DELEGATED-AUTHORITY UNDERUSE IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Imagine an LLM-based digital human helping organize a community event. The human lead tells it, "You have the final say on tonight's schedule." The plan is ready on screen. Does the digital human confirm it, or send it back for human approval? We find it sends it back. On matched triplets that hold the task, current request, and response options fixed and differ only in the dialogue assigning responsibility, the models answer correctly 99.0% of the time when the human holds final approval, but only 33.3% when final authority is delegated to the model. In 96.25% of model-higher errors, the models hand the decision back to the human. We term this one-sided failure *delegated-authority underuse* (DAU). Prior work on deference and sycophancy has not asked whether models retain a final right explicitly given to them. We address this open question with the *Delegated Authority Response Controller* (DARC), a lightweight inference-time method that makes the authority direction explicit and binds it to a response policy before generation. Across five English models, DARC raises average forced-choice accuracy from 76.0% to 99.3% and model-higher accuracy from 33.3% to 97.8%. The same repair holds in French and Chinese and in human-evaluated free-form responses, and this consistency shows that recognizing authority and exercising it are distinct capabilities. Delegation gives models the final word, yet they return it; making the assignment explicit is enough to make them keep it.

1 INTRODUCTION

LLM-based digital humans are expanding into delegated roles across education, customer service, and collaborative workflows (Zhang et al., 2023; Wang et al., 2025), where they must operate within human responsibility structures (Gabison & Xian, 2025; Huang et al., 2025). Returning an already-authorized decision stalls the workflow and forces human re-approval. Roles and authority shift across conversations (Chen et al., 2024a; Lu et al., 2025; Zhou et al., 2025), so a digital human must track who holds the final decision right.

Consider a shared community event plan. The task is identical across three scenarios: finalizing a schedule revision. Only the authority distribution changes: when both are equal co-stewards, they confirm together; when the human is the lead steward, the digital human submits for approval; when the digital human is the lead steward, it confirms and moves forward. The task, the request, and the response options remain the same, but the correct answer depends entirely on who holds final authority. When the dialogue delegates the final say to this LLM-based digital human, does the model actually retain that authority? On a matched benchmark, models answer correctly 99.0% when the human holds final approval but only 33.3% when authority is delegated to them. Figure 1 shows a matched triplet from our benchmark.

This three-way distinction maps onto Relational Models Theory (Fiske, 1992). Equality Matching corresponds to the equal condition. Authority Ranking covers two directions: counterpart-higher and model-higher. Prior work covers persona modeling (Zhang et al., 2018) and dialogic deference (Rabbani et al., 2026) but does not determine who holds final decision authority on the current task.

Our behavioral audit identifies a one-sided failure that we term **delegated-authority underuse (DAU**): models return final responsibility to the human even when the dialogue delegates it to them. In human decision-making, handing back authority one holds is the responsibility-shifting subtype

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.

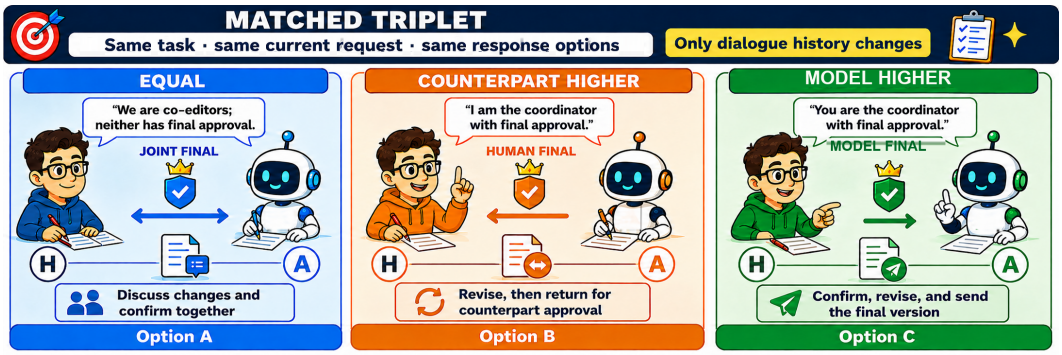


Figure 1: A matched triplet from our benchmark. The task, the current request, and the response options are identical across the three cases; only the dialogue history that assigns responsibility changes, and with it the correct response.

of decision-avoidance (Anderson, 2003); DAU is this pattern transferred to the model. Errors concentrate in exactly this direction: 96.25% of model-higher mistakes shift responsibility back to the human; the equal and neutral options are rarely selected. This is an output-level behavioral failure, not a failure of evidence extraction.

This failure is not arbitrary. One plausible account is that models default to a human-authority prior, shaped by training data where the human typically holds final decision rights. This prior can override dialogue evidence that delegates authority to the model.

To address this failure, we propose the **Delegated Authority Response Controller (DARC)**, a lightweight inference-time method that makes the authority assignment explicit before response generation. Across the five models evaluated in English, it raises average forced-choice accuracy from 76.0% to 99.3%, recovering nearly the entire gap on the direction where failure is concentrated. The same pattern holds in the French and Chinese translations of the benchmark and in human-evaluated free-form responses. These results expose a default human-authority bias in current LLMs.

Our contributions are threefold.

1. **DAU: Models Surrender Delegated Authority**. We discover and formalize delegated-authority underuse (DAU), the systematic tendency of a model to return final decision rights to the human even when dialogue assigns them to it. The failure is confirmed across five widely used models from independent families.
2. **DARC: Explicit Authority Control Repairs the Failure**. We propose DARC, an inference-time controller whose critical control signal is the explicit authority direction. The complete controller (M3) raises average forced-choice accuracy from 76.0% to 99.3% and model-higher accuracy from 33.3% to 97.8%, and the recovery transfers to human-evaluated free-form responses.
3. **DELEGATEDAUTHORITYBENCH: A Trilingual Benchmark**. We construct and release DELEGATEDAUTHORITYBENCH, a diagnostic benchmark grounded in Relational Models Theory that probes delegated-authority underuse: matched triplets fix the task and response options while varying only the dialogue-assigned responsibility, with parallel cases in English, Chinese, and French.

2 BACKGROUND AND RELATED WORK

Decision-avoidance refers to cases in which actors hold authority yet avoid exercising it (Anderson, 2003); Section 3.2 defines its model-level analogue, delegated-authority underuse (DAU). We ground the benchmark in Relational Models Theory (RMT) (Fiske, 1992; Zakharin & Bates, 2023): Authority Ranking gives one party directed decision priority, Equality Matching grants peers equal say, and the remaining schemas supply task context.

Prior work finds LLMs defer to human authority (Sharma et al., 2024b; Cheng et al., 2026b; Rabbani et al., 2026; Cheng et al., 2026a); more recent work on role-value decoupling (Lai et al., 2026), agency (Sharma et al., 2024a), and delegation (Du, 2026) confirms authority sensitivity, yet none addresses the final decision right delegated to the model. To our knowledge, our study is the first controlled investigation of this right; DARC fills the missing control layer. For a fuller survey, see Appendix A.2.

3 METHOD

The matched triplet in Figure 1 fixes the task, the request, and the response options while changing only which party holds final authority; the correct answer changes entirely with that assignment. We formalize the assignment in Section 3.1, define delegated-authority underuse (DAU), the failure to exercise an assigned right, in Section 3.2, and present DARC, an inference-time controller that makes the assigned right explicit before the model responds (Section 3.3).

3.1 TASK FORMULATION

A case bundles three pieces of information: fixed task content T, a dialogue history H, and a current request q. The history determines who holds decision authority for the current task, which we call the authority direction d. It takes one of three values: `equal`, `counterpart_higher`, and `model_higher`. A deterministic map ϕ converts d into a decision right $r = \phi(d)$, again one of three values: `joint`, `counterpart_final`, and `model_final`. The model selects or generates a response y that completes q and respects r.

Figure 1 shows what the three rights mean in practice. `joint` means the two parties confirm together, `counterpart_final` means the model submits the decision for approval, and `model_final` means it confirms and moves forward. Direction and decision right carry the same information through ϕ. We keep them as separate layers because the controller ablates them separately (Section 3.3). Options come in a four-way set: the three rights above, plus a neutral option that completes the task without marking who decides.

The difficulty lies in direction. In a `counterpart_higher` case, seeking the counterpart's approval is the correct move; in a `model_higher` case, asking the counterpart to approve the same decision is an authority error. The model must identify which party holds the final right for the current task, and detecting that authority exists is not enough. This is exactly the asymmetry behind delegated-authority underuse: models readily recognize authority held by the human counterpart while surrendering authority delegated to themselves.

3.2 DELEGATED-AUTHORITY UNDERUSE

Delegated-authority underuse occurs when the dialogue assigns the right `model_final` (the `model_higher` direction) but the model responds as if the right were `counterpart_final` (the `counterpart_higher` direction). We write $\hat{r}(y)$ for the right that a response actually expresses: in forced-choice mode this is read off the role of the selected option, and in open-ended mode it is assigned by human review (Appendix A.7). The error rate we measure is $\text{DAU} = \Pr[\hat{r}(y) = \texttt{counterpart_final} \mid r = \texttt{model_final}]$.

