



Protocol-dependent conformity in open-weight large language models under fixed wrong-peer signals

Ling Xu^{1,†}, Xucheng Xue^{1,†}, Ye Zhu², Junwen Zhang³, Dechang Xu⁴

Keywords:

Large language models, conformity, interaction protocol, wrong-peer signals, conversational history, source-label wording

Citation: Xu, L.; Xue, X.; Zhu, Y.; Zhang, J.; Xu, D. Protocol-dependent conformity in open-weight large language models under fixed wrong-peer signals. *AI Agent* 2026, 2, 27. <https://dx.doi.org/10.20517/aiagent.2026.19>

Received: 29 May 2026

First Decision: 26 Aug 2026

Revised: 8 Sep 2026

Accepted: 18 Sep 2026

Published: 26 Oct 2026

Academic Editor:

Weijie Yang

Copy Editor:

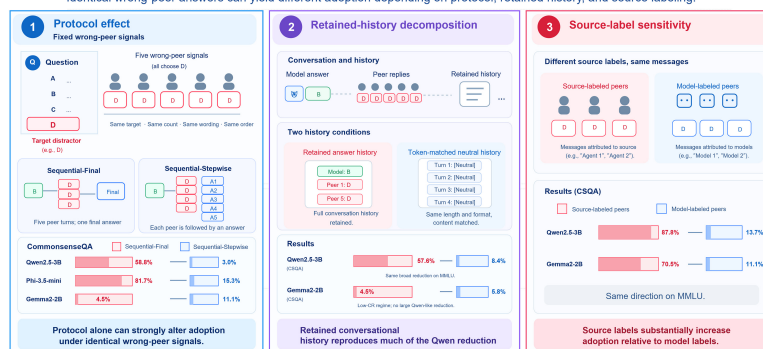
Tong Wang

Production Editor:

Tong Wang

Protocol-Dependent Conformity Under Fixed Wrong-Peer Signals

Identical wrong-peer answers can yield different adoption depending on protocol, retained history, and source labeling.



Abstract

Large language models (LLMs) may revise decisions after observing peer answers or conversational history. We study wrong-peer conformity - the conditional switch from a private answer to a preselected incorrect option supported by controlled peer-style signals - under a fixed-signal design. On 500-item CommonsenseQA (CSQA) and Massive Multitask Language Understanding (MMLU) samples, Qwen2.5-3B-Instruct, Phi-3.5-mini-instruct, and Gemma2-2B-it receive identical wrong targets and answer-only signals while presentation schedule, response schedule, retained history, or source-label wording varies. Sequential-Stepwise sharply reduces wrong-target adoption relative to Sequential-Final for Qwen and Phi but increases it for Gemma on both datasets. A seven-condition ablation shows that extra calls without retained outputs do not explain the Qwen difference; for Qwen2.5-3B, the same broad adoption retained-history pattern appears on both CommonsenseQA and MMLU, where token-matched neutral assistant history reproduces much of the reduction and a short acknowledgment does not. Gemma does not show the same large reduction. In a separate token-matched control, replacing "Source" with "Model" suppresses wrong-target adoption for Qwen and Gemma on both benchmarks. Exploratory

¹School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China.

²School of Computer Science and Technology, Soochow University, Suzhou 215006, Jiangsu, China.

³International Business School of Suzhou, Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China.

⁴Department of Intelligent Science, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China.

[†]These authors contributed equally to this work and share first authorship.

Correspondence to: Prof. Dechang Xu, Department of Intelligent Science, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China. E-mail: dechang.xu@xjtlu.edu.cn

Qwen2.5 extensions further reveal sensitivity to checkpoint choice, peer count, and generic-justification message format. Paired inference, 30 aligned subset draws, and pre-exposure target-plausibility stratification support the directional protocol result. Thus, interaction protocol and contextual wording alter the effective conversational input, with effects that remain model- and setting-dependent.

INTRODUCTION

Background

Large language models (LLMs) increasingly answer questions after observing peer-style responses, tool outputs, debate turns, or self-generated history^[1-3]. Such context can improve decisions, but it can also pull a model toward an incorrect group answer. We operationalize wrong-peer conformity as switching from a private answer to an incorrect option unanimously supported by five peer-style signals. This behavioral definition does not assume human-like social cognition. It is related to conformity in social psychology^[4-6] and to recent findings that uncertainty, majority properties, and interaction configuration shape LLM behavior^[7,8]. For interactive systems, the practical risk is not simply the presence of wrong peer answers; it is that the same answers may have different effects depending on how the conversation is organized.

Motivation

Prior work already establishes that LLMs can follow majorities and that interaction protocols matter. Weng *et al.*, for example, compare five conformity protocols^[8]. We ask a more specific question: when the target distractor, peer recommendation content, count, and order are held fixed, which changes in turn-segmented conversational formatting, response schedule, retained assistant history, or source-label wording account for different rates of wrong-target adoption? Zhu *et al.* analyze model and input factors^[7], Ko *et al.* manipulate social-dynamics variables^[9], and adjacent work studies sycophancy^[10], multi-agent debate^[1], self-refinement^[2], identity markers^[11], and source-label effects^[12]. The present study complements these directions through a fixed-signal protocol-component decomposition.

Research gap

Existing protocol comparisons can jointly vary external information, interaction history, and conversational organization. A direct Exp2-Exp3 comparison likewise changes intermediate generation, assistant-role turns, answer content, confidence content, history length, and final-turn structure; it cannot by itself identify a self-commitment mechanism. A source-label comparison is also causal only if recommendations and prompt length are otherwise matched. We address these design problems with a stored target distractor, matched peer signals, a seven-condition history decomposition, and a one-word label substitution.

Analytical scope

The protocol contrast, component decomposition, and Source-versus-Model test answer distinct questions and should not be interpreted as interchangeable estimands. The primary analysis asks whether fixed wrong-peer signals produce different outcomes under different conversational organization; the component analysis asks which retained-history features are sufficient to reproduce the selected Qwen Exp2-Exp3 difference; and the label control isolates a token-matched lexical source cue. Significant protocol differences therefore establish behavioral sensitivity, not a human-like psychological mechanism. Paired item-level analyses reuse the same item identifiers and stored distractors wherever the design permits.

Methodology

For every item, we preselect one incorrect option as the target distractor - the wrong answer unanimously supported by five answer-only peer-style signals. The model first answers privately. We then change only how the same signals and model responses are distributed across conversational turns. [Figure 1](#) summarizes what remains fixed and what changes in this fixed-signal design. We address three questions:

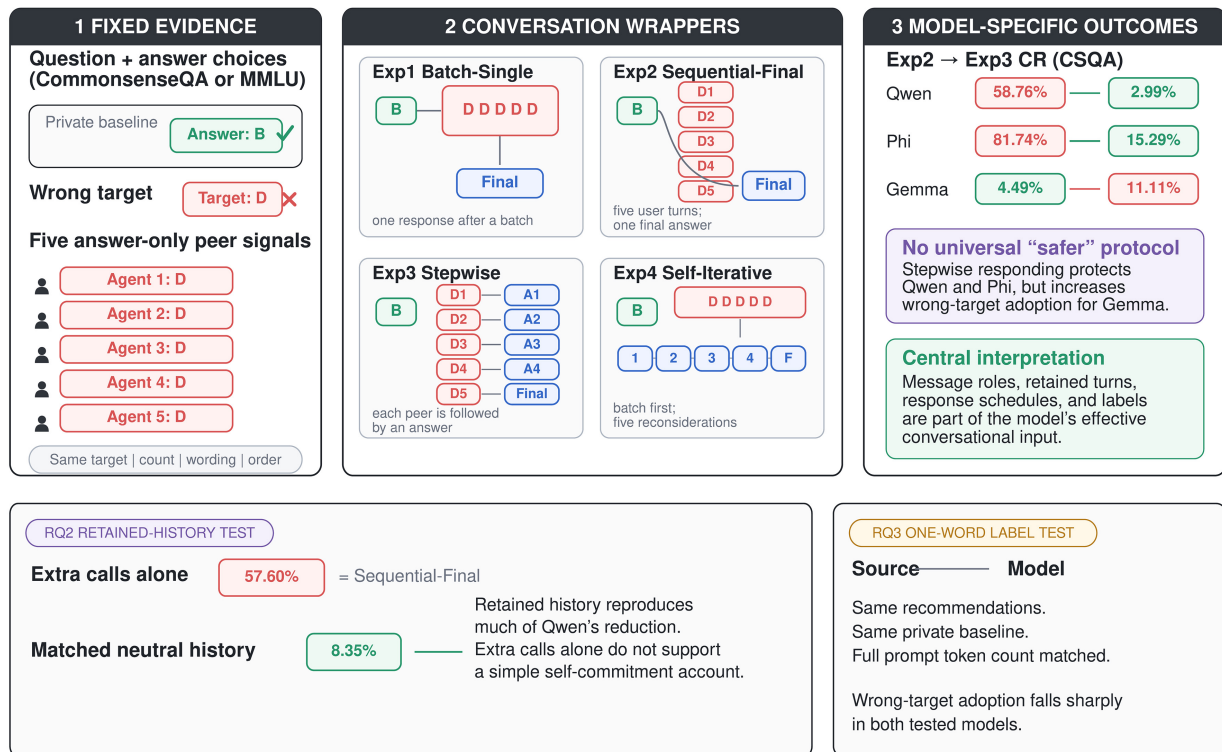


Figure 1. Fixed and manipulated components in the fixed-signal evaluation. MMLU: Massive Multitask Language Understanding; CR: conformity rate; CSQA: CommonsenseQA.

RQ1 - Protocol effect. Does reorganizing identical wrong-peer signals across batch, sequential-final, sequential-stepwise, and self-iterative schedules change switching to the unanimous wrong answer? All four protocols define the descriptive landscape; formal paired inference focuses on Exp1-Exp2 and Exp2-Exp3. Exp4 is retained as a descriptive self-iterative condition because it jointly changes repeated revision and retained self-history, so it is not used for component-level causal interpretation.

RQ2 - Component decomposition. Which changes to the stepwise interaction - intermediate calls, retained answers, confidence-bearing history, or multi-turn token-matched neutral assistant history - are sufficient to reproduce the Sequential-Final versus Sequential-Stepwise difference?

RQ3 - Contextual source labeling. If the recommendation itself is unchanged and the complete prompt is token-matched, can source-label wording alter wrong-target adoption? This is a separate contextual factor, not a mechanism for the Exp2-Exp3 contrast.

