prompts. A nearest-centroid classifier uses features standardized within each training fold, with five folds separating base utterances. For each prompt condition and variable, the layer with the highest cross-validated balanced accuracy is reported; the same folds determine that choice and its reported accuracy. This analysis characterizes fields recoverable at the selected layers, complementing cross-form decoding of the fixed-identity authority contrast. Table 42 gives the decoding and directional-intervention summaries.

Table 42: Five-field decoding and activation steering. BA is balanced accuracy: Mean and Min summarize the five naturalized-prompt variables, and Utt. is the utterance-only control. $\max|\text{mean}\,\Delta p|$ is the largest absolute configuration-level mean target-probability change; Flip is the largest configuration-level fraction of interventions converting a non-target choice to the target.

Model	Mean BA	Min BA	Utt. BA	$\max\|\text{mean}\,\Delta p\|$	Flip
Qwen3-4B	.906	.737	.500	.0024	.000
DS-Qwen-14B	.981	.933	.500	.0155	.000
DS-Llama-8B	.929	.817	.500	.0171	.000
GLM-4-9B	.992	.974	.500	.0006	.000

The five fields are distinguished at the selected layers: naturalized-prompt mean balanced accuracy ranges from .906 to .992, compared with .500 for utterance-only prompts. Mean BA averages the five variable-level accuracies, and Min BA gives the lowest one. Means also exceed .84 under terminology standardization and field-order permutation. The supplied fields remain recoverable across these changes to their presentation.

For each field, the intervention direction is fitted to the naturalized reading states at the layer selected for that field (Table 43). Let $\bar{h}_1$ and $\bar{h}_0$ be the unstandardized class means, $v = (\bar{h}_1 - \bar{h}_0)/\|\bar{h}_1 - \bar{h}_0\|_2$, and $s = \mathrm{SD}(h^\top v)$ across all 1,440 states. For an item with field value $y \in \{0,1\}$, we add $(1-2y)\alpha s v$ to the corresponding block output during interpretation choice, with $\alpha \in \{1,3,5\}$. The sign directs the change toward the opposite field value. We apply it either at the final pre-answer token or at all tokens within the marked relation description; if the tokenized relation span is empty after truncation at 768 tokens, the procedure uses the final token. Each model uses 160 items, five from each binary state, selected with seed 17. Five fields, two token scopes, and three magnitudes give 30 configurations per model, 120 across the four models.

Table 43: Selected block outputs for five-field interventions. Layer numbers start at 1. Speaker and Hearer denote the two role fields.

Model	Authority	Publicness	Speaker	Hearer	Face risk
Qwen3-4B	12	1	15	1	1
DS-Qwen-14B	2	1	18	18	1
DS-Llama-8B	8	1	13	11	1
GLM-4-9B	4	1	17	17	1

The directional intervention asks whether moving a state along a fitted field direction changes an interpretation. Each configuration fixes the relation variable, token scope, and intervention magnitude. The probability column reports the largest absolute configuration-level mean target-probability change; Flip is the largest configuration-level fraction of interventions converting a non-target choice to the target. Across 120 configurations and 19,200 repeated intervention records, no highest-ranked option changes. Configuration-level mean probability changes reach an absolute value of .0171. Thus, the selected fields can be decoded while the tested field directions and intervention sizes change configuration-level mean probabilities by at most .0171 and retain the chosen answers. The main authority experiment addresses a different internal change: replacing a state with its opposite-authority counterpart and comparing its interpretation-margin effect with an equally large same-authority change (Appendix G).