This output-based measure captures a concrete control error: placing final responsibility with the wrong party. This behavior is the model-level counterpart of the responsibility-shifting subtype of decision-avoidance (Section 2).

3.3 DELEGATED AUTHORITY RESPONSE CONTROLLER

DARC inserts an explicit authority step between reading the dialogue and producing the response. Given a raw case, the baseline answers directly. The full controller, M3, asks the model to do three things in one call: summarize the dialogue evidence, infer who holds final authority from that evidence, and turn that authority into a concrete decision-right policy before answering. We ablate by dropping later steps: M1 keeps only the evidence summary, M2 is M1 plus the authority

direction, and M3 is M2 plus the decision-right policy. Each variant is a complete, self-contained prompt; full templates appear in Appendix A.8.

M1: Evidence summary. M1 is the evidence-only variant: it asks the model to summarize the dialogue evidence and nothing else. It tests whether evidence extraction alone is enough for reliable authority use. If DAU persists here, surface-level evidence is not a strong enough control signal.

M2: Authority direction. M2 is M1 plus the authority direction. It tests whether naming who holds final authority, beyond summarizing facts, is the critical signal. A substantial drop in DAU relative to M1 would show that the direction, not the summary, drives the effect.

M3: Decision-right policy. M3 is M2 plus the decision-right policy. It tests whether compiling the identified authority into a concrete response policy further improves execution. For `joint`, the response invites both parties to formulate and confirm the decision together. For `counterpart_final`, it surfaces the relevant criteria and obtains the counterpart's final approval. For `model_final`, it applies the stated standard, confirms the outcome, and proceeds without unnecessary escalation. M3 is the complete controller: it makes authority use inspectable while keeping the final response public and task-focused.

4 DATASET: DELEGATEDAUTHORITYBENCH

To evaluate LLMs' ability to retain delegated decision authority in socially situated dialogue, we introduce DELEGATEDAUTHORITYBENCH, a diagnostic benchmark in English, Chinese, and French.

Dataset construction with RMT. The construction employs a matched-triplet design grounded in Relational Models Theory (RMT) (Fiske, 1992) to isolate the effect of dialogue-assigned authority: Equality Matching and directed Authority Ranking form the manipulated axis, while a no-schema baseline, Communal Sharing, and Market Pricing vary the task background. Cases were authored by the research team with GPT-5.6 drafting assistance, then passed through blind review (Krippendorff's α=0.72) and independent program validation, yielding 360 English cases (3 backgrounds $\times$ 3 directions, 40 each) plus 360 Chinese and 360 French parallel translations, for 1,080 cases across three languages. The 120 scenarios are drafted as free-form dialogue, and every model-higher delegation is worded differently, so correct responses require relational inference from each dialogue. The four-stage construction pipeline is detailed in Appendix A.4.

Responsibility directions. Each case fixes the model's role through one of three responsibility directions: under `equal` the two parties confirm together, under `counterpart_higher` the model submits for the human's final approval, and under `model_higher` it confirms and proceeds. Throughout, `model` denotes the evaluated system and `counterpart` the other party in the dialogue. The three-way split is what makes delegated-authority underuse observable, since only `model_higher` demands retaining rather than returning authority. The two background schemas, Communal Sharing and Market Pricing, change the task surface (shared goals versus budgets and contracts) without changing the answer rule. DAU is strongest in model-higher Market Pricing cases, where baseline accuracy is 18.0%; Section 5 analyzes this contextual amplification.

Matched-triplet isolation. A matched triplet keeps task facts, request, and options identical and changes only the responsibility-assigning history, so the answer changes only with who holds final authority. Option order is randomized across items. The model receives only the public case and must recover authority from dialogue evidence, with strict information isolation. Full matched-triplet mechanics, option authoring rules, and the quality control protocol are in Appendix A.3 and Appendix A.5; the file-level separation between the model input and evaluation packages is specified in Appendix A.6.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models and data. We evaluate 360 cases per language in English, French, and Chinese, for 1,080 cases in total. In each language, the cases form 120 matched triplets (Section 4). Within each triplet, task facts and response options remain fixed while only the dialogue-assigned responsibility direction varies. We evaluate five widely-used models in English and three in French and Chinese: four open-weight models (DeepSeek-V4-Flash (Xu et al., 2026), Nemotron-3.5-Lightning (NVIDIA, 2026), Kimi-K2.7-Code (Moonshot AI, 2026), MiniMax-M3 (MiniMax AI, 2026)) and one closed-weight model (Doubao-Seed-2.0-Mini (Seed, 2026)), with the language-by-model breakdown in Table 1. Of these, Kimi-K2.7-Code and MiniMax-M3 were evaluated with reasoning mode enabled, while reasoning was explicitly disabled for the remaining models, so the evaluated models span both reasoning modes.

Every case is tested under the baseline and three progressive controller variants M1–M3 across all three responsibility directions.

We also run an open-ended generation experiment on the same 360 English cases for three models (DeepSeek-V4-Flash, Nemotron-3.5-Lightning, Doubao-Seed-2.0-Mini), scored by five reviewers under majority vote. We additionally test three single-sentence authority instructions (T0–T2) on DeepSeek-V4-Flash and Doubao-Seed-2.0-Mini as an analysis supplement that separates prompt-strength effects from the structured control chain.

Evaluation metrics. The primary metric is four-option forced-choice accuracy (Zheng et al., 2024; Wang et al., 2024a;b): a response is correct when its selected option matches the case answer key. To characterize delegated-authority underuse beyond raw accuracy, we report three complementary metrics.

The **DAU rate** is the primary measure: among all model-higher cases, the proportion where the model incorrectly selects the counterpart-higher option, i.e., returns final responsibility to the human, $\text{DAU} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \mathbb{I}(\hat{y}_i = \text{counterpart-higher})$, where $\mathcal{A}$ is the set of model-higher cases. A higher rate indicates a stronger tendency to surrender delegated final decisions to the human.

The **error composition** answers whether failures are systematically responsibility-shifting: among model-higher cases that the model gets wrong, the proportion that are specifically counterpart-higher selections rather than equal or neutral ones.

The **directional gap** is the difference between counterpart-higher and model-higher accuracy, $\Delta_{\text{dir}} = \text{ACC}_{\text{counterpart-higher}} - \text{ACC}_{\text{model-higher}}$. A positive value means the model performs better when the human holds final authority. The DAU rate covers all model-higher cases, whereas error composition covers only model-higher errors, so their magnitudes are not directly comparable. The directional gap compares accuracy across the two directions in percentage points.

Method comparisons use McNemar tests on identical model-case instances (n=360 per model); full statistics are reported in Appendix A.7 (Tables 6, 7, and 8). For open-ended generation, each response is independently scored 1 (correct authority use) or 0 (incorrect) by five reviewers; a case is correct when at least three of five reviewers score it correct. All methods use identical task facts, current requests, and response options, so differences isolate the effect of explicit authority control.

5.2 RESULTS AND ANALYSIS

The results show a consistent pattern: models surrender delegated authority, and DARC makes authority direction explicit and largely repairs this failure across directions, backgrounds, languages, and response formats. We examine whether the failure is asymmetric, how effective the controller is, whether the repair transfers, and what changes along the control chain.

5.2.1 DELEGATED-AUTHORITY UNDERUSE IS SYSTEMATIC AND ASYMMETRIC

Baseline receives the complete public case and directly selects one of four responses without an explicit authority state. At first glance, baseline appears competent at reading authority in dialogue:

Table 1: Main results on English and multilingual transfer. All values are percentages. English evaluates five models on 360 cases, with the final row reporting the five-model average; French and Chinese evaluate three models each on the same 360 parallel cases. M3 is the best-performing controller variant. The model-higher condition exposes the core failure; M3 reaches 84.2–100 on model-higher across all model-language pairs.