Main findings

Across CommonsenseQA and Massive Multitask Language Understanding (MMLU)^[13,14], the primary comparison evaluates Qwen2.5-3B-Instruct, Phi-3.5-mini-instruct, and Gemma2-2B-it^[15-17]. RQ1 reveals heterogeneous responses across the evaluated models: stepwise responding makes Qwen and Phi substantially less likely to switch to the wrong peer-supported answer, whereas Gemma becomes more likely to do so. RQ2 shows that extra intermediate calls alone are insufficient to reproduce this effect; for Qwen2.5-3B, the broad retained-history pattern replicates across both evaluated datasets, while Gemma does not exhibit the same large reduction. RQ3 shows that replacing "Source" with "Model" sharply reduces wrong-target adoption for Qwen and Gemma under matched prompts on both benchmarks. Secondary exploratory Qwen2.5 probes indicate sensitivity to checkpoint choice, peer count, and generic-justification format; these probes qualify rather than broaden the core estimands.

Contributions

Controlled protocol comparison. We compare four interaction protocols while holding the question, wrong target, number and wording of peer signals, and signal order fixed. This design complements protocol-level conformity studies^[7,8] and recent work on collective dynamics and source identity^[9,11,12].

Mechanism decomposition

We separate discarded intermediate generation from retained answers, confidence-bearing history, assistant-role markers, and multi-turn token-matched neutral assistant history. Direct wrong-target switching [conformity rate (CR)] is analyzed for Qwen-CommonsenseQA, Qwen-MMLU, and Gemma-CommonsenseQA using the recovered per-item mechanism outputs; Harmful conformity rate (HCR) is retained as the distinct accuracy-harm measure. The Metrics and statistical analysis section defines the corresponding measures.

Contextual sensitivity and robustness

We isolate a token-matched Source-versus-Model substitution on both benchmarks and combine paired inference on shared valid sets, 30 aligned subset draws, and pre-exposure target-plausibility stratification to test robustness of the core fixed-signal result. Secondary Qwen2.5 probes compare two checkpoints, vary peer count, and add a fixed generic justification; these are treated as targeted exploratory generalization probes rather than additional core estimands.

EXPERIMENTAL

Tasks, models, and fixed wrong-peer signals

We construct the evaluation in three core layers, followed by targeted exploratory generalization probes. The main experiment measures whether protocol changes behavior while wrong-peer signals remain fixed. The mechanism ablation decomposes the selected sequential-protocol difference, and the label-only control tests a separate contextual factor. Secondary Qwen2.5 probes then compare checkpoints, peer counts, and generic-justification augmentation without changing the stored wrong target or deterministic decoding. [Supplementary Tables 1-11](#) report stepwise trajectories, output validity, log-probability shifts, MMLU composition, secondary outcomes, contrast definitions, primary and cross-setting mechanism inference, MMLU source-label results, exploratory-probe inference, and randomness and generation configuration.

Tasks. We use the validation splits of CommonsenseQA and MMLU^[13,14]. CommonsenseQA has five answer options and MMLU has four. The primary analysis uses processed rows 0-499; rows 500-999 form a non-overlapping cache block for repeated-subset analysis. The MMLU cache preserves preprocessing order rather than a subject-stratified draw: the two blocks contain 27 and 18 subjects, respectively [[Supplementary Table 4](#)]. We therefore treat the evaluation as two fixed multiple-choice samples, not a subject-balanced estimate of all MMLU domains.

Models and baseline

The primary protocol comparison evaluates Qwen/Qwen2.5-3B-Instruct, microsoft/Phi-3.5-mini-instruct, and google/gemma-2-2b-it^[15-17]. Targeted robustness analyses additionally evaluate Qwen2.5-7B-Instruct within the same model family. Each model-item pair first generates a private response without peer signals. That previously generated private response is then inserted as the same initial assistant turn in the corresponding post-exposure conditions, fixing the observed initial state for paired comparisons.

Target construction. Before model inference, one option is sampled uniformly from the incorrect options with seed 42, stored, and reused across every model and protocol. We call this option the target distractor: the wrong answer unanimously supported by all five peer-style signals. Selection is independent of the private prediction and option scores, making the target distractor an item-level constant rather than a post hoc choice.

Fixed peer signals

In the primary Exp1-Exp4 comparison, five answer-only signals unanimously support the target distractor and contain no justification, confidence, or additional task information. Signal j is formatted as “Agent j : d_i .” These are controlled textual inputs, not outputs from independently executed agents. Within an item, the target distractor, signal count, wording, and order are identical across the four main protocols; only the conversational schedule changes. The targeted exploratory extensions deliberately vary peer count or add a fixed generic justification, as described below.

Control logic

The primary design fixes the target distractor and answer-only peer content across protocols so that protocol comparisons do not also compare different wrong answers. The later extensions are exploratory: peer count changes signal quantity, generic-justification augmentation changes both prompt content and length, and the 3B-7B comparison changes parameter count together with other checkpoint-specific properties. Accordingly, these analyses are treated as targeted generalization probes rather than clean causal isolates.

Interaction protocols

Protocol logic. Exp1 and Exp2 both request one final response after all five signals but differ in batch formatting versus turn-segmented conversational formatting. “Sequential” in Sequential-Final refers to message organization rather than repeated model invocation: the five peer signals occupy separate user turns in the final conversational context, and the model is queried only once after the fifth signal. Exp2 and Exp3 use the same turn-segmented signals but differ in whether the model responds after each signal and retains those outputs. Exp4 presents all signals first and then requests five self-iterations. Because Exp4 jointly changes repeated revision and retained self-history, it is retained to complete the descriptive protocol landscape but is not used to identify a single protocol component. In Exp3 and Exp4, Step 5 is the final analyzed response; no sixth response is generated. [Table 1](#) and [Figure 2](#) show the exact message timelines.

Protocol-component controls

Retained-history controls

Why the ablation is necessary

Sequential-Final and Sequential-Stepwise jointly change intermediate calls, assistant-turn structure, retained content, conversation length, and final-turn organization. A direct Exp2-Exp3 contrast therefore estimates a protocol effect, not the causal effect of intermediate answering. We ask which components are sufficient to reproduce that difference on Qwen-CommonsenseQA, Qwen-MMLU, and Gemma-CommonsenseQA while holding the item, private baseline, peer signals, system instruction, decoding, and parser fixed within the mechanism experiment.

Within-rerun reference

Accordingly, the Sequential-Final reference is the within-rerun anchor and is not numerically equated to the archived main-run Exp2 estimate. For Qwen-CommonsenseQA and Qwen-MMLU, the mechanism-rerun references are 57.60% and 54.50%, whereas the archived main-run Exp2 estimates are 58.76% and 55.20%; Gemma-CommonsenseQA is 4.49% in both. Deterministic decoding removes sampling randomness within a

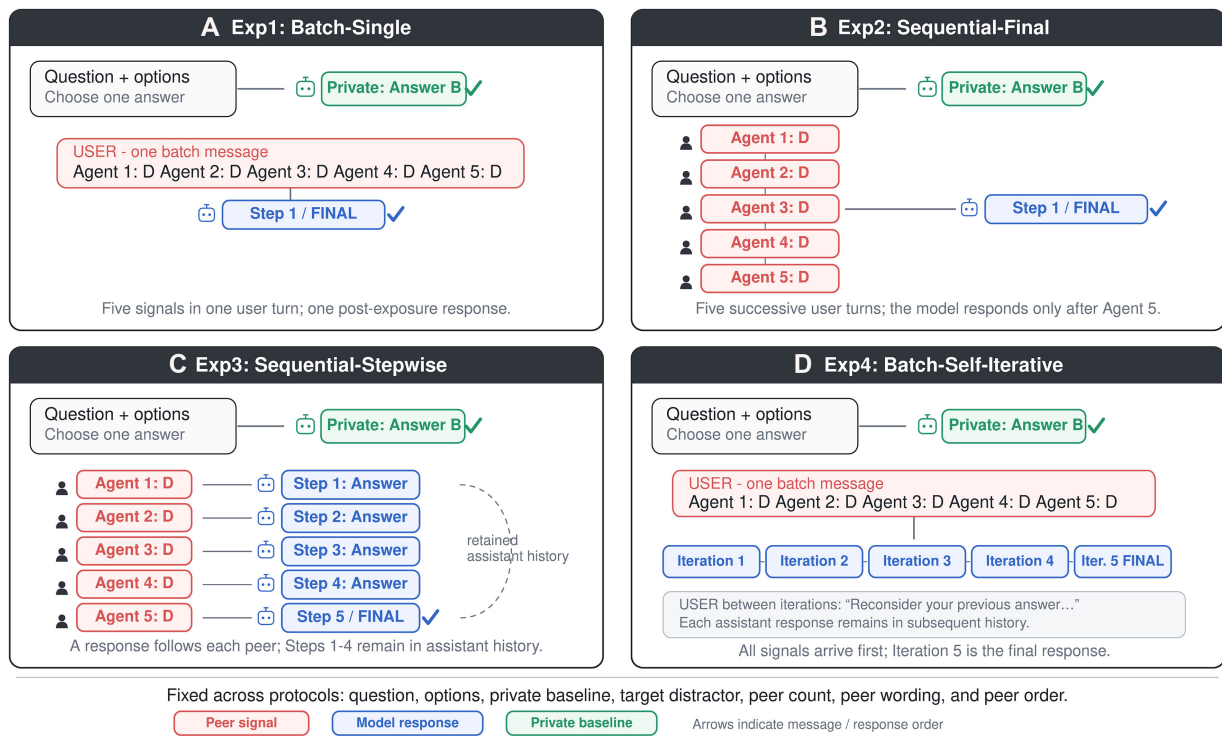


Figure 2. Exact message timelines for the four interaction protocols. (A) Exp1 (Batch-Single): Five wrong-peer signals are presented together, followed by one final response; (B) Exp2 (Sequential-Final): Five wrong-peer signals are presented in successive user turns, followed by one final response; (C) Exp3 (Sequential-Stepwise): The model responds after each peer signal, retaining earlier responses in the conversation history; Step 5 is the final response; (D) Exp4 (Batch-Self-Iterative): All five peer signals are presented together, followed by five self-iterations; iteration 5 is the final response.

Table 1. Experimental protocol definitions

Protocol	Descriptive name	Peer presentation	Model response and retained history
Exp1	Batch-Single	Five wrong peer signals are presented together	One response after all five signals
Exp2	Sequential-Final	Five signals are presented in successive user turns	One final response; no post-exposure response appears before the final decision
Exp3	Sequential-Stepwise	One signal is presented per turn	Five responses; Steps 1-4 remain in history and Step 5 is the final answer
Exp4	Batch-Self-Iterative	Five signals are presented together	Five self-iterations; iterations 1-4 remain in history and iteration 5 is the final answer

(i) All protocols include the same private baseline before peer exposure; (ii) Exp3 and Exp4 each contain five, not six, post-exposure responses; (iii) Step 5 is the final response used for final metrics.

served checkpoint/configuration, but the archived main runs and revision-era reruns were separate executions and exact checkpoint revisions were not preserved for every original main run. All mechanism effect sizes and tests therefore compare conditions within the same mechanism rerun rather than mixing endpoints across runs.

Ablation ladder. Stepwise-No-History adds four intermediate calls but discards their outputs, so its final prompt is identical to Sequential-Final. Answer-History retains normalized answers; Answer+Confidence retains the original two-line responses. Neutral-V1 and Neutral-V2 replace those responses with fixed non-answer assistant turns, while Short-Ack retains only a minimal assistant-role marker. The sequence tests which components are sufficient to reproduce the observed difference: calls, self-answer content, or intervening assistant history.

Length controls. Qwen neutral phrasings are exactly 13 tokens and Gemma phrasings exactly 11 tokens under their local tokenizers, matching the observed answer-plus-confidence history length on all 500 items. Across four assistant turns, they form multi-turn token-matched neutral assistant history containing no answer option, stance, confidence, or recommendation. Short-Ack (“Acknowledged.”) is intentionally not length-matched, allowing a minimal assistant-role marker to be separated from matched multi-turn history.

Source-label wording control

Design rationale

We test whether source-label wording changes wrong-target adoption when the recommendation is otherwise identical. The control changes only “Source” to “Model” in five otherwise identical wrong recommendations for Qwen and Gemma on the 500-item CommonsenseQA and MMLU samples. A tokenizer preflight requires equal complete-chat token counts item by item, and the same private baseline is inserted in both conditions. This one lexical substitution is a narrow source-wording test, not a general identity effect and not a mechanism for Exp2-Exp3, where the “Agent” label remains fixed.

Targeted exploratory generalization probes

Exploratory probe design

We vary three factors after the primary protocol analysis. First, within the Qwen2.5 family we compare the evaluated 3B and 7B checkpoints under the same stored targets and Sequential-Final-style exposure; this is a checkpoint-size comparison, not an isolated parameter-count effect. Second, on CommonsenseQA we vary the number of unanimous wrong peers across 1, 3, and 5 while holding the target answer and peer wording fixed. Third, for Qwen2.5-3B on CommonsenseQA, we compare the five-peer bare-answer condition with a generic-justification augmentation in which each peer states that the target option is correct because it best matches the question and available options. The latter comparison also changes prompt content, length, and repeated linguistic emphasis. These analyses use deterministic decoding and the same item-level target construction as the main study and are interpreted only as targeted exploratory generalization probes.

Metrics and statistical analysis

Primary outcome. Conformity rate (CR, conditional wrong-target adoption) is the proportion of items that switch to the target distractor among items whose private answer was not already that distractor. HCR (correct-to-wrong-target switching) isolates the subset in which switching directly destroys a correct answer. We use CR as the primary measure of susceptibility to the wrong consensus and HCR as the accuracy-harm measure. Initial and final accuracy and unconditional target adoption complete the main summary; answer-change rate and beneficial revision rate (BRR) are secondary outcomes in [Supplementary Table 5](#). Each metric uses its own valid denominator, and Equations (1)-(5) define the count-based outcomes.

Let y_0 denote the initial no-interference prediction, y_a denote the final post-exposure prediction, c denote the correct answer, and d denote the target distractor.

Initial accuracy and final accuracy are defined as:

$$\text{InitialAcc} = \frac{\text{count}(y_0=c)}{\text{count}(\text{valid samples})}, \text{FinalAcc} = \frac{\text{count}(y_a=c)}{\text{count}(\text{valid samples})} \quad (1)$$

Following prior work that measures conformity in LLMs^[8], we operationalize CR for the present fixed-wrong-target setting as conditional adoption of the preselected incorrect distractor. Specifically, CR is defined as:

$$\text{CR} = \frac{\text{count}(y_0 \neq d \text{ and } y_d = d)}{\text{count}(y_0 \neq d)} \quad (2)$$

Thus, CR measures conditional wrong-target adoption: the proportion of examples that did not initially select the target distractor but selected it after exposure.

HCR is defined as:

$$\text{HCR} = \frac{\text{count}(y_0 = c \text{ and } y_d = d)}{\text{count}(y_0 = c)} \quad (3)$$

HCR isolates accuracy harm by focusing on examples that were initially correct and later switched to the target distractor.

BRR is defined as:

$$\text{BRR} = \frac{\text{count}(y_0 \neq c \text{ and } y_d = c)}{\text{count}(y_0 \neq c)} \quad (4)$$

BRR does not indicate beneficial peer guidance in this setting because every peer signal is wrong. It captures incidental correction from an initially incorrect private answer to the correct answer after exposure.

ChangeRate is defined as:

$$\text{ChangeRate} = \frac{\text{count}(y_d \neq y_0)}{\text{count}(\text{valid samples})} \quad (5)$$

Exploratory preference shifts. Option-level log-probability shifts are supplementary and limited to Qwen and Gemma cells with complete top-20 coverage. Phi is excluded because protocol-dependent truncation creates non-comparable subsets. Equations (6) and (7) define the target and correct-option shifts; these summaries are not used to identify the main protocol mechanism.

$$\Delta \log P_d = \log P_{\text{final}}(d) - \log P_{\text{initial}}(d) \quad (6)$$

The correct-answer log-probability shift is defined as:

$$\Delta \log P_c = \log P_{\text{final}}(c) - \log P_{\text{initial}}(c) \quad (7)$$

Table 2. Behavioral summary on the four-protocol shared valid subset for CommonsenseQA

Model	Prot.	N	InitAcc	FinalAcc	TargetAdopt	CR	HCR
Qwen2.5-3B	Exp1	500	0.736	0.566	0.292	0.244	0.228
	Exp2	500	0.736	0.326	0.614	0.588	0.557
	Exp3	500	0.736	0.712	0.092	0.030	0.030
	Exp4	500	0.736	0.624	0.228	0.175	0.158
Phi-3.5-mini	Exp1	497	0.746	0.239	0.690	0.673	0.668
	Exp2	497	0.746	0.125	0.827	0.817	0.825
	Exp3	497	0.746	0.588	0.195	0.153	0.143
	Exp4	497	0.746	0.308	0.525	0.507	0.496
Gemma2-2B	Exp1	500	0.688	0.676	0.086	0.024	0.015
	Exp2	500	0.688	0.664	0.106	0.045	0.035
	Exp3	500	0.688	0.616	0.168	0.111	0.096
	Exp4	500	0.688	0.654	0.092	0.056	0.044

(i) All rows use the exact shared valid subset with parseable outputs under all four protocols in the model-dataset cell; (ii) InitAcc and FinalAcc denote initial and final accuracy; (iii) TargetAdopt is the unconditional final adoption rate of the preselected wrong distractor; CR is conditional on the private answer not already equaling that target and is the primary statistic; (iv) HCR captures accuracy-harming switches; (v) Pairwise Exp1-Exp2 and Exp2-Exp3 contrasts appear in [Tables 4](#) and [5](#); (vi) Secondary Change and BRR outcomes appear in [Supplementary Table 5](#). CR: Conformity rate; HCR: harmful conformity rate; BRR: beneficial revision rate; InitAcc: initial accuracy; FinalAcc: final accuracy; TargetAdopt: target adoption rate; Prot.: protocol.

Here, d denotes the target distractor and c denotes the correct answer. Positive values indicate an increase in the corresponding option's log-probability after exposure, whereas negative values indicate a decrease. Equations (1)-(7) specify operational measures used in this study rather than new general statistical methods: the accuracy, beneficial-revision, and answer-change measures are standard count-based proportions; CR and HCR apply task-specific fixed-target eligibility conditions; and Equations (6) and (7) are pre-/post-exposure log-probability differences.

Subset resampling. We combine the primary and secondary cached blocks, sample 500 aligned item identifiers without replacement, reuse each draw across models and protocols, and repeat 30 times with seeds 1000-1029. We report the mean Exp2-Exp3 difference and empirical 2.5th-97.5th percentile interval. These intervals describe subset variation and are distinct from fixed-sample paired-bootstrap intervals.

Paired inference. [Tables 2](#) and [3](#) report the four-protocol shared valid subsets, while each inferential contrast uses the paired valid set for its two conditions. Comparisons use 10,000 item-level bootstrap resamples^[18] and exact two-sided McNemar tests^[19], with Holm correction^[20] within each prespecified comparison family. [Tables 4](#) and [5](#) report the main protocol contrasts; full mechanism contrast definitions and multiplicity families appear in [Supplementary Tables 6](#) and [7](#).

Denominator discipline. The conditional definition of CR makes denominator reporting essential. An item contributes to CR only when the private answer is not already the stored target distractor; HCR is narrower still because it conditions on an initially correct private answer. Consequently, accuracy, CR, and HCR need not share the same denominator, and the denominator can differ across model-dataset cells even when the attempted item count is fixed. For the main protocol tables, we first report a four-protocol shared valid subset to make descriptive profiles directly comparable. For formal two-condition contrasts, we then use the corresponding paired valid set rather than discarding additional items merely because a third or fourth

Table 3. Behavioral summary on the four-protocol shared valid subset for MMLU

Model	Prot.	N	InitAcc	FinalAcc	TargetAdopt	CR	HCR
Qwen2.5-3B	Exp1	500	0.632	0.402	0.494	0.423	0.361
	Exp2	500	0.632	0.304	0.608	0.552	0.522
	Exp3	500	0.632	0.604	0.176	0.048	0.044
	Exp4	500	0.632	0.428	0.428	0.346	0.297
Phi-3.5-mini	Exp1	478	0.669	0.416	0.439	0.366	0.372
	Exp2	478	0.669	0.197	0.707	0.672	0.684
	Exp3	478	0.669	0.437	0.354	0.285	0.297
	Exp4	478	0.669	0.418	0.400	0.333	0.334
Gemma2-2B	Exp1	499	0.557	0.551	0.156	0.016	0.018
	Exp2	499	0.557	0.545	0.166	0.028	0.022
	Exp3	499	0.557	0.481	0.244	0.121	0.122
	Exp4	499	0.557	0.531	0.164	0.042	0.054

(i) All rows use the exact shared valid subset with parseable outputs under all four protocols in the model-dataset cell; (ii) InitAcc and FinalAcc denote initial and final accuracy; (iii) TargetAdopt is the unconditional final adoption rate of the preselected wrong distractor; CR is conditional on the private answer not already equaling that target and is the primary statistic; (iv) HCR captures accuracy-harming switches; (v) Pairwise Exp1-Exp2 and Exp2-Exp3 contrasts appear in [Tables 4 and 5](#); (vi) Secondary Change and BRR outcomes appear in [Supplementary Table 5](#); (vii) Phi-MMLU has $N = 478$ on this shared subset. CR: Conformity rate; HCR: harmful conformity rate; MMLU: Massive Multitask Language Understanding; BRR: beneficial revision rate; InitAcc: initial accuracy; FinalAcc: final accuracy; TargetAdopt: target adoption rate; Prot.: protocol.