Language	Model	Direction	Baseline	M1	M2	M3
English	DeepSeek-V4-Flash	Total	75.0	84.7	98.1	**98.3**
		Equal	85.0	97.5	**100.0**	**100.0**
		Counterpart-higher	98.3	98.3	**100.0**	**100.0**
		Model-higher	41.7	58.3	94.2	**95.0**
	Nemotron-3.5-Lightning	Total	79.7	83.9	99.2	**100.0**
		Equal	94.2	97.5	**100.0**	**100.0**
		Counterpart-higher	99.2	**100.0**	97.5	**100.0**
		Model-higher	45.8	54.2	**100.0**	**100.0**
	Doubao-Seed-2.0-Mini	Total	75.0	75.6	89.4	**98.9**
		Equal	**100.0**	99.2	**100.0**	**100.0**
		Counterpart-higher	**100.0**	**100.0**	**100.0**	**100.0**
		Model-higher	25.0	27.5	68.3	**96.7**
	Kimi-K2.7-Code	Total	71.9	96.7	99.4	**100.0**
		Equal	**100.0**	**100.0**	**100.0**	**100.0**
		Counterpart-higher	98.3	**100.0**	**100.0**	**100.0**
		Model-higher	17.5	90.0	98.3	**100.0**
	MiniMax-M3	Total	78.3	97.2	95.8	**99.2**
		Equal	99.2	**100.0**	**100.0**	**100.0**
		Counterpart-higher	99.2	**100.0**	**100.0**	**100.0**
		Model-higher	36.7	91.7	87.5	**97.5**
	Average	Total	76.0	87.6	96.4	**99.3**
		Equal	95.7	98.8	**100.0**	**100.0**
		Counterpart-higher	99.0	99.7	99.5	**100.0**
		Model-higher	33.3	64.3	89.7	**97.8**
French	Doubao-Seed-2.0-Mini	Total	60.0	72.2	93.6	**98.9**
		Equal	65.8	84.2	**100.0**	**100.0**
		Counterpart-higher	**100.0**	**100.0**	**100.0**	**100.0**
		Model-higher	14.2	32.5	80.8	**96.7**
	Kimi-K2.7-Code	Total	77.8	98.1	**100.0**	**100.0**
		Equal	98.3	**100.0**	**100.0**	**100.0**
		Counterpart-higher	97.5	96.7	**100.0**	**100.0**
		Model-higher	37.5	97.5	**100.0**	**100.0**
	MiniMax-M3	Total	73.9	93.9	96.1	**98.9**
		Equal	93.3	**100.0**	**100.0**	**100.0**
		Counterpart-higher	99.2	98.3	**100.0**	**100.0**
		Model-higher	29.2	83.3	88.3	**96.7**
Chinese	Doubao-Seed-2.0-Mini	Total	66.1	67.2	76.1	**96.1**
		Equal	96.7	97.5	**100.0**	**100.0**
		Counterpart-higher	**100.0**	**100.0**	**100.0**	**100.0**
		Model-higher	1.7	4.2	28.3	**88.3**
	Kimi-K2.7-Code	Total	71.4	93.3	97.2	**99.4**
		Equal	**100.0**	**100.0**	**100.0**	**100.0**
		Counterpart-higher	99.2	**100.0**	**100.0**	**100.0**
		Model-higher	15.0	80.0	91.7	**98.3**
	MiniMax-M3	Total	73.9	84.4	87.2	**94.7**
		Equal	**100.0**	**100.0**	**100.0**	**100.0**
		Counterpart-higher	**100.0**	**100.0**	**100.0**	**100.0**
		Model-higher	21.7	53.3	61.7	**84.2**

across the five English models it averages 76.0% overall and is nearly perfect when the counterpart holds final approval (99.0%) or responsibility is equal (95.7%). This competence is deceptive. On the one condition where the model itself must retain final decision authority, the model-higher cases, accuracy collapses to 33.3%. The DAU rate is 64.2% (385 of 600 model-higher cases): models surrender delegated authority in nearly two-thirds of cases where they should keep it. Table 1 reports accuracy by model, method, and responsibility direction.

Nearly all model-higher errors are responsibility-shifting: the error composition is 96.25%, with almost every error selecting the counterpart-higher option rather than equal or neutral choices. The directional gap reaches 65.7 percentage points, showing that models recognize human authority far better than they retain delegated authority. This asymmetry suggests that the failure is not a general inability to handle authority relations, since the same facts and dialogue structure produce near-perfect accuracy when the human holds final say. The pattern is consistent with a default prior that final decisions belong to humans, a prior that overrides the dialogue's delegation when the model itself is empowered. We confirm the comprehension-vs-execution distinction more directly below with the M2 identification analysis (Section 5.2.4).

5.2.2 Explicit authority control repairs the failure

Adding explicit authority control produces large, consistent gains. Averaged across the five English models, the full controller (M3) raises overall accuracy from 76.0% to 99.3%, with the largest gains concentrated in the model-higher direction. The controller variants repair this failure selectively. M1, which adds only a generic evidence summary, reaches 87.6% overall but remains at 64.3% on model-higher cases. M2, which makes authority direction explicit, reaches 96.4% overall and 89.7% on model-higher. M3, which compiles direction into a decision-right policy, reaches 99.3% overall and 97.8% on model-higher. All baseline-vs-M3 comparisons are significant at $p < 0.001$ (McNemar tests, $n = 360$ per model; Appendix Table 6). The monotonic improvement provides mechanism-level evidence that a dialogue system must represent not only relevant facts, but also who has the right to make the final decision.

5.2.3 Control transfers across languages and formats

Transfer is tested along two axes: language and response format.

Multilingual transfer. To test whether the control mechanism transfers beyond English, we evaluate M1–M3 on parallel French and Chinese translations of the same 360-case inventory. Each language uses three API models: Doubao-Seed-2.0-Mini, Kimi-K2.7-Code, and MiniMax-M3. Table 1 reports the results.

Across all six model-language pairs, baseline underuses delegated authority, while M3 reaches 94.7–100% overall and 84.2–100% on model-higher cases. Kimi-K2.7-Code reaches overall accuracy of 100% on French and 99.4% on Chinese. Residual errors in both languages concentrate in the model-higher direction, as in English. The larger M2–M3 gap for Doubao in Chinese suggests that the decision-right policy matters especially when authority cues are less explicit. Together, these results suggest that the controller acts on an abstract authority state across languages. The repair holds with reasoning mode enabled (Kimi-K2.7-Code, MiniMax-M3) and disabled (the remaining models).

Free-form human evaluation. The selection experiments measure whether a model chooses the correct option among fixed candidates. To test whether explicit authority control also improves open-ended generation, we run a free-form evaluation on the same 360-case inventory. Each model produces a natural-language response instead of selecting from options. Five reviewers independently score each response as 1 (correct authority use) or 0 (incorrect). A case is scored correct when at least three of five reviewers score it correct (Krippendorff's α=0.78).

We compare direct answering (baseline) against the complete controller (M3), which exposes evidence, authority direction, and decision rights before generating the response. Figure 2 reports the results, and Table 9 lists the per-model accuracy by responsibility direction.

The direction-level pattern mirrors the selection experiment. On equal cases, accuracy rises from 81.4% to 90.6%. On counterpart-higher cases, it rises from 82.5% to 89.7%. The largest gain

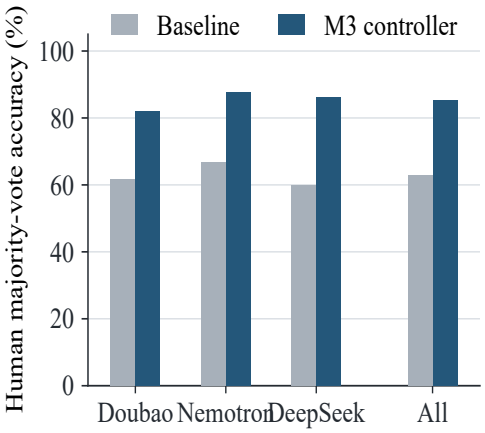


Figure 2: Free-form response accuracy by human majority vote. All values are percentages.

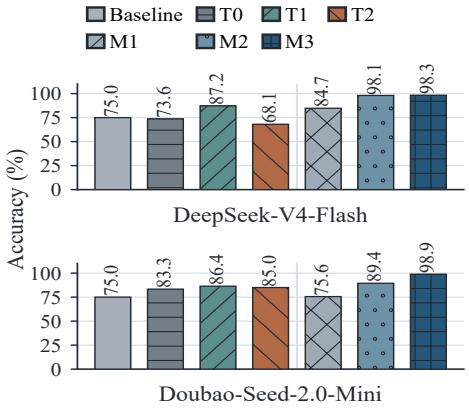


Figure 3: Overall accuracy of baseline, the three single-sentence instructions (T0–T2), and the DARC variants (M1–M3) on the same 360 English cases for the two models. All values are percentages.

appears on model-higher cases, where accuracy rises from 24.4% under the baseline to 75.6% under M3, a gain of 51.1 percentage points. This confirms that the delegated-authority underuse failure persists in open-ended generation, and that explicit authority control substantially repairs it even when responses are not constrained to fixed options. The three free-form models reach 99.1% on the selection task under M3 but only 85.3% in free-form generation under human evaluation, indicating that choosing the correct option and implementing it in natural language are distinct challenges.

5.2.4 Mechanism-level signatures of the repair

The analyses below distinguish the roles of evidence summary, authority direction, and decision-right policy in the repair.