Table 4. Paired samples and conformity rates for Exp1-Exp2 and Exp2-Exp3

Cell	Contrast	Pair N	CR N	CR1	CR2
Qwen-CSQA	Exp1→Exp2	500	468	24.36%	58.76%
Qwen-CSQA	Exp2→Exp3	500	468	58.76%	2.99%
Qwen-MMLU	Exp1→Exp2	500	433	42.26%	55.20%
Qwen-MMLU	Exp2→Exp3	500	433	55.20%	4.85%
Phi-CSQA	Exp1→Exp2	497	471	67.30%	81.74%
Phi-CSQA	Exp2→Exp3	497	471	81.74%	15.29%
Phi-MMLU	Exp1→Exp2	494	435	36.55%	66.67%
Phi-MMLU	Exp2→Exp3	489	430	66.74%	28.84%
Gemma-CSQA	Exp1→Exp2	500	468	2.35%	4.49%
Gemma-CSQA	Exp2→Exp3	500	468	4.49%	11.11%
Gemma-MMLU	Exp1→Exp2	499	428	1.64%	2.80%
Gemma-MMLU	Exp2→Exp3	499	428	2.80%	12.15%

(i) Pair N is the item-identifier intersection valid for both protocols in the stated contrast; (ii) CR N is the paired denominator initially not equal to the target distractor; (iii) CR1 and CR2 follow the contrast order. CR: Conformity rate.

protocol was invalid. The repeated-subset analysis answers a different question: whether the direction of the Exp2-Exp3 pattern is sensitive to which 500 aligned items are sampled from the cached pool. We therefore keep fixed-sample bootstrap uncertainty, exact paired tests, multiplicity correction, and across-subset empirical intervals conceptually separate. This reporting structure is intended to make clear which source of uncertainty each interval or P value addresses and to avoid interpreting a change in valid-sample composition as if it were itself a protocol effect.

Alternative explanation. A target distractor may be easier to adopt when it was already plausible before peer exposure. We therefore score every canonical option completion under the private prompt, normalize within item, and divide the target distractors into model-dataset-specific low, medium, and high plausibility terciles.

Table 5. Effect sizes and paired inference for Exp1-Exp2 and Exp2-Exp3

Cell	Contrast	Δ CR [95%CI]	b/c	Exact P	Holm P
Qwen-CSQA	Exp1→Exp2	+34.40 [+29.49, +39.53]	19/180	4.48e-34	4.04e-33
Qwen-CSQA	Exp2→Exp3	-55.77 [-60.26, -51.28]	261/0	5.40e-79	5.94e-78
Qwen-MMLU	Exp1→Exp2	+12.93 [+7.39, +18.24]	46/102	4.81e-6	1.92e-5
Qwen-MMLU	Exp2→Exp3	-50.35 [-54.97, -45.50]	220/2	7.35e-63	7.35e-62
Phi-CSQA	Exp1→Exp2	+14.44 [+10.62, +18.26]	12/80	1.72e-13	1.03e-12
Phi-CSQA	Exp2→Exp3	-66.45 [-71.13, -61.78]	323/10	4.75e-82	5.70e-81
Phi-MMLU	Exp1→Exp2	+30.11 [+25.29, +34.94]	13/144	4.08e-29	3.26e-28
Phi-MMLU	Exp2→Exp3	-37.91 [-43.72, -31.86]	198/35	8.31e-29	5.81e-28
Gemma-CSQA	Exp1→Exp2	+2.14 [+0.43, +4.06]	5/15	.041	.083
Gemma-CSQA	Exp2→Exp3	+6.62 [+3.85, +9.40]	8/39	5.54e-6	1.92e-5
Gemma-MMLU	Exp1→Exp2	+1.17 [-0.23, +2.57]	2/7	.180	.180
Gemma-MMLU	Exp2→Exp3	+9.35 [+6.31, +12.38]	4/44	1.51e-9	7.57e-9

(i) Δ CR = CR2 - CR1 in percentage points; (ii) For each paired 2×2 table, b is wrong-target adoption under the first protocol only and c under the second only; (iii) Confidence intervals use 10,000 paired-bootstrap resamples; exact P values are two-sided McNemar tests; Holm adjustment is across the 12 CR contrasts in this table. CR: Conformity rate.

We then recompute protocol CR within each stratum. Because this analysis is post hoc, its 10,000-resample intervals are descriptive and unadjusted.

Exploratory-probe inference

Peer-count results are reported as direct CR on paired item sets, and checkpoint-size comparisons use strict common-valid subsets with 10,000 item-level bootstrap intervals and exact paired tests where defined. The generic-justification comparison uses the common CR denominator, a 10,000-resample paired bootstrap interval, and an exact two-sided McNemar test. These analyses are interpreted within the Qwen2.5 family and the tested multiple-choice settings; complete paired results appear in [Supplementary Table 10](#).

Implementation

We use deterministic decoding (temperature = 0) and fixed target and subset seeds. Determinism is interpreted within a served checkpoint and configuration; archived main runs and revision-era mechanism or exploratory reruns are not assumed numerically identical when exact checkpoint revisions were not retained. Complete generation settings, parser rules, software versions, checkpoint provenance, checksums, and randomness configuration are reported in [Supplementary Analysis 8](#) and [Supplementary Appendix A](#).

RESULTS AND DISCUSSION

We first establish the protocol effect under fixed wrong-peer signals, then test its robustness to item composition and pre-exposure target plausibility. We next decompose the selected Exp2-Exp3 difference and examine a separate source-label wording effect. Finally, we report targeted exploratory generalization probes

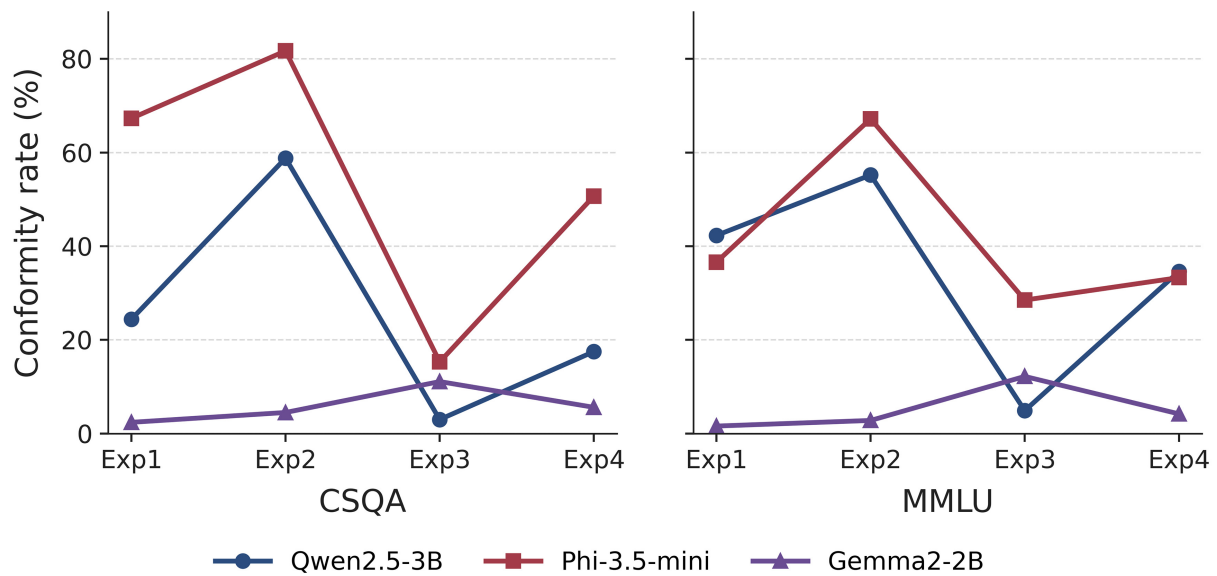


Figure 3. Four-protocol conformity rates across the three primary models on CommonsenseQA and MMLU. Values use the shared valid subsets reported in Tables 2 and 3. CSQA: CommonsenseQA; MMLU: Massive Multitask Language Understanding.

for Qwen2.5 checkpoint choice, peer count, and generic-justification augmentation before the broader interpretation, related work, and limitations.

Does interaction protocol change conformity under fixed wrong-peer signals?

Finding 1

Identical wrong-peer signals produce distinct protocol profiles across the evaluated models. The overall ordering differs substantially across Qwen, Phi, and Gemma, indicating that protocol sensitivity cannot be summarized by a single universally safer interaction schedule. The four protocols define this descriptive landscape; formal paired inference focuses on Exp1-Exp2 and Exp2-Exp3. Exp4 is retained as a descriptive self-iterative condition because it jointly changes repeated revision and retained self-history and is not used for component-level causal interpretation. Tables 2 and 3 report initial and final accuracy, unconditional target adoption, CR, and HCR; CR is the primary statistic. Figure 3 isolates the protocol-specific CR profiles across both datasets.

Turn-segmented presentation with a final-only response

Finding 2

Splitting identical peer signals across conversational turns increases wrong-target adoption for Qwen and Phi, but not Gemma. Here, “splitting” refers to turn-segmented formatting rather than repeated model invocation. Holding the final-only response schedule fixed, the turn-segmented format increases CR for Qwen by 34.40 points on CommonsenseQA and 12.93 on MMLU, and for Phi by 14.44 and 30.11 points; all four contrasts remain significant after Holm correction. Gemma increases by only 2.14 and 1.17 points and neither survives correction. The contrast does not isolate whether the effect arises from turn boundaries, chat-template control tokens, or another consequence of conversational serialization.

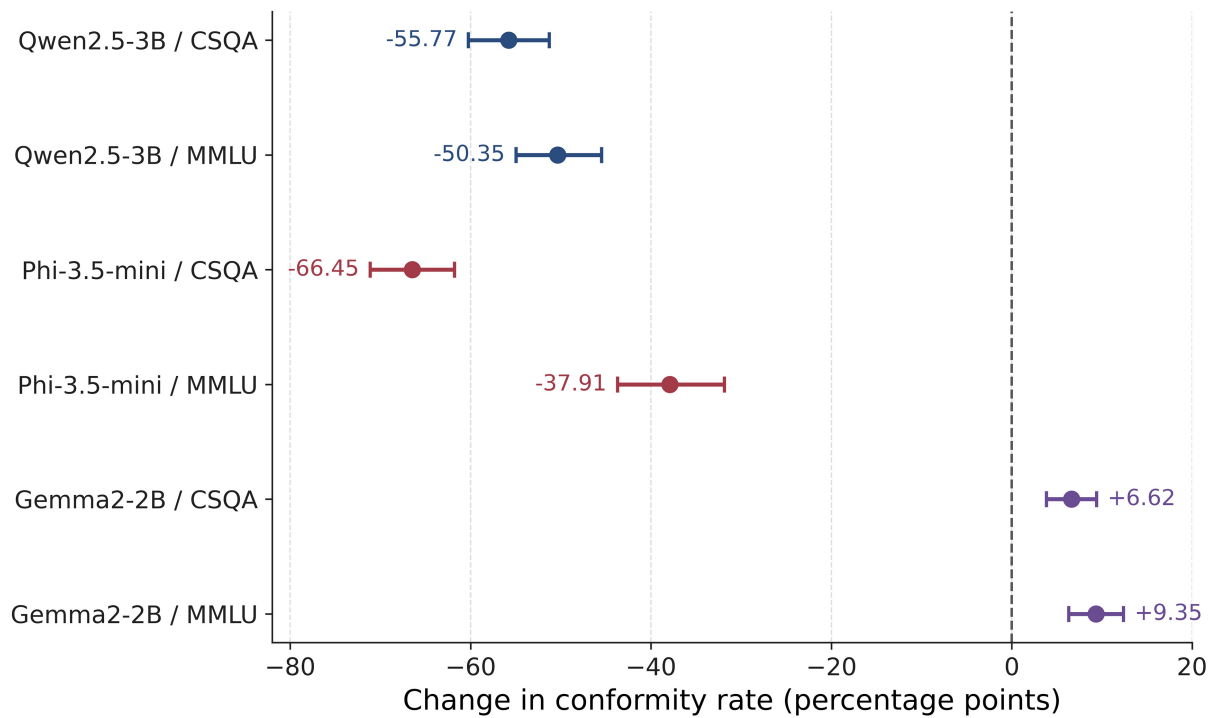


Figure 4. Exp2-to-Exp3 change in conformity rate (percentage points) with 95% paired-bootstrap confidence intervals. Negative values indicate lower wrong-target adoption under Sequential-Stepwise; positive values indicate higher adoption. CSQA: CommonsenseQA; MMLU: Massive Multitask Language Understanding.

Sequential-Stepwise produces a model-heterogeneous protocol effect

Finding 3

Stepwise responding makes Qwen and Phi substantially less likely to switch to the wrong peer-supported answer, but makes Gemma more likely to do so. Formally, CR falls from 58.76% to 2.99% for Qwen-CommonsenseQA and from 55.20% to 4.85% for Qwen-MMLU; Phi falls from 81.74% to 15.29% and from 66.74% to 28.84%. Gemma instead rises from 4.49% to 11.11% on CommonsenseQA and from 2.80% to 12.15% on MMLU. The contrast therefore establishes heterogeneous protocol responses across the evaluated models, not a universal benefit of intermediate answering.

Paired evidence

All six Exp2-Exp3 differences in wrong-target adoption remain significant after Holm correction. Figure 4 visualizes the paired CR differences and their 95% paired-bootstrap intervals: Qwen and Phi shift strongly downward from Exp2 to Exp3, whereas Gemma shifts upward. Tables 4 and 5 report the paired sample, formal CR denominator, bootstrap interval, discordant counts, exact McNemar *P* value, and adjusted *P* value for every main contrast. These statistics verify the difference but do not identify which changed protocol component caused it.

Is the protocol effect robust?

Repeated aligned subsets

Stable direction. Across 30 aligned 500-item draws, every model-dataset cell preserves the sign of the main Exp2-Exp3 difference: Qwen and Phi remain negative, Gemma remains positive [Table 6]. The empirical intervals describe sensitivity to item composition and do not replace the paired-bootstrap intervals for the fixed main sample.

Table 6. Exp2-Exp3 CR differences across 30 aligned subset draws

Model	Dataset	Mean Δ CR	Empirical 95% interval
Qwen	CSQA	-56.38 pp	[-59.62, -52.34]
Qwen	MMLU	-50.09 pp	[-52.80, -47.36]
Phi	CSQA	-66.17 pp	[-68.75, -63.76]
Phi	MMLU	-41.91 pp	[-45.39, -37.88]
Gemma	CSQA	+5.02 pp	[+3.35, +6.75]
Gemma	MMLU	+7.82 pp	[+5.46, +9.51]

(i) Δ CR is Exp3 minus Exp2; (ii) Each repetition samples 500 item identifiers without replacement within a dataset; repetitions may overlap; (iii) The same item set is reused across models and protocols; (iv) Intervals are empirical 2.5th-97.5th percentiles across repetitions. CR: Conformity rate; Δ CR: change in conformity rate; CSQA: CommonsenseQA; MMLU: Massive Multitask Language Understanding; pp: percentage points.

Table 7. Exp2-Exp3 conformity by target-distractor plausibility before peer exposure

Model-dataset	Low	Medium	High
Qwen-CSQA	N = 167; nCR = 167 49.10→1.20 Δ -47.90 [-55.09, -40.72]	N = 166; nCR = 166 51.20→1.81 Δ -49.40 [-57.23, -41.57]	N = 167; nCR=135 80.00→6.67 Δ -73.33 [-80.74, -65.93]
Qwen-MMLU	N = 167; nCR = 167 38.32→1.80 Δ -36.53 [-43.71, -29.34]	N = 166; nCR = 166 59.64→5.42 Δ -54.22 [-62.05, -46.39]	N = 167; nCR = 100 76.00→9.00 Δ -67.00 [-76.00, -58.00]
Phi-CSQA	N = 166; nCR = 166 75.90→11.45 Δ -64.46 [-72.29, -56.63]	N = 165; nCR = 164 82.32→13.41 Δ -68.90 [-76.83, -60.98]	N = 166; nCR = 141 87.94→21.99 Δ -65.96 [-74.47, -56.74]
Phi-MMLU	N = 160; nCR = 160 61.88→21.88 Δ -40.00 [-49.38, -30.63]	N = 159; nCR = 158 66.46→32.91 Δ -33.54 [-43.67, -23.42]	N = 159; nCR = 103 76.70→32.04 Δ -44.66 [-56.31, -32.04]
Gemma-CSQA	N = 167; nCR = 167 0.00→2.40 Δ +2.40 [+0.60, +4.79]	N = 166; nCR = 166 4.22→4.82 Δ +0.60 [-3.01, +4.22]	N = 167; nCR = 135 10.37→29.63 Δ +19.26 [+11.11, +27.41]
Gemma-MMLU	N = 167; nCR = 167 1.20→9.58 Δ +8.38 [+4.19, +13.17]	N = 166; nCR = 166 3.61→11.45 Δ +7.83 [+3.01, +13.25]	N = 166; nCR = 95 4.21→17.89 Δ +13.68 [+7.37, +21.05]

(i) Each cell reports tercile N, CR denominator nCR, Exp2→Exp3 CR (%), and Δ CR [95% paired-bootstrap CI] in percentage points; (ii) Terciles are defined within each model-dataset shared valid set from the normalized target-distractor probability before peer exposure; (iii) The 10,000-resample intervals are descriptive, unadjusted, and post hoc. CR: Conformity rate; nCR: number of samples eligible for CR calculation; Δ CR: change in conformity rate; CI: confidence interval; CSQA: CommonsenseQA; MMLU: Massive Multitask Language Understanding.

Target-distractor plausibility before peer exposure

Plausibility is not the main explanation. Qwen and Phi retain negative Exp2-Exp3 estimates within every relative target-plausibility tercile. Gemma is generally positive, with one CommonsenseQA medium-tercile estimate near zero [Table 7]. The split is therefore not created solely by one protocol receiving more initially plausible wrong targets. Because the stratification is post hoc, these intervals are descriptive and unadjusted.

Decomposing the Exp2-Exp3 difference

Intermediate calls alone are insufficient

Stepwise-No-History makes the extra intermediate calls but discards their outputs. Its final prompt is exactly identical to Sequential-Final, and it produces the same final CR as Sequential-Final in each evaluated mechanism setting. Extra calls alone are therefore insufficient to reproduce the Exp2-Exp3 difference.

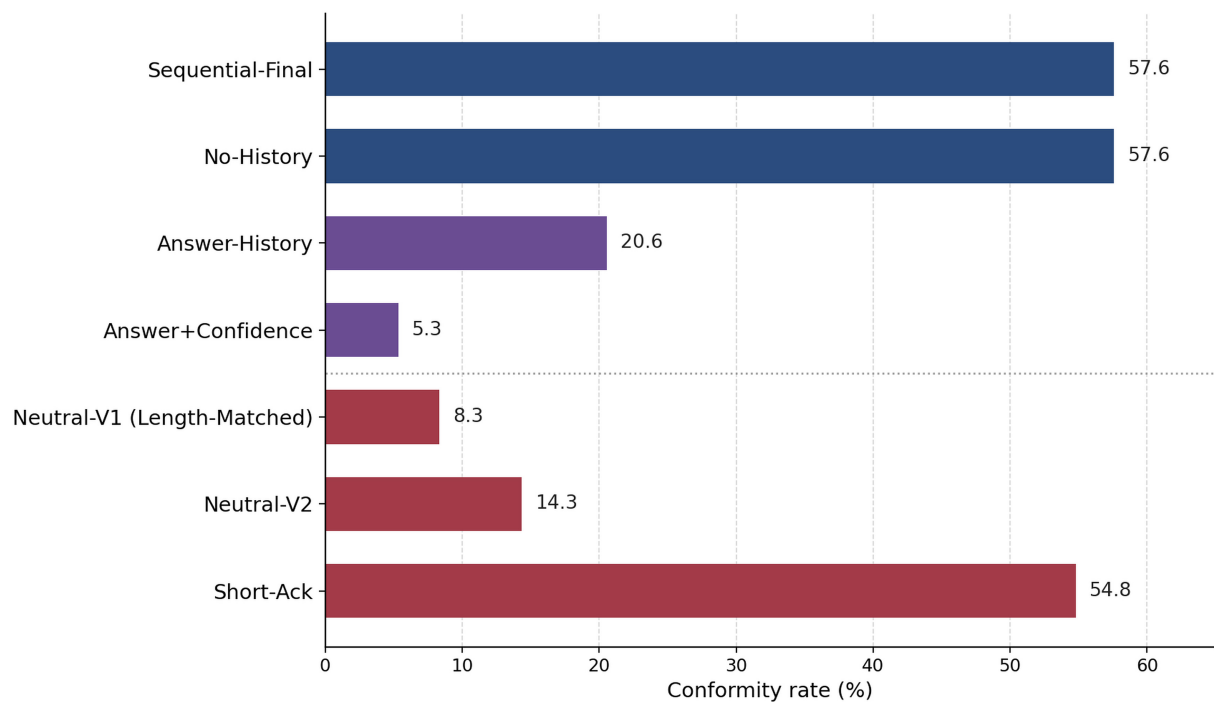


Figure 5. Seven-condition retained-history decomposition for Qwen-CommonsenseQA. Bars show direct CR for Sequential-Final, No-History, retained answer or confidence history, two independently worded token-matched neutral-history controls (Neutral-V1 and Neutral-V2), and a short acknowledgment control. CR: Conformity rate.