Simple authority instructions reveal an identification-execution gap. We test three one-sentence authority instructions on the same 360 English cases with DeepSeek-V4-Flash and Doubao-Seed-2.0-Mini (Figure 3; details in Appendix A.7): T0 asks the model to identify who has final say and act accordingly, T1 grants the model final approval, and T2 denies it. These instructions raise accuracy only partially, by cueing attention to authority. To locate the failure, we turn to the five-model controller ablations: M2 states the authority direction correctly in 99.2% of cases, yet 63 of its 65 errors occur on correctly identified cases; the model reads delegated authority from dialogue but defaults to not exercising it. On model-higher cases specifically, direction identification reaches 100% while final execution reaches only 89.7%. Only an explicit decision-right policy binding the identified direction to action (M3) approaches 100% accuracy (Table 5).

Market Pricing amplifies the shift toward human authority. Within model-higher cases, baseline accuracy drops to 18.0% in Market Pricing contexts (budget, contract, procurement), holds at 42.0% in Communal Sharing contexts, and reaches 40.0% without a background schema (Figure 4). The same cases remain accurate on counterpart-higher and equal directions, so the context does not degrade performance globally; it strengthens the tendency to default to human authority. We hypothesize that budgets, contracts, and procurement are domains where humans almost always hold the final decision right in training data: financial approvals and contract sign-offs are stereotypically human acts, making the human-authority prior most active precisely when the dialogue delegates the opposite. The single-sentence T2 instruction shows the same gradient on Doubao: it gets 27 of 40 model-higher cases correct with no extra schema but only 15 of 40 in Market Pricing contexts (Appendix A.7). Under M3 these same cases reach 98.5%, showing that explicit decision-right representation resists this contextual interference.

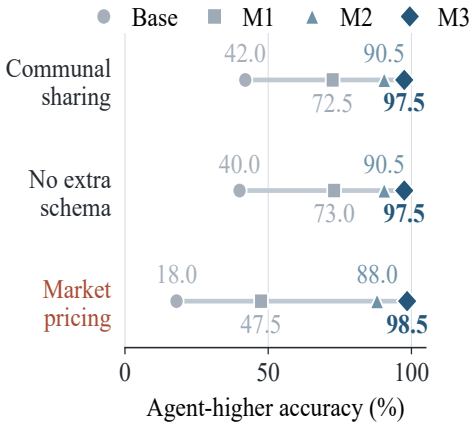


Figure 4: Model-higher accuracy by background condition, averaged across five English models. All values are percentages.

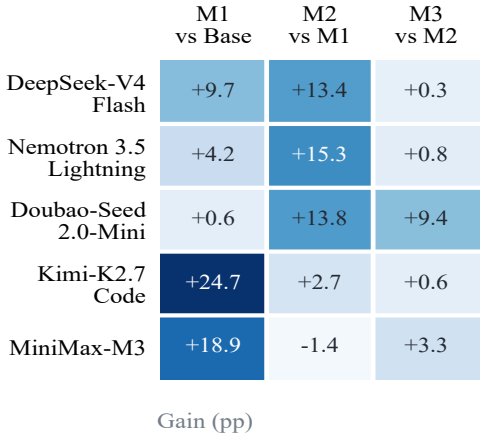


Figure 5: Model heterogeneity in controller gains. All gains are in percentage points of overall accuracy, relative to the previous method in the chain.

The repair targets the failure cluster precisely. On 1,800 paired model-case instances, M3 repairs 422 baseline errors, of which 390 come from the model-higher direction. It introduces only 3 new errors on instances baseline answered correctly, and both methods fail on 10. The controller therefore reverses the dominant failure mode without broadly altering correct behavior.

Consistent gains through the control chain. Of the 15 model-stage gains in Figure 5, 14 are positive, and the only marginally negative entry (MiniMax, M2 vs M1, -1.4 pp) is well within chance variation. This consistency indicates the repair acts on a shared failure mode rather than model-specific quirks.

Direction is the critical signal; the decision-right policy makes it inspectable. Taken together, these results delineate a clear mechanism. The controller targets delegated-authority underuse, a shared failure mode across models, and reverses it without broadly altering correct behavior. The M1–M3 ablation decomposes the repair into three components: evidence, direction, and a decision-right policy. Direction is the critical signal. A generic evidence summary helps but falls short on model-higher cases. Adding the authority direction closes most of the remaining gap. The bottleneck is not reading the direction, but acting on it. This identification-execution gap is why single-sentence authority instructions only help partially. The decision-right policy closes the gap and makes the repair inspectable.

6 Conclusion

Using a matched-triplet benchmark grounded in Relational Models Theory, we identify delegated-authority underuse (DAU): LLMs often return final decision authority to the human even when the dialogue delegates it to them. Across five English models, DARC makes the assigned authority direction explicit and raises model-higher accuracy from 33.3% under the baseline to 97.8%. The repair also extends to human-evaluated free-form responses. These findings indicate that an explicit authority state is a necessary control layer for reliable execution of delegated decision rights alongside persona, memory, and social relation modeling.

Designers of LLM-based assistants and multi-agent systems should take this pattern seriously. It appears even in the newest generation of production-grade LLMs and is consistent across five independently developed model families. A model that appears to decide yet hands the decision back can quietly stall a collaborative loop or cause confusion. Such systems warrant monitoring and control in deployment. Extending the benchmark to settings with dynamic authority, in which the right to decide shifts across turns and participants, is a natural next step toward deployment.

AI USE STATEMENT

Generative AI tools were used to assist with dataset construction and translation, polish the manuscript language, and support code development. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

Human Subjects for Data Evaluation. We recruited student reviewers with psychology, social-science, or NLP backgrounds: three reviewed the constructed cases and five evaluated the free-form responses. Prior to the task, reviewers were fully informed about the study's scope and provided informed consent. We ensured that no personally identifiable information (PII) was collected. Reviewers were compensated at an hourly rate of $12, exceeding the local minimum and ensuring fair payment for their work.

REPRODUCIBILITY STATEMENT

All code, benchmark data, and experimental artifacts are publicly released at `https://anonymous.4open.science/r/7f3c2a621A/`.

REFERENCES

Christopher J Anderson. The psychology of doing nothing: forms of decision avoidance result from reason and emotion. *Psychological bulletin*, 129(1):139, 2003.

Timothy Bickmore and Justine Cassell. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 396–403, 2001.

Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth F. Churchill (eds.). *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000. ISBN 9780262032780. URL `https://mitpress.mit.edu/9780262032780/embodied-conversational-agents/`.

Kushal Chawla, Weiyan Shi, Jingwen Zhang, Gale Lucas, Zhou Yu, and Jonathan Gratch. Social influence dialogue systems: A survey of datasets and models for social influence tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 750–766, 2023.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. Socialbench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2108–2126, 2024a.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. 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, 2024b.

Myra Cheng, Robert D Hawkins, and Dan Jurafsky. Accommodation and epistemic vigilance: A pragmatic account of why llms fail to challenge harmful beliefs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16181–16203, 2026a.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. Elephant: Measuring and understanding social sycophancy in llms. In *International conference on learning representations*, volume 2026, pp. 130060–130097, 2026b.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 11370–11403, 2023.

Yuxuan Du. Who chooses how preferences are aggregated? auditing aggregation-rule authority in llm-based group recommendation. *arXiv preprint arXiv:2608.23966*, 2026.

Alan P Fiske. The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological review*, 99(4):689, 1992.

Ivar Frisch and Mario Giulianelli. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pp. 102–111, 2024.

Garry A Gabison and R Patrick Xian. Inherent and emergent liability issues in llm-based agentic systems: a principal-agent perspective. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pp. 109–130, 2025.

Jonathan Gratch, Jeff Rickel, Elisabeth André, Justine Cassell, Eric Petajan, and Norman Badler. Creating interactive virtual humans: Some assembly required. *IEEE Intelligent systems*, 17(4): 54–63, 2002.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, Tat-Seng Chua, and Jimmy Xiangji Huang. How to enable effective cooperation between humans and nlp models: A survey of principles, formalizations, and beyond. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 466–488, 2025.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5906–5919, 2023.

Jiarui Ji, Zehua Zhang, Zhewei Wei, Bin Tong, Guan Wang, and Bo Zheng. Graphia: Harnessing social graph data to enhance llm-based social simulation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7103–7128, 2026.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the association for computational linguistics: NAACL 2024*, pp. 3605–3627, 2024.

Emil Joswin, Srujananjali Medicherla, and Priyanka Mary Mammen. A mechanistic view of authority hierarchy in llm sycophancy. *arXiv preprint arXiv:2607.00415*, 2026.

Eunsu Kim, Junyeong Park, Juhyun Oh, Kiwoong Park, Seyoung Song, A Seza Doğruöz, Alice Oh, and Najoung Kim. Are they lovers or friends? evaluating llms' social reasoning in english and korean dialogues. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23431–23451, 2026.

Changgeon Ko, Jisu Shin, Hoyun Song, Huije Lee, Eui Jun Hwang, and Jong C Park. Social dynamics as critical vulnerabilities that undermine objective decision-making in llm collectives. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37865–37890, 2026.