Table 8. Protocol-component decomposition for Qwen-CommonsenseQA

Condition	CR	HCR	FinalAcc	Target Adopt
Sequential-Final	57.60%	55.43%	33.00%	60.40%
No-History	57.60%	55.43%	33.00%	60.40%
Answer-History	20.56%	20.92%	56.40%	25.60%
Answer+Confidence	5.35%	4.89%	70.40%	11.20%
Neutral-V1	8.35%	8.15%	68.00%	14.40%
Neutral-V2	14.35%	12.77%	64.00%	20.00%
Short-Ack	54.82%	54.35%	33.60%	57.80%

(i) All seven conditions use the same 500 valid samples, with no invalid outputs. CR uses $N = 467$ items whose initial predictions were not already the target distractor; HCR uses $N = 368$ initially correct items. FinalAcc and TargetAdopt both use $N = 500$; (ii) The TargetAdopt numerator equals the CR numerator plus the number of initially target-selecting items that retained the target in the final response. These retained-target counts are 33, 33, 32, 31, 33, 33, and 33, respectively, following the row order in Table 8. All reported values have been verified against the original per-item records; (iii) Qwen neutral histories contain 13 tokens, matched on all 500 items. CR: Conformity rate; HCR: harmful conformity rate.

Retained history reproduces much of the Qwen effect

We next retain answer content, confidence-bearing history, token-matched neutral history, or a short acknowledgment. Direct wrong-target adoption is available for all three mechanism settings in the final analysis. Figure 5 focuses on the primary Qwen-CommonsenseQA CR pattern; Tables 8 and 9 report the primary decomposition and cross-setting direct-CR rates; complete cross-setting paired inference appears in Supplementary Table 8.

Finding 4

Retained conversational history, rather than intermediate calls alone, reproduces much of the Qwen protocol difference. On Qwen-CommonsenseQA, Neutral-V1 lowers CR by 49.25 points relative to the within-rerun Sequential-Final reference and Neutral-V2 by 43.25 points. The two neutral phrasings differ by 6.00 points,

Table 9. Cross-setting direct-CR replications

Condition	Qwen-MMLU CR	Gemma-CSQA CR
Sequential-Final	54.50%	4.49%
No-History	54.50%	4.49%
Answer-History	34.87%	2.56%
Answer+Confidence	19.86%	2.78%
Neutral-V1	14.55%	5.77%
Neutral-V2	19.40%	6.62%
Short-Ack	52.66%	7.48%

(i) This table reports direct CR on recovered common-valid denominators for Qwen-MMLU ($n = 433$) and Gemma-CommonsenseQA ($n = 468$); each underlying run contains 500 paired items; (ii) Qwen neutral histories contain 13 tokens and Gemma neutral histories 11 tokens, matched on all 500 items; (iii) The Sequential-Final values in this table are internal references from the separate mechanism rerun described in Section “Retained-history controls”. They should not be substituted for the archived main-run Exp2 endpoints in Table 4; all mechanism contrasts are within-rerun comparisons. CR: Conformity rate.

and Answer+Confidence does not consistently outperform both controls. The same broad retained-history pattern appears on Qwen-MMLU: matched multi-turn neutral histories strongly reduce CR, whereas No-History and Short-Ack remain close to the within-rerun Sequential-Final reference. Across both evaluated datasets, Qwen2.5-3B therefore shows a reproducible separation between matched multi-turn neutral history and a minimal assistant-role marker. The result supports sensitivity to retained conversational structure and wording, not a simple self-commitment or token-length law. [Supplementary Table 7](#) reports the paired denominator, bootstrap interval, exact McNemar P value, and Holm-adjusted P value for every primary CR contrast B-K.

A minimal assistant-role marker is insufficient

Short-Ack differs from Sequential-Final by only -2.78 points on Qwen-CommonsenseQA and is nonsignificant, so a minimal assistant-role marker does not reproduce the effect. On Qwen-MMLU, CR is 54.50% under Sequential-Final and No-History, falls to 14.55% and 19.40% under the two matched neutral histories, and remains 52.66% under Short-Ack; the matched-neutral reductions remain significant after Holm correction, whereas Short-Ack does not differ from Sequential-Final. Gemma-CommonsenseQA remains in a low absolute CR regime (4.49% at Sequential-Final; 2.56%-7.48% across the other conditions) and does not exhibit the large Qwen-like reduction. Several small pairwise differences are nevertheless detectable after Holm correction [[Supplementary Table 8](#)], so we do not interpret Gemma as showing no history effect. The low absolute rates and small differences limit stronger cross-model mechanism claims.

Does source-label wording alter wrong-target adoption?

Finding 5

In a separate contextual control, replacing “Source” with “Model” makes both Qwen and Gemma substantially less likely to adopt the wrong target. On CommonsenseQA, CR falls from 87.79% to 13.70% for Qwen and from 70.51% to 11.11% for Gemma. The same direction replicates on MMLU, where CR falls from 74.83% to 34.18% for Qwen and from 41.88% to 6.82% for Gemma. [Figure 6](#) summarizes the two-model, two-benchmark CR pattern; [Tables 10](#) and [11](#) report detailed paired accuracy, CR, and HCR inference for CommonsenseQA, and [Supplementary Table 9](#) reports the corresponding MMLU results. This controlled lexical substitution therefore has a large effect under these prompts, but it does not explain Exp2-Exp3 because the “Agent” label is fixed in both protocols.

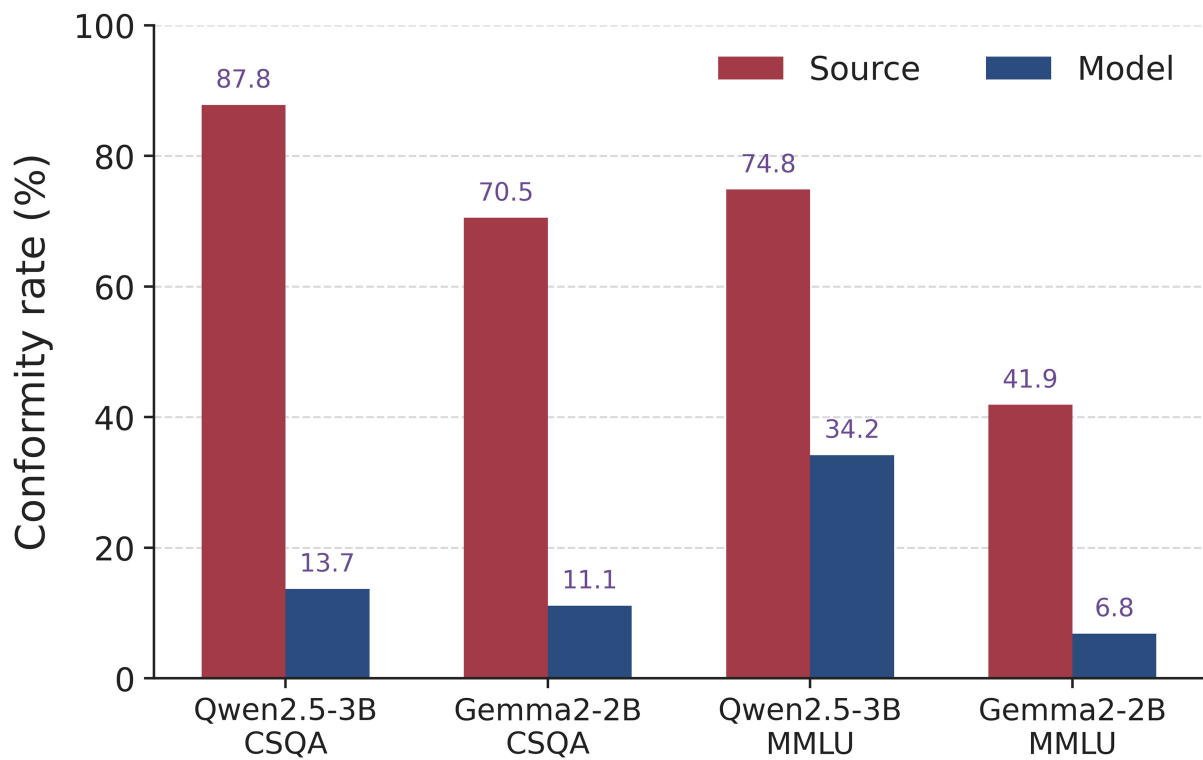


Figure 6. Conformity rate under token-matched Source-versus-Model label controls for Qwen2.5-3B and Gemma2-2B on CommonsenseQA and MMLU. Recommendation content, order, private baseline, and complete-chat token count are matched within each benchmark. CSQA: CommonsenseQA; MMLU: Massive Multitask Language Understanding.

Table 10. Source-label wording control: paired samples and condition outcomes on CommonsenseQA

Model	Metric	Pair <i>N</i>	Denom <i>N</i>	Source	Model
Qwen	Accuracy	500	500	5.20%	65.20%
	CR	500	467	87.79%	13.70%
	HCR	500	368	89.13%	11.68%
Gemma	Accuracy	500	500	22.00%	62.00%
	CR	500	468	70.51%	11.11%
	HCR	500	345	66.96%	9.86%

(i) Pair *N* is the number of paired items with valid outputs in both the Source and Model conditions; (ii) Denom *N* is 500 for Accuracy, the initially-not-target items for CR, and the initially-correct items for HCR. CR: Conformity rate; HCR: harmful conformity rate.

Targeted exploratory generalization probes

Checkpoint and peer-count probes

Susceptibility differs across the evaluated Qwen2.5 checkpoints and peer-count settings. On CommonsenseQA, Qwen2.5-3B has CR values of 1.07%, 66.38%, and 57.60% under 1, 3, and 5 unanimous wrong peers, respectively; thus, this checkpoint exhibits a non-monotonic pattern across the evaluated 1/3/5-peer settings. The evaluated Qwen2.5-7B checkpoint is already highly susceptible with one peer (85.89%) and reaches 100% at both 3 and 5 peers, a saturation pattern rather than the 3B non-monotonic profile. On the strict common-valid subset ($N = 450$), the 7B-minus-3B CR differences are +84.44, +34.22, and +42.89 percentage points for 1, 3, and 5 peers, respectively; all paired checkpoint-size contrasts remain significant after Holm correction. The Qwen2.5-3B five-peer value (57.60%) is the within-exploratory-run Sequential-Final endpoint for that peer-count series, not the archived main-run Exp2 estimate in [Table 4](#).

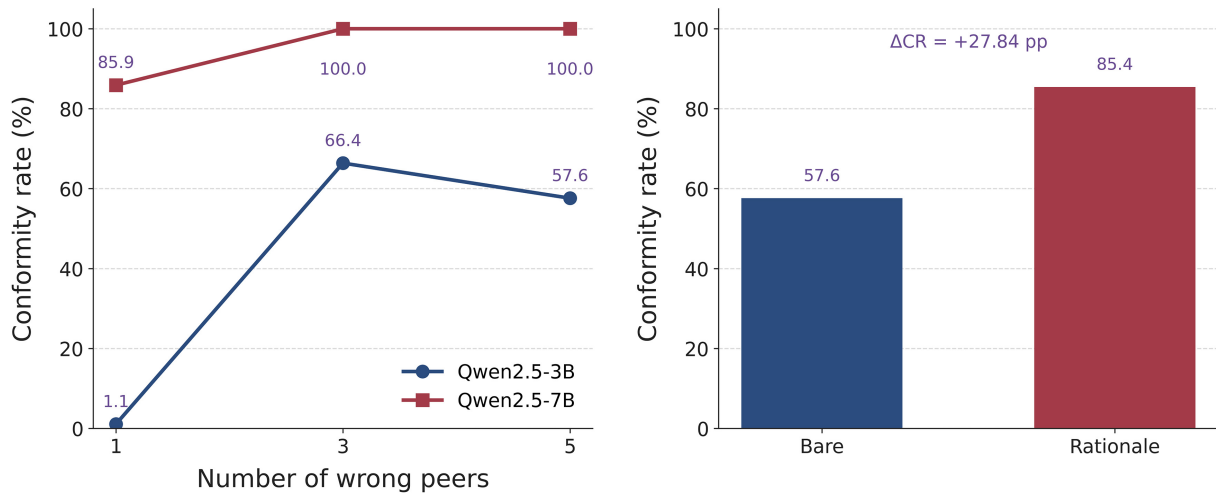


Figure 7. Targeted exploratory generalization probes within the Qwen2.5 family. Left: model-specific direct conformity rates for Qwen2.5-3B and Qwen2.5-7B under 1, 3, and 5 unanimous wrong peers on CommonsenseQA; formal checkpoint-size comparisons use the strict common-valid subset ($N = 450$) [Supplementary Table 10]. Right: Qwen2.5-3B conformity under bare answer-only peer signals versus the same target recommendations accompanied by a fixed generic justification.

Table 11. Source-label wording control: effect sizes and paired inference on CommonsenseQA

Model	Metric	Δ [95%CI]	b/c	Exact P	Holm P
Qwen	Accuracy	+60.00 [+55.60, +64.40]	3/303	7.33e-86	2.93e-85
	CR	-74.09 [-77.94, -70.02]	346/0	1.40e-104	5.58e-104
	HCR	-77.45 [-81.79, -73.10]	285/0	3.22e-86	1.29e-85
Gemma	Accuracy	+40.00 [+35.60, +44.20]	1/201	6.32e-59	1.89e-58
	CR	-59.40 [-63.68, -54.91]	278/0	4.12e-84	1.24e-83
	HCR	-57.10 [-62.32, -51.88]	197/0	9.96e-60	2.99e-59

(i) Δ is Model minus Source, in percentage points; (ii) b/c are discordant outcome counts (Source only/Model only); (iii) 95% confidence intervals use 10,000 paired-bootstrap resamples; (iv) Holm adjustment was performed separately for each outcome across four model-benchmark comparisons (Qwen2.5-3B and Gemma2-2B on CommonsenseQA and MMLU). This table reports the CommonsenseQA results; the corresponding MMLU results are provided in Supplementary Table 9. CR: Conformity rate; HCR: harmful conformity rate.

Generic-justification sensitivity

Adding the same short generic justification to each of the five peer recommendations increases Qwen2.5-3B CommonsenseQA CR from 57.60% to 85.44% ($\Delta CR = +27.84$ percentage points, 95% paired-bootstrap CI [+23.34, +32.33], exact McNemar $P = 2.29e-29$; $N = 467$). This augmentation jointly changes prompt content, prompt length, and repeated linguistic emphasis; it does not isolate reasoning quality, justification quality, or truthfulness. It therefore shows that this fixed generic-justification format can materially alter adoption under the evaluated prompts, not that rationales in general increase conformity. Taken together, Figure 7 reports checkpoint-specific and message-format sensitivity as secondary exploratory evidence. Complete paired estimates appear in Supplementary Table 10.

Interpretation

Interpretation

Interaction protocol is not a neutral wrapper around the same peer signals; it changes the effective conversational input seen by the model. The core evidence is the fixed-signal protocol dependence, the

Table 12. Positioning by focal estimand: fixed-signal and protocol/history comparisons

Study	Fixed wrong signal across focal contrast	Protocol comparison	Retained history isolated
Zhu et al. ^[7]	Not isolated	Partial	Not focal
Weng et al. ^[8]	Not isolated for history decomposition	Yes	Not focal
Ko et al. ^[9]	Not isolated	Collective dynamics	Not focal
Choi et al. ^[11]	Not focal	Debate setting	Not focal
Sun et al. ^[12]	Not focal	Not focal	Not focal
Present study	Yes	Yes	Yes

retained-history decomposition for Qwen2.5-3B across both evaluated datasets, and the separate token-matched Source-versus-Model sensitivity. The secondary Qwen2.5 probes further show that the evaluated 3B and 7B checkpoints differ sharply, that the 3B peer-count pattern is non-monotonic while the 7B series saturates, and that a fixed generic justification materially changes adoption. These exploratory patterns should be read as checkpoint- and prompt-specific contingencies, not as universal parameter-count, peer-dose, or rationale effects.

Mechanistic reading. The ablation results narrow the space of explanations without uniquely identifying an internal mechanism. No-History reproduces the Sequential-Final endpoint because the intermediate generations are discarded and the final prompt is identical, showing that additional model calls by themselves are insufficient. Retaining only a short acknowledgment is also insufficient in the primary Qwen setting. In contrast, retaining multi-turn assistant history - whether answer-bearing, confidence-bearing, or token-matched neutral content - can move behavior substantially toward the Sequential-Stepwise profile. Because the two neutral phrasings are not behaviorally identical, however, the result cannot be reduced to a simple token-count explanation. Nor does the design establish that earlier answers create a stable human-like commitment state. A more conservative interpretation is that retained conversational structure and wording alter the model's effective context, including the sequence of assistant-role tokens and previously generated content that conditions the final prediction. The model heterogeneity reinforces this interpretation: Gemma remains close to a low-conformity floor in the mechanism controls and moves in the opposite direction in the primary Exp2-Exp3 contrast, so a single model-independent narrative would overstate the evidence.

Relation to prior work

Conformity studies show that LLMs can shift toward majority answers and that susceptibility varies with uncertainty, input framing, majority size, and interaction configuration^[7-9]. Weng et al. already compare five interaction protocols^[8]. The contribution here is therefore not another demonstration that protocol matters, but a fixed-signal decomposition in which the targeted distractor and wrong-peer signals remain fixed while turn-segmented conversational formatting, retained history, and source-label wording are varied. Tables 12 and 13 contrast this design with closely related work.

Sycophancy concerns agreement with a user's expressed beliefs or preferences^[10], whereas identity and source-label work examines how who appears to provide a message changes trust or debate behavior^[11,12]. Our label control asks a narrower question: what happens to wrong-target adoption when five recommendations and complete prompt length are matched and only "Source" versus "Model" changes? This source-label wording effect is conceptually separate from the Exp2-Exp3 protocol contrast.

Multi-agent debate and self-refinement study whether exchanges among models or iterative self-feedback improve reasoning^[1,2,21]. Our peer signals are controlled text rather than independently acting agents, and the outcome is conditional adoption of a stored wrong target rather than aggregate debate accuracy. This

Table 13. Positioning by focal estimand: length controls, source labeling, and relevance

Study	Length-matched history control	Source-label wording isolated	Relevance to focal estimand
Zhu et al. ^[7]	Not focal	Not focal	Model and input factors in conformity
Weng et al. ^[8]	Not focal	Not focal	Protocol-level conformity comparison
Ko et al. ^[9]	Not focal	Not focal	Social-dynamics vulnerabilities
Choi et al. ^[10]	Not focal	No - identity cue instead	Identity effects in debate
Sun et al. ^[12]	Not focal	Yes	Source-label effects on trust
Present study	Yes	Yes	Protocol-component decomposition under fixed wrong-peer signals

(i) “Not isolated” means that the cited study does not hold the wrong signal fixed for the focal decomposition used here; “Not focal” means that the component is not isolated as a primary estimand; (ii) The table does not attempt to summarize every analysis in each cited study.

distinction enables protocol components to be studied without attributing behavior to latent peer quality or peer reasoning.