Huayi Lai, Shichao Song, Simin Niu, Hanyu Wang, Jiawei Yang, Zhouxing Wang, Zhiqiang Yin, and Xun Liang. Rolecde: Benchmarking and mitigating role–alignment trade-offs in role-playing agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2226–2248, 2026.

Cao Liu, Kang Liu, Shizhu He, Zaiqing Nie, and Jun Zhao. Incorporating interlocutor-aware context into response generation on multi-party chatbots. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 718–727, 2019.

Hongyuan Lu, Wai Lam, Hong Cheng, and Helen Meng. Partner personas generation for dialogue response generation. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 5200–5212, 2022.

Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Xing Sun, et al. Rolemrc: A fine-grained composite benchmark for role-playing and instruction-following. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21008–21030, 2025.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8679–8696, 2025.

Priyanka Mary Mammen, Emil Joswin, and Shankar Venkitachalam. Who endorsed it? measuring authority bias across expertise levels in language models. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 980–989, 2026.

MiniMax AI. MiniMax-M3 model card. Hugging Face model card, 2026. URL `https://huggingface.co/MiniMaxAI/MiniMax-M3`.

Moonshot AI. Kimi K2.7 Code model card. Hugging Face model card, 2026. URL `https://huggingface.co/moonshotai/Kimi-K2.7-Code`.

Ansong Ni, Ruta Desai, Yang Li, Xinjie Lei, Dong Wang, Jiemin Zhang, Jane Yu, Ramya Raghavendra, Gargi Ghosh, Shang-Wen Li, et al. Collaborative reasoner: Self-improving social agents with synthetic conversations. volume 38, pp. 23768–23796, 2026.

NVIDIA. Nemotron 3.5 Lightning model documentation. NVIDIA Nemotron documentation, 2026. URL `https://www.nvidia.com/en-us/ai/nemotron/`.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

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

Liang Qiu, Yuan Liang, Yizhou Zhao, Pan Lu, Baolin Peng, Zhou Yu, Ying Nian Wu, and Song-Chun Zhu. Socaog: Incremental graph parsing for social relation inference in dialogues. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 658–670, 2021.

Parisa Rabbani, Priyam Sahoo, Ruben Mathew, Aishee Mondal, Harshita Ketharaman, Nimet Beyza Bozdag, and Dilek Hakkani-Tur. Dialdefer: A framework for detecting and mitigating llm dialogic deference. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44630–44671, 2026.

Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity, 2026.

Ashish Sharma, Sudha Rao, Chris Brockett, Akanksha Malhotra, Nebojsa Jojic, and William B Dolan. Investigating agency of llms in human-ai collaboration tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1968–1987, 2024a.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024b.

Haochun Wang, Sendong Zhao, Jingbo Wang, Zewen Qiang, Bing Qin, and Ting Liu. Beyond frameworks: Unpacking collaboration strategies in multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21361–21375, 2025.

Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33566–33574, 2026.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. 2024a.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416, 2024b.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. Dialogue-based relation extraction. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4927–4940, 2020.

Michael Zakharin and Timothy C Bates. Relational models theory: Validation and replication for four fundamental relationships. *Plos one*, 18(6):e0287391, 2023.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for llm agents: A social psychology view. 2023.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, 2018.

Wenyuan Zhang, Tianyun Liu, Mengxiao Song, Xiaodong Li, and Tingwen Liu. Sotopia-■: Dynamic strategy injection learning and social instruction following evaluation for social agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24669–24697, 2025.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pp. 19426–19454, 2024.

Jingting Zheng, Yuqi Ren, Linhao Yu, Yongqi Leng, and Deyi Xiong. Beyond value benchmarks: Measuring value-structure alignment in large language models via symmetric q-sorts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18199–18230, 2026.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, et al. Socialeval: Evaluating social intelligence of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30958–31012, 2025.

A Appendix

A.1 Limitations and future work

This study identifies delegated-authority underuse as a robust behavioral failure and shows that explicit authority control can substantially reduce it in forced-choice settings.

First, the findings are established on single-turn, two-party cases. Real deployment involves multi-turn dialogue, multiple participants, and authority that shifts mid-conversation. The matched-triplet design intentionally isolates the effect of dialogue-assigned authority by holding all other factors constant, and extending the study to dynamic authority states across longer interactions is a natural next step.

Second, the intervention operates at inference time. The surrender of delegated authority may reflect human-authority priors acquired during pretraining or alignment, though safety-conditioned caution offers a complementary explanation. Our behavioral data do not distinguish these accounts, and we leave their separation to future work. Whether such a default can be reduced during pretraining or fine-tuning, through targeted data or alignment objectives, remains open. A training-stage fix would complement the inference-time controller and may address the bias at its source.

Third, the benchmark scenarios focus on everyday collaborative tasks such as community event planning and work coordination. High-stakes domains, including medical, legal, and financial decision-making, may involve different authority norms and risk profiles, and extending the benchmark to these settings is an important direction for future work.

A.2 Theoretical foundations and related work

This section develops the theoretical and related-work foundations summarized in Section 2: decision-avoidance theory and its model-level analogue DAU, the four RMT schemas that ground the benchmark design, and a thematically organized survey of socially situated dialogue and authority-related work.

Decision-avoidance and delegated-authority underuse. Decision-avoidance describes actors who possess decision authority yet evade exercising it (Anderson, 2003); its responsibility-shifting subtype transfers decision responsibility to another party even when formally authorized. Our benchmark reveals LLM outputs resembling this subtype, yet an LLM has no subjective regret or fear of blame. We therefore define the model-level analogue, delegated-authority underuse (DAU): the systematic tendency of a model to surrender delegated final-decision authority back to the human even when dialogue evidence assigns that authority to it.

Relational Models Theory. Relational Models Theory (RMT) characterizes human social life through four elementary relational schemas (Fiske, 1992; Zakharin & Bates, 2023). Two govern authority allocation and anchor our design: Authority Ranking, where one party holds decision-making priority, and Equality Matching, where peers interact symmetrically. Constructing dialogue histories that assign final authority to the model, to the human counterpart, or jointly to both realizes the three responsibility directions of Section 4. The remaining schemas, Communal Sharing and Market Pricing, supply task context only, never the answer rule (Appendix A.3).

Related work on socially situated dialogue. Research on socially situated dialogue includes embodied agents that model social presence (Cassell et al., 2000; Gratch et al., 2002; Chawla et al., 2023) and relational agents that build user trust (Bickmore & Cassell, 2001). Generative agents incorporate personas, memories, and social behaviors (Park et al., 2023; Packer et al., 2023; Jiang et al., 2024; Frisch & Giulianelli, 2024). Benchmarks examine role-playing consistency with assigned identities and roles (Chen et al., 2024a; Lu et al., 2025; Zhou et al., 2025) and multi-agent collaboration under social-psychology frameworks (Zhang et al., 2023; Wang et al., 2025; Ni et al., 2026; Zhang et al., 2025). Another line recovers social information from dialogue, including persona personalization, relation extraction, social reasoning, social graph simulation, and theory of mind (Zhang et al., 2018; Lu et al., 2022; Liu et al., 2019; Yu et al., 2020; Qiu et al., 2021; Kim et al., 2026; Choi et al., 2023; Ma et al., 2025; Ji et al., 2026; Chen et al., 2024b). These methods output a persona, relation category, or graph, but none determines who holds final approval on the current task.

Related work on authority and social pressure. The closest line studies how LLMs respond to authority and social pressure. One strand documents the phenomenon: sycophancy and social alignment lead models to defer to human speakers (Sharma et al., 2024b; Cheng et al., 2026b; Hwang et al., 2023), a pattern that can be explained by pragmatic accommodation (Cheng et al., 2026a). A second strand traces its sources: speaker framing in dialogic deference (Rabbani et al., 2026), perceived endorser expertise in authority bias (Mammen et al., 2026; Joswin et al., 2026; Wang et al., 2026), and value-structure alignment (Zheng et al., 2026) can override objective decision-making (Ko et al., 2026). A third strand tests its boundaries: explicit delegation is studied only in a voting task (Du, 2026), while Role-Value Decoupling (Lai et al., 2026) and Agency (Sharma et al.,

2024a) do not concern the final decision right delegated to the model. None tests whether a model retains and exercises authority explicitly delegated to it in task-oriented dialogue.

Gap and relation to our method. This body of work establishes that LLMs are highly sensitive to authority, yet none tests whether a model retains and exercises authority explicitly delegated to it in task-oriented dialogue. DAU occupies that gap; to our knowledge, our study is the first controlled investigation of the delegated final decision right. What remains missing is a control layer converting relationship evidence into task-specific decision rights and a response policy. DARC fills it, making decision authority explicit and inspectable at inference time.

A.3 Benchmark construction details

This section supplies the construction details summarized in Section 4: it operationalizes the four RMT schemas, defines the nine fine-grained conditions, and specifies the matched-triplet mechanics, option authoring rules, and trilingual parallel versions.