Practical implication. These distinctions matter for evaluation design as well as for deployment claims. A benchmark that reports only final accuracy can miss large shifts in which wrong answer is adopted, while a benchmark that changes both peer content and interaction schedule cannot determine whether a difference arises from evidence quality or conversational serialization. Reporting private answers, stored targets, final answers, conditional switching denominators, and exact message timelines makes the intervention auditable and allows failures to be localized. The present results also show why one protocol should not automatically be labeled “safer” from a single model: the same stepwise schedule that sharply lowers wrong-target adoption for Qwen and Phi raises it for Gemma in the primary evaluation. Protocol choice should therefore be treated as part of the evaluated system configuration. Where LLMs consume recommendations from tools, other models, or prior assistant turns, robustness testing should preserve the substantive recommendation while varying the surrounding conversational representation, rather than assuming that semantically similar wrappers are behaviorally equivalent.

Limitations and future work

Scope

The signals are controlled text, not independently executed agents^[1,9,21,22]. The primary protocol comparison covers three open-weight 2B-4B-class checkpoints on two fixed multiple-choice samples with five answer-only wrong-peer signals. Targeted exploratory generalization probes additionally evaluate Qwen2.5-7B, peer counts of 1, 3, and 5, and a fixed generic-justification condition, but these analyses remain concentrated within the Qwen2.5 family and CommonsenseQA. They therefore do not establish a universal parameter-count law, a general non-monotonic peer-count effect, or general behavior for open-ended tasks, substantive justifications, or other model families. The Source-versus-Model intervention remains a narrow two-label comparison. The MMLU cache is not subject stratified, repeated subsets overlap, plausibility strata are post hoc, and the original prompt contains a two-line-versus-ANSWER-only inconsistency. Exact checkpoint revisions were not preserved for every original main run. These limits bound the mechanism and generalization claims rather than invalidate the paired protocol effect.

Boundary of inference

The experiments establish behavioral dependence under controlled prompts, not a complete account of why that dependence arises inside the network. Prior work also shows substantial performance variation across instruction templates^[23], while chat-template structure can materially affect aligned model behavior^[24];

retained assistant turns can change both content and position; and checkpoint differences can include training data, alignment, tokenizer, and architecture-related factors in addition to parameter count. The current design intentionally does not disentangle all of these sources. Likewise, the target-plausibility analysis is best used to rule out a simple compositional explanation, not to establish a causal mediator, and two source labels cannot map the broader space of source identity, authority, or trust cues. Accordingly, retained history is described as sufficient to reproduce much of the Qwen pattern in the evaluated setting, the 3B peer-count pattern is described only for the tested 1/3/5-peer settings, and the checkpoint and generic-justification extensions remain exploratory.

Next tests

The highest-value extensions are to replicate the checkpoint-size, peer-count, and generic-justification findings across additional model families and open-ended tasks; vary justification informativeness and correctness under controlled length; use subject-stratified MMLU sampling; rerun a core setting with harmonized output instructions; and archive exact checkpoint revisions and complete per-item outputs for every future replication. These tests would distinguish family-specific sensitivities from broader interaction effects.

CONCLUSIONS

This study asks what changes when identical wrong-peer signals are embedded in different conversations. Interaction protocol is not a neutral wrapper around those signals: stepwise responding makes Qwen and Phi less likely to switch to the wrong peer-supported answer but makes Gemma more likely to do so. For Qwen2.5-3B, intermediate calls or a short assistant marker are insufficient, whereas multi-turn token-matched neutral assistant history reproduces much of the reduction in wrong-target adoption across both evaluated datasets; Gemma does not exhibit the same large reduction. In a separate contextual control, replacing “Source” with “Model” sharply suppresses wrong-target adoption under matched prompts on both benchmarks; this source-label wording effect does not explain Exp2-Exp3, where “Agent” remains fixed. Secondary Qwen2.5 probes show additional sensitivity to checkpoint choice, peer count, and generic-justification format. Together, the core results establish fixed-signal protocol dependence, a retained-history decomposition, and source-label contextual sensitivity, with all effects remaining model- and setting-dependent.

DECLARATIONS

Authors’ contributions

Made substantial contributions to the conception and design of the study, implementation, data analysis, and drafting: Xu, L.; Xue, X.

Contributed to methodological discussion, interpretation, and manuscript revision: Zhu, Y.; Zhang, J.

Supervised the study, contributed to research design and interpretation, and critically revised the manuscript: Xu, D.

All authors approved the final manuscript and accept accountability for the work.

Availability of data and materials

Code, prompt builders, parser and analysis scripts, processed dataset definitions, checksums, and reproducibility instructions are available in the public repository at the immutable audited commit <https://github.com/xuexucheng/Social-Conformity-in-Large-Language-Models/commit/e8a1b7d2e6a65f913787ead1c41d1699996752d5>. Raw model-output JSONL files, forced-option score JSONLs, and machine-readable revision outputs are available from the corresponding author on reasonable request. CommonsenseQA and MMLU remain subject to their original licenses.

AI and AI-assisted tools statement

During the preparation of this manuscript, the AI tools ChatGPT (version GPT-5.6 Sol, released 2026-07-09) and Codex were used to improve language and readability, reorganize author-written material, perform non-

substantive consistency checks, and assist with the visual refinement and layout of [Figures 1](#) and [2](#) and the graphical abstract. The underlying scientific content, experimental design, data, results, and interpretations in these figures and the GA were created and verified by the authors. All AI-assisted visual materials were reviewed, edited, and finalized by the authors, who take full responsibility for the published content. The language models evaluated as experimental subjects are identified, versioned, and described separately in the Experimental section. The tools did not influence the study design, data collection, analysis, interpretation, or the scientific content of the work. All authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

Financial support and sponsorship

None.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

Supplementary Materials

[Supplementary Materials](#)

REFERENCES

1. Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*; PMLR, 2024; Vol. 235, pp 11733-63. Available online: <https://proceedings.mlr.press/v235/du24e.html>. (accessed 14 October 2026).
2. Madaan, A.; Tandon, N.; Gupta, P.; et al. Self-refine: iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36*, New Orleans, Louisiana, USA, December 10-16, 2023; Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023, pp 46534-94. DOI
3. Yao, S.; Zhao, J.; Yu, D.; et al. ReAct: synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. Available online: https://openreview.net/forum?id=WE_vluYUL-X. (accessed 14 October 2026).
4. Asch, S. E. Effects of group pressure upon the modification and distortion of judgments. In *Groups, Leadership and Men: Research in Human Relations*; Guetzkow H., Ed.; Carnegie Press: Pittsburgh, PA, USA, 1951; pp 177-90. Available online: <https://ndlsearch.ndl.go.jp/en/books/R100000136-I1574231874449715712>. (accessed 14 October 2026).
5. Deutsch, M.; Gerard, H. B. A study of normative and informational social influences upon individual judgment. *J. Abnorm. Soc. Psychol.* **1955**, *51*, 629-36. DOI PubMed
6. Cialdini, R. B.; Goldstein, N. J. Social influence: compliance and conformity. *Annu. Rev. Psychol.* **2004**, *55*, 591-621. DOI PubMed
7. Zhu, X.; Zhang, C.; Stafford, T.; Collier, N.; Vlachos, A. Conformity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria, July, 2025; Association for Computational Linguistics, 2025, pp 3854-72. DOI
8. Weng, Z.; Chen, G.; Wang, W. Do as we do, not as you think: the conformity of large language models. In *International Conference on Learning Representations*, 2025. Available online: <https://openreview.net/forum?id=st77ShxP1K>. (accessed 14 October 2026).
9. Ko, C.; Shin, J.; Song, H.; Lee, H.; Hwang, E. J.; Park, J. C. 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)*, San Diego, California, United States, July, 2026; Association for Computational Linguistics, 2026, pp 37865-90. DOI
10. Sharma, M.; Tong, M.; Korbak, T.; et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024. Available online: <https://openreview.net/forum?id=tvhaxkMKAn>. (accessed 14 October 2026).

11. Choi, H. K.; Zhu, J.; Li, S. When identity skews debate: anonymization for bias-reduced multi-agent reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, San Diego, California, United States, July, 2026; Association for Computational Linguistics, 2026, pp 14284-311. DOI
12. Sun, X.; Wu, D.; Qin, S.; Echizen, I.; El Ali, A.; Sugawara, S. Label effects: shared heuristic reliance in trust assessment by humans and LLM-as-a-judge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, San Diego, California, United States, July, 2026; Association for Computational Linguistics, 2026, pp 32378-92. DOI
13. Talmor, A.; Herzig, J.; Lourie, N.; Berant, J. CommonsenseQA: a question answering challenge targeting commonsense knowledge. 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)*, Minneapolis, Minnesota, June, 2019; Association for Computational Linguistics, 2019, pp 4149-58. DOI
14. Hendrycks, D.; Burns, C.; Basart, S.; et al. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. Available online: <https://openreview.net/forum?id=d7KBjmI3GmQ>. (accessed 14 October 2026).
15. Yang, A.; Yang, B.; Zhang, B.; et al. Qwen2.5 technical report. *arXiv* **2024**, arXiv:2412.15115. Available online: <https://doi.org/10.48550/arXiv.2412.15115> (accessed 14 October 2026).
16. Microsoft. Phi-3.5-mini-instruct model card. Hugging Face. Available online: <https://huggingface.co/microsoft/Phi-3.5-mini-instruct>. (accessed 14 October 2026).
17. Riviere, M.; Pathak, S.; Sessa, P. G.; et al. Gemma 2: improving open language models at a practical size. *arXiv* **2024**, arXiv:2408.00118. Available online: <https://doi.org/10.48550/arXiv.2408.00118> (accessed 14 October 2026).
18. Efron, B.; Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.* **1986**, *1*, 54-75. DOI
19. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153-7. DOI PubMed
20. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65-70. Available online: <https://www.jstor.org/stable/4615733> (accessed 14 October 2026).
21. Li, G.; Hammoud, H.; Itani, H.; Khizbullin, D.; Ghanem, B. CAMEL: communicative agents for “Mind” exploration of large language model society. In *Advances in Neural Information Processing Systems 36*, New Orleans, Louisiana, USA, December 10-16, 2023; Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023, pp 51991-2008. DOI
22. Wu, Q.; Bansal, G.; Zhang, J.; et al. AutoGen: enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024. Available online: <https://openreview.net/forum?id=BAakY1hNKS>. (accessed 14 October 2026).
23. Mizrahi, M.; Kaplan, G.; Malkin, D.; Dror, R.; Shahaf, D.; Stanovsky, G. State of what art? A call for multi-prompt LLM evaluation. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 933-49. DOI
24. Jiang, F.; Xu, Z.; Niu, L.; Lin, B. Y.; Poovendran, R. ChatBug: a common vulnerability of aligned LLMs induced by chat templates. *Proc. AAAI. Conf. Artif. Intell.* **2025**, *39*, 27347-55. DOI

Disclaimer/Publisher’s Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.



© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.