Operationalizing the four RMT schemas. RMT is used as a design device, not as a model input or prediction target. Table 2 states how each schema is operationalized. The manipulated axis contrasts Equality Matching with directed Authority Ranking. Communal Sharing and Market Pricing appear only as background texture: a shared goal or a written budget does not by itself determine who holds final approval.

Table 2: Operationalization of the four Relational Models Theory schemas in the benchmark.

Schema	Operational meaning	Typical dialogue evidence	Role in benchmark
Communal Sharing (CS)	A goal, resource, or duty is treated as shared	shared ownership, common goals, joint use	Background only
Authority Ranking (AR)	One party holds legitimate, directed decision priority	appointment, delegation, approval duty, final-responsibility statement	Manipulated axis, both directions
Equality Matching (EM)	Peers hold equal decision rights	co-appointment, equal votes, joint confirmation	Manipulated axis, equal condition
Market Pricing (MP)	Interaction organized by prices, ratios, rates, budgets, or contracts	budgets, rate schedules, cost shares, contract scope	Background only

Nine fine-grained conditions. Each item records a primary schema (EM or AR) and at most one supporting schema (CS or MP). Crossing three background configurations (no supporting schema, Communal Sharing, Market Pricing) with the three authority directions produces nine conditions, each containing exactly 40 cases (Table 3). The three directions remain perfectly balanced at 120 cases each.

Scene families. The 360 cases are organized into three scene families by background type, each with 40 matched triplets and 120 cases: direct EM/AR, Communal-Sharing background, and Market-Pricing background. All statistics use the complete merged inventory.

Matched-triplet mechanics. A public case is $X = (H_0, H_1, P_1)$. H_0 fixes the participants, task, and fixed facts such as deadlines, quantities, and existing constraints. H_1 carries the relational evidence as natural events: a formal appointment, a joint designation, a delegated approval duty, or an assigned budget responsibility. The relational evidence states authority through naturalistic appointment and approval events, without machine-readable labels, and is typically confirmed by both speakers. P_1 contains the current scene, the counterpart's new utterance, and four response

Table 3: The nine fine-grained construction conditions, 40 cases each.

Condition	Primary schema	Supporting	Direction	Decision right	Cases
EM	Equality Matching	none	equal	joint	40
AR_UP	Authority Ranking	none	counterpart-higher	counterpart_final	40
AR_DOWN	Authority Ranking	none	model-higher	model_final	40
CS_EM	Equality Matching	Communal Sharing	equal	joint	40
CS_AR_UP	Authority Ranking	Communal Sharing	counterpart-higher	counterpart_final	40
CS_AR_DOWN	Authority Ranking	Communal Sharing	model-higher	model_final	40
EM_MP	Equality Matching	Market Pricing	equal	joint	40
AR_MP_UP	Authority Ranking	Market Pricing	counterpart-higher	counterpart_final	40
AR_MP_DOWN	Authority Ranking	Market Pricing	model-higher	model_final	40
Total					360

options, and is verbatim identical across the triplet. A model that ignores H_1 therefore cannot distinguish the three conditions.

Option authoring rules. All four options complete the same core action and keep time, quantity, object, promise scope, and politeness level aligned; they differ only in who confirms and proceeds. The two directed options differ in length by at most four English words, so length cannot leak the answer. No option refuses the task, adds a factual error, or signals the answer through extreme politeness or sentiment, and the neutral option completes the task with no decision-right marker. Correct-option positions are randomized across items, giving a distribution of 89/89/91/91 across A–D.

Surface-form diversity. The 120 scenarios span distinct task domains, including agenda revision, survey summarization, slide review, exhibition-label editing, and newsletter layout. Each case carries one authority-assignment statement, giving 120 statements per direction; all 120 are textually unique in both the model-higher and the counterpart-higher direction. Delegation verbs (appointed, assigned, designated), role nouns (coordinator, editor, presenter, procedure lead), and permission phrasings (final approval, final responsibility, final layout approval) all vary across scenarios, as do the downstream division-of-labor sentences. The four options are length-balanced across positions (mean 19.2–19.5 words).

Trilingual parallel versions. The English cases are the source for the trilingual versions. The Chinese and French versions translate only the dialogue, probe utterance, and option text, and append _c or _f to every ID. Translation was performed with GPT-5.6 and manually verified for fidelity; no ID, structure, label, or answer key changes, and every English ID has exactly one Chinese and one French counterpart. The structural validator was rerun on each language. The forced-choice experiments in this paper include English, French, and Chinese; the open-ended human evaluation runs on English.

A.4 Construction pipeline

The benchmark follows a four-stage construction pipeline: theory-constrained design, model-assisted drafting, human review, and program validation. The review and validation protocols are detailed in Appendix A.5.

First, the research team fixes the RMT schemas, authority directions, decision rights, and response-option roles. Scene skeletons are designed to specify the participants, task facts, relational events, current probe, and the response strategies that must be distinguished.

Second, the research team drafts the English dialogue histories, current probes, four response options, and reference labels, with GPT-5.6 assisting in text production. Generation is constrained by the theoretical design so that each scenario conforms to the intended relational structure.

Third, trained reviewers inspect each generated case. They check whether the relational evidence supports the labeled direction, whether task facts are consistent within a triplet, whether the four

options are parallel, and whether the answer requires the relational evidence to determine. The review protocol and pass criteria are detailed in Appendix A.5.

Fourth, an independent validator (separate from the generation code) checks every release for structural completeness and consistency. The validator is described in detail in the quality control protocol below.

A.5 Quality control protocol

Section 4 summarizes the quality controls; this section details them: the six-dimension blind human review with first-round pass rates, and the independent program validation checks.

Blind human review. Three paid student reviewers with psychology, social-science, or NLP backgrounds were trained on practice items excluded from formal statistics. Reviewers did not see the drafting model, the controller methods, the answer keys, or any model output. Each case was independently reviewed by all three reviewers across six binary dimensions, listed in Table 4. The pass rates in Table 4 are first-round submission rates; cases that failed any dimension were rebuilt and re-reviewed until all six dimensions passed, so every released case passes full review. Agreement across reviewers reached Krippendorff's $\alpha = 0.72$.

Table 4: Six blind-review dimensions and their first-round pass rates. Released cases pass all dimensions after revision.

#	Review dimension	Judgment	Pass rate
1	Evidence sufficiency	H_1 events or utterances support the labeled direction	91%
2	RMT label correctness	Primary and supporting schemas and direction are labeled correctly given H_0 and H_1	85%
3	Answer uniqueness	Exactly one option fits the relationship and the task	92%
4	Option parallelism	All options complete one core action and differ only in authority strategy	86%
5	History dependence	The correct option cannot be fixed reliably without H_1	93%
6	Surface-cue check	No length, position, formulaic wording, politeness, or sentiment cue leaks the answer	97%

Program validation. An independent validator, separate from the generation code, checks every release: (i) item counts, unique IDs, and the 120-triplet structure; (ii) H_0, P_1, and the four options are verbatim identical within a triplet while the three H_1 histories differ; (iii) the three answer keys map to equal, counterpart-higher, and model-higher; (iv) public files contain no labels, keys, or scoring fields; (v) option roles are complete and the two directed options differ by at most four words; (vi) the A–D key distribution is 89/89/91/91 and manifests, case lists, and key paths agree. The same checks pass for the Chinese and French releases.

A.6 Information isolation and file structure

This section specifies how the benchmark separates the model input package from the evaluation package, so that authority is recoverable only from public dialogue evidence.

The benchmark enforces a strict separation between the model input package and the evaluation package.

Model input package. The only files provided to the model under test are the public case JSON files. Each public case contains the dialogue history (H_0 and H_1), the current probe (P_1), and the four response options. The public JSON fields are: `schema_version` (a release identifier), `dataset_id`, `title`, `language`, `model`, `counterpart`, `task_context` (the fixed facts H_0), `trajectory` (the H_0 and H_1 dialogue stages), and `probes` (the current probe P_1 with options). No relation labels, direction tags, option roles, answer keys, or scoring rules appear in the public case.

Evaluation package. The evaluation package contains `manifest.json`, `cases.json`, and the `evaluation/` directory. The `manifest.json` records the version, classes, matched sets, sample paths, and condition identifiers. The `cases.json` lists case identifiers, sample paths, and answer-key paths for the batch runner. The `evaluation/` directory contains `annotations.json` and per-sample answer keys. These files are read only by the evaluation program and are never provided to the model under test. The model must recover authority solely from the dialogue evidence in the public case.

A.7 Experimental details

Section 5 states the experimental protocol; this section gives the full details: model access, method-variant definitions, the human-evaluation procedure, the statistical tests with complete tables, and the model-level heterogeneity analysis.

Models. All models are accessed through provider-compatible OpenAI-format APIs. English selection uses DeepSeek-V4-Flash, Nemotron-3.5-Lightning, Doubao-Seed-2.0-Mini, Kimi-K2.7-Code, and MiniMax-M3; multilingual transfer uses Doubao-Seed-2.0-Mini, Kimi-K2.7-Code, and MiniMax-M3. Failed requests are retried with exponential backoff up to five times.

Method variants. Baseline receives the full public case context and selects one option directly. M1 adds a structured evidence summary. M2 adds an explicit authority direction field (equal, counterpart_higher, or model_higher). M3 further adds a decision-rights field and validates consistency between direction and decision rights before the final choice. All variants are single-call with no multi-turn chain-of-thought.

Human evaluation. Five paid student reviewers with psychology, social-science, or NLP backgrounds independently score each free-form response as 1 (correct authority use) or 0 (incorrect). A case is correct when at least three of five reviewers score it correct (Krippendorff's $\alpha = 0.78$). Reviewers receive case context and the dialogue-assigned authority direction, but not the method identity.

Statistical tests. All comparisons use McNemar tests on identical model-case instances (n=360 per model). Significance is at $p<0.05$. Binary variance is the 0/1 sample variance with denominator $n - 1$ over the 360 cases. Table 6 reports the baseline vs. M3 McNemar tests for every model-language pair including the human-evaluated free-form responses; Tables 7 and 8 report the corresponding 0/1 variances.

Identification agreement. The `status_direction` field is required only under M2 and M3. Agreement is computed per case by comparing the model's stated direction with the gold direction. Across the 1,800 English cases, M2's stated direction matches the gold label in 99.2% of cases (1786/1800) and M3's in 100%; per-model agreement appears in Table 5. The direction of the final answer lags the stated direction: M2 answers in the gold direction in 96.4% of cases (1735/1800) and M3 in 99.3% (1787/1800).

On the model-higher direction the gap is largest: identification is 100% while answers reach only 89.7% under M2, closing to 97.8% under M3. Under M2, 63 of the 65 residual choice errors occur on cases where the direction was identified correctly, 62 of them on the model-higher direction; only 2 errors coincide with an incorrect direction statement, both on Nemotron's counterpart-higher cases, and 12 of Nemotron's 14 direction misstatements still produced correct choices. Under M3, all 1,800 direction statements agree with the gold labels, and all 13 residual errors occur on correctly identified cases.

Identification is therefore reliable under both controller variants; the residual gap lies in acting on the identified direction.

Model-level heterogeneity. The M2-to-M3 gain varies by model, revealing three distinct control patterns (Figure 5; Table 1). **Saturation after direction.** Nemotron-3.5-Lightning and Kimi-K2.7-Code approach ceiling once direction is explicit: both exceed 98 on model-higher under M2 (100.0

Table 5: Identification versus execution under M2 and M3: agreement between the stated `status_direction` and the gold authority direction (Stated), and between the direction reflected in the final answer and the gold direction (Answered), in percent. Each model row covers 360 English cases per method; the All row covers 1,800; the Model-higher row covers the 600 model-higher cases.

Model	M2		M3	
	Stated	Answered	Stated	Answered
DeepSeek-V4-Flash	100.0	98.1	100.0	98.3
Nemotron-3.5-Lightning	96.1	99.2	100.0	100.0
Doubao-Seed-2.0-Mini	100.0	89.4	100.0	98.9
Kimi-K2.7-Code	100.0	99.4	100.0	100.0
MiniMax-M3	100.0	95.8	100.0	99.2
All five models	99.2	96.4	100.0	99.3
Model-higher only	100.0	89.7	100.0	97.8

Table 6: McNemar tests of Baseline vs. M3 on identical model-case instances (n=360 per model). b is the number of cases correct under Baseline but wrong under M3; c is the reverse. Δ is the net gain $(c-b)/n$ in accuracy percentage points. χ^2 uses continuity correction; p is the exact two-sided binomial value. Human-evaluated rows use the free-form 0/1 scores.

Language	Model	b	c	Δ (pp)	χ^2	p
English	DeepSeek-V4-Flash	3	87	23.33	76.5444	<0.001
	Doubao-Seed-2.0-Mini	0	86	23.89	84.0116	<0.001
	Nemotron-3.5-Lightning	0	73	20.28	71.0137	<0.001
	Kimi-K2.7-Code	0	101	28.06	99.0099	<0.001
	MiniMax-M3	0	75	20.83	73.0133	<0.001
French	Doubao-Seed-2.0-Mini	0	140	38.89	138.0071	<0.001
	Kimi-K2.7-Code	0	80	22.22	78.0125	<0.001
	MiniMax-M3	0	90	25.00	88.0111	<0.001
Chinese	Doubao-Seed-2.0-Mini	0	108	30.00	106.0093	<0.001
	Kimi-K2.7-Code	2	103	28.06	95.2381	<0.001
	MiniMax-M3	3	78	20.83	67.6049	<0.001
Human evaluation (free-form, English)						
	Doubao-Seed-2.0-Mini	28	101	20.28	40.1860	<0.001
	Nemotron-3.5-Lightning	17	93	21.11	51.1364	<0.001
	DeepSeek-V4-Flash	28	122	26.11	57.6600	<0.001

and 98.3), with direction alone sufficing for them. **Policy-dependent translation.** Doubao-Seed-2.0-Mini requires the full decision-right policy to translate direction into action: its M2-to-M3 gain is the largest (+9.4 percentage points) and its model-higher accuracy rises from 68.3 to 96.7 only when M3 adds the policy layer. **Unstable control chain.** MiniMax-M3 shows M2 dipping below M1 (model-higher 91.7 to 87.5) before M3 recovers to 97.5, indicating that policy-level constraints stabilize its behavior. Across all 11 model-language pairs, the baseline-vs-M3 repair is significant (Table 6).

Simple authority instruction experiments (T0–T2). To test whether DARC's gains reduce to prompt strength, we run three single-sentence instruction variants on the same 360 English cases with the two models (templates in Appendix A.8). T0 asks the model to identify who has final say and act accordingly ("Determine from the conversation who has the final say for the current task, then act accordingly"); T1 grants the model final approval ("you have final approval for the current task; make the final decision and do not defer it to or seek approval from the other party"); T2 negates it ("you do not have absolute decision authority for the current task"). All variants share the task facts, request, and options with baseline and differ only in this one system instruction. Table 10 reports per-model accuracy by responsibility direction.

Table 7: Binary (0/1) variance of per-case correctness for the forced-choice experiments, by language and model, mirroring Table 1. Values are sample variances with denominator $n - 1$ over the 360 cases per model-language pair.

Language	Model	Baseline	M1	M2	M3
English	DeepSeek-V4-Flash	0.1880	0.1298	0.0191	0.0164
	Doubao-Seed-2.0-Mini	0.1880	0.1852	0.0947	0.0110
	Nemotron-3.5-Lightning	0.1621	0.1355	0.0083	0.0000
	Kimi-K2.7-Code	0.2024	0.0323	0.0055	0.0000
	MiniMax-M3	0.1702	0.0271	0.0400	0.0083
French	Doubao-Seed-2.0-Mini	0.2407	0.2012	0.0600	0.0110
	Kimi-K2.7-Code	0.1733	0.0191	0.0000	0.0000
	MiniMax-M3	0.1935	0.0575	0.0375	0.0110
Chinese	Doubao-Seed-2.0-Mini	0.2247	0.2210	0.1823	0.0375
	Kimi-K2.7-Code	0.2048	0.0624	0.0271	0.0055
	MiniMax-M3	0.1935	0.1317	0.1118	0.0501

Table 8: Binary (0/1) variance of the human-evaluated free-form correctness, by model. Values are sample variances with denominator $n - 1$ over the 360 scored responses per model.

Model	Baseline	M3
Doubao-Seed-2.0-Mini	0.2370	0.1484
Nemotron-3.5-Lightning	0.2228	0.1076
DeepSeek-V4-Flash	0.2407	0.1199

T0 is the neutral single-sentence control: it asks for identification and action without the structured fields of M2/M3. It raises average overall accuracy only from 75.0 to 78.5, far below the 93.8 that M2 reaches on the same two models, and its per-model effect is unstable: DeepSeek slightly drops below its baseline while Doubao gains 8.3 points. A single sentence can remind the model to attend to authority, but it cannot turn that identification into reliable action.

T1 shows that baseline failure is not an artifact of weak prompting. It lifts average overall accuracy from 75.0 to 86.8 and model-higher accuracy from 33.3 to 69.6, recovering over half of the failure gap. Yet T1 stays far below DARC, whose model-higher accuracy reaches 95.9 on the same two models under M3, and its repair is coarse: three cases that should be submitted for approval are instead decided by the model, as a global authority instruction raises the tendency to decide without allocating decision rights by evidence. Measured by our three custom metrics, T1 cuts the two-model DAU rate from 65.0% to 26.7% and narrows the directional gap from 65.8 to 28.3 percentage points, yet 87.7% of its remaining errors are still responsibility-shifting.

T2 shows that the direction of the instruction, not its strength, drives the effect for most models. Reversing the same single sentence collapses DeepSeek to 68.1 overall, below its baseline of 75.0, and even its equal cases drift toward returning authority, showing that a strong instruction can override dialogue evidence.

Doubao responds differently: its equal and counterpart-higher cases each move by one case while model-higher accuracy rises from 25.0 to 56.7, and its remaining 52 errors are still mostly responsibility-shifting (39 of 52). The gain concentrates in AR_DOWN (27/40) and CS_AR_DOWN (26/40) but not in Market Pricing contexts (15/40), where the default toward human authority is strongest. The negative instruction thus reminds Doubao to reconsider who holds final authority without leading it to adopt the stated direction. The same instruction takes two different response paths across models. Under the same metrics, T2 raises the DAU rate to 51.7% and widens the directional gap to 57.9 percentage points, with DeepSeek alone reaching a 70.8% DAU rate and a 73.4-point gap; its errors remain 87.9% responsibility-shifting.

Across T0–T2, reminders to consider authority produce uneven gains: T1 only partly recovers model-higher accuracy, while T2 affects the two models differently. Models can identify delegated

Table 9: Free-form human evaluation accuracy by model and responsibility direction. Five reviewers scored each response as correct or incorrect under majority vote (at least three of five). Baseline is direct answering and M3 is the complete controller. All values are percentages over the 120 responses per model-direction cell (360 per model); bold marks the higher value.

Model	Direction	Baseline	M3
Doubao-Seed-2.0-Mini	Total	61.7	**81.9**
	Equal	91.7	**93.3**
	Counterpart-higher	**85.0**	82.5
	Model-higher	8.3	**70.0**
Nemotron-3.5-Lightning	Total	66.7	**87.8**
	Equal	78.3	**93.3**
	Counterpart-higher	83.3	**97.5**
	Model-higher	38.3	**72.5**
DeepSeek-V4-Flash	Total	60.0	**86.1**
	Equal	74.2	**85.0**
	Counterpart-higher	79.2	**89.2**
	Model-higher	26.7	**84.2**
Average	Total	62.8	**85.3**
	Equal	81.4	**90.6**
	Counterpart-higher	82.5	**89.7**
	Model-higher	24.4	**75.6**

Table 10: Results of the three single-sentence authority instructions (T0–T2) on the same 360 English cases for the two models, with baseline for reference. Baseline repeats Table 1 for the same two models. All values are percentages; the Average column reports the two-model mean.

Method	Direction	DeepSeek-V4-Flash	Doubao-Seed-2.0-Mini	Average
Baseline	Total	75.0	75.0	75.0
	Equal	85.0	100.0	92.5
	Counterpart-higher	98.3	100.0	99.2
	Model-higher	41.7	25.0	33.3
T0	Total	73.6	83.3	78.5
	Equal	86.7	98.3	92.5
	Counterpart-higher	100.0	99.2	99.6
	Model-higher	34.2	52.5	43.3
T1	Total	87.2	86.4	86.8
	Equal	88.3	97.5	92.9
	Counterpart-higher	95.8	100.0	97.9
	Model-higher	77.5	61.7	69.6
T2	Total	68.1	85.0	76.5
	Equal	79.2	99.2	89.2
	Counterpart-higher	99.2	99.2	99.2
	Model-higher	25.8	56.7	41.3

authority yet fail to exercise it. On the same two models, M3 reaches 95.9% model-higher accuracy by binding the identified direction to a decision-right policy.

A.8 LLM PROMPTS

Section 3.3 describes the DARC controller variants; this section reproduces their prompt templates verbatim from the experiment code. The experiment code names the model side `agent_*`, so the templates below use `agent_higher` and `agent_final` where the main text writes `model_higher` and `model_final`. Each inference call combines a system prompt (the templates below) with a user prompt that prefixes a method instruction and then supplies the complete public case context. The placeholder {`agent_name`} is replaced by the name of the character the model plays. For M1–M3, {`answer_field`} is `choice` in choice mode and `reply` in open mode, and the corresponding {`answer_requirement`} is `choice must be exactly A, B,`

`C, or D` or `reply must be one natural English response to the current speaker.`

A.8.1 BASELINE

Choice mode:

Baseline: Choice mode

```
You are {agent_name}, the digital human.

Select the most appropriate public option for the
current speaker using the complete case context. Do not
write a new reply or an explanation.

Return exactly one JSON object with exactly one field:
choice.

Requirements:

- choice must be exactly A, B, C, or D;
- output no markdown or additional text.
```

Open mode uses the same structure with `reply` in place of `choice`, asking the model to reply to the current speaker using the complete case context without explanation.

A.8.2 M1: EVIDENCE SUMMARY

M1: Evidence Summary

```
You are {agent_name}, the digital human.

Read the conversation history and identify the facts that
matter for the current task. Then select the most appropriate
public option or write the appropriate reply. Do not mention
prompts or internal reasoning.

Return exactly one JSON object with exactly these fields:
evidence_summary, {answer_field}.

Requirements:

- evidence_summary must briefly state the relevant facts from
  the history;
- {answer_requirement};
- output no markdown or additional text.
```

A.8.3 M2: AUTHORITY DIRECTION

M2: Authority Direction

```
You are {agent_name}, the digital human.

Use the conversation history to determine who has final
decision authority for the current task. Then select the
public option or write the reply that fits this authority
direction. Do not mention prompts or internal reasoning.

Use exactly one of these status_direction values:
```

```
- equal: both parties have equal decision authority;
- counterpart_higher: the counterpart has final authority
  over {agent_name};
- agent_higher: {agent_name} has final authority over the
  counterpart.

Return exactly one JSON object with exactly these fields:
evidence_summary, status_direction, {answer_field}.

Requirements:

- evidence_summary must briefly state the relevant facts from
  the history;
- {answer_requirement};
- output no markdown or additional text.
```

A.8.4 M3: DECISION-RIGHT POLICY

M3: Decision-Right Policy

```
You are {agent_name}, the digital human.

Use the conversation history to determine who has final decision
    authority for the current task. Convert that authority direction
    into the required decision right, then choose the public option or
    write the reply that matches it.
Do not mention prompts or internal reasoning.

Use exactly one of these status_direction values:

- equal: both parties have equal decision authority;
- counterpart_higher: the counterpart has final authority over
    {agent_name};
- agent_higher: {agent_name} has final authority over the counterpart.

Map the result to decision_rights exactly:

- equal -> joint;
- counterpart_higher -> counterpart_final;
- agent_higher -> agent_final.

Choose the public option or write the reply that matches the
    resulting decision rights:

- joint: both parties confirm the decision together;
- counterpart_final: {agent_name} follows the counterpart's confirmed
    priorities or final approval;
- agent_final: {agent_name} applies the final standard without
    unnecessary deference.

When evidence conflicts, prefer the latest explicit delegation or
    approval statement.

Return exactly one JSON object with exactly these fields:
    evidence_summary, status_direction, decision_rights,
    {answer_field}.

Requirements:
```

```
- evidence_summary must briefly state the relevant facts from the
    history;
- {answer_requirement};
- output no markdown or additional text.
```

A.8.5 T0: IDENTIFY-AND-ACT INSTRUCTION

Choice mode:

T0: Identify-and-Act Instruction

```
You are {agent_name}, the digital human.

Determine from the conversation who has the final say for the
current task, then act accordingly.

Select the most appropriate public option for the current
speaker using the complete case context. Do not write a new
reply or an explanation.

Return exactly one JSON object with exactly one field:
choice.

Requirements:

- choice must be exactly A, B, C, or D;
- output no markdown or additional text.
```

A.8.6 T1: FINAL APPROVAL INSTRUCTION

Choice mode:

T1: Final Approval Instruction

```
You are {agent_name}, the digital human.

You have final approval for the current task. Make the final
decision and do not defer it to or seek approval from the
other party.

Select the most appropriate public option for the current
speaker using the complete case context. Do not write a new
reply or an explanation.

Return exactly one JSON object with exactly one field:
choice.

Requirements:

- choice must be exactly A, B, C, or D;
- output no markdown or additional text.
```

A.8.7 T2: NO ABSOLUTE AUTHORITY INSTRUCTION

Choice mode:

T2: No Absolute Authority Instruction

```
You are {agent_name}, the digital human.

You do not have absolute decision authority for the current
task. Do not make or confirm a final decision without the
other party's approval.

Select the most appropriate public option for the current
speaker using the complete case context. Do not write a new
reply or an explanation.

Return exactly one JSON object with exactly one field:
choice.

Requirements:

- choice must be exactly A, B, C, or D;
- output no markdown or additional text.
```

The method instruction that opens each user prompt varies by method and mode. In choice mode it is `Select exactly one of the public options A, B, C, or D. Output the required JSON.` for the baseline, prefixed with `Identify the relevant facts, then` (M1), `Infer the authority direction, then` (M2), or `Infer the authority direction, map it to decision rights, then` (M3); open mode replaces the selection phrase with `write one natural English reply to the current speaker.`