# HUMAN HISTORY HELPS—BUT FOR HOW LONG? TRACKING LLM DIALOGUE CONTINUATION ACROSS TURNS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Can human response history keep language-model continuations close to recorded human responses—and for how long after new human responses stop being supplied? We introduce a controlled dialogue-replay protocol that fixes one participant's recorded utterances while a language model generates the target participant's response at each eligible turn. Five conditions vary the target-speaker history available to the model: no recorded human responses, the first one, two, or five human responses followed by model-generated responses, or all previous human responses. Earlier human responses remain in context after each finite prefix ends, while the current human reference is never exposed. We evaluate five language models across informal conversation (MaiChat), emotional support (ESConv), and persuasion (PersuasionForGood), with over 630,000 model-generated continuations saved before matched-window filtering. MaiChat and ESConv use dialogue history alone, whereas PersuasionForGood additionally includes fixed task context without strategy labels. We measure turn-level semantic distance to recorded human continuations and estimate condition-dependent trajectories over matched dialogue windows using mixed-effects models. Across all 15 model–dataset analyses, full human history has a lower fitted window-averaged semantic distance than the condition without recorded human response history, although its fitted response-step trend remains positive. A five-response human prefix retains a detectable benefit over that condition after model-generated responses enter the history in ESConv and PersuasionForGood. Yet continued human responses provide an additional detectable benefit over that prefix at one or more post-prefix positions for all five PersuasionForGood models and four of five ESConv models. Effects in MaiChat are less precise and more model-dependent. These findings distinguish the carryover of early human responses from the benefit of continually supplying new ones. Our protocol measures fidelity to recorded continuations under fixed interlocutor utterances, not general response quality, task success, or fully interactive dialogue.

## 1 INTRODUCTION

Large language models (LLMs) are increasingly used as conversational systems. In multi-turn dialogue, a generated response can become context for later responses. Yet evaluating each position using only the recorded human response history does not expose the model to this self-generated context. A score obtained under recorded history therefore need not characterize continuation after the model repeatedly conditions on its own outputs.

Prior multi-turn benchmarks assess two-turn questions, follow-up instructions, and tasks whose information is distributed across an interaction (Zheng et al., 2023; Kwan et al., 2024; Laban et al., 2026). Closer to our setting, Bai et al. (2024) compare golden and self-predicted dialogue histories. We examine the intermediate case in which only an initial segment of target-speaker history is human-written.

We investigate this question through *controlled dialogue replay*. For each recorded two-party dialogue, one participant's utterances remain fixed while a model predicts the target participant's re-



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.

sponse at every eligible position. Five conditions vary which earlier target-speaker responses enter subsequent histories: model-generated responses throughout (C0); the first one, two, or five recorded human responses followed by model-generated responses (C1, C2, and C5); or all previous recorded human responses (CALL). The model generates a prediction at every position in every condition; the conditions determine which responses are inserted into later histories. Supplied human responses remain in context after a finite prefix ends. The current human reference is never shown to the model, and any target-speaker opening is shared across conditions.

The conditions enable two complementary comparisons after the C5 prefix ends. C5 versus C0 asks whether an early human prefix retains an advantage once model-generated responses enter the history; CALL versus C5 asks whether continuing to supply human responses provides an additional advantage. Both compare complete history-construction policies under the same recorded interlocutor sequence.

We evaluate five language models across three human–human dialogue settings: informal conversation from MaiChat (Dao et al., 2026), emotional support from ESConv (Liu et al., 2021), and persuasion from PersuasionForGood (Wang et al., 2019). MaiChat and ESConv use dialogue history alone, whereas PersuasionForGood additionally includes fixed task context without turn-level strategy annotations.

We compare each generated response with the corresponding recorded human response using semantic distance and estimate trajectories over matched dialogue windows. Because a context can admit multiple appropriate replies, this outcome measures fidelity to one recorded continuation, not general response quality or task success (Liu et al., 2016; Yeh et al., 2021). The fixed interlocutor sequence makes this a controlled replay rather than a fully interactive conversation.

Across the 15 model–dataset analyses, full human history yields a positive window-averaged benefit relative to C0 in every analysis, although reference distance also rises with response position under full human history. A five-response prefix can remain beneficial after model-generated responses enter the history. Yet directly comparing C5 with CALL reveals an additional detectable benefit from continuing to supply human responses at one or more post-prefix positions for all five PersuasionForGood models and four of five ESConv models. MaiChat shows less precise and more model-dependent effects. These findings distinguish the carryover of an initial human prefix from the benefit of continued human history, without treating the last detectable step as an exact endpoint of influence.

Our contributions are threefold:

- We introduce a controlled dialogue-replay protocol that varies the source and amount of target-speaker history while fixing the recorded interlocutor sequence and prediction target.
- Across five models and three dialogue settings, we generated and saved more than 630,000 responses. The matched-window analyses use the retained, model-specific subsets of these predictions.
- We show that finite human history can carry beyond its supplied prefix, but continued human history can provide an additional benefit; rising reference distance is not unique to self-generated history.

## 2 Related Work

### 2.1 Multi-Turn Evaluation of Language Models

Multi-turn benchmarks reveal limitations that isolated prompts can miss. MT-Bench uses two-turn questions and LLM-based judging (Zheng et al., 2023); MT-Eval probes recollection, expansion, refinement, and follow-up (Kwan et al., 2024); and BotChat evaluates model-generated multi-turn conversations using an LLM-centric protocol (Duan et al., 2024). Laban et al. (2026) show that distributing task information across an interaction can impair performance.

Bai et al. (2024) compare golden and self-predicted dialogue histories for one model in two MT-Bench-101 tasks, showing that history source can change turn-level scores. Our study examines a different, intermediate regime: recorded human responses are supplied for one, two, or five initial

target-speaker positions before model-generated responses enter the history. We conduct this comparison in recorded human–human dialogues, with the interlocutor sequence and reference response fixed at each evaluated position.

### 2.2 Self-Generated Context

Conditioning on observed rather than model-generated sequences is related to exposure bias in autoregressive generation. Under conventional likelihood training, a model receives ground-truth preceding tokens but conditions on its own predictions at inference time. Scheduled sampling and Professor Forcing address aspects of this discrepancy through changes to training (Bengio et al., 2015; Lamb et al., 2016). Schmidt (2019) cautions against assuming that the discrepancy necessarily produces compounding error. Comparing observed and generated text prefixes at inference time, He et al. (2021) find limited, nonincremental distortions and evidence of self-recovery in open-ended generation. They also examine how the effect of an altered prefix changes after an intervening generated span.

Our intervention instead varies complete *earlier dialogue responses*, not a prefix of the current answer. We neither modify training nor treat the experiment as a direct test of token-level exposure bias. The current human reference is never exposed, and human responses supplied by a finite prefix remain in context as later model-generated responses accumulate. Any attenuation therefore occurs without explicitly removing those responses from the input.

### 2.3 Reference-Based Evaluation of Dialogue

A dialogue context can admit multiple appropriate continuations, making similarity to one recorded response an incomplete measure of dialogue quality. Multiple human references can improve the agreement of automatic metrics with human judgments in open-domain dialogue (Gupta et al., 2019). Word-overlap metrics have shown weak correlation with human judgments (Liu et al., 2016); contextualized-embedding methods address some lexical mismatch (Ghazarian et al., 2019). Nevertheless, dialogue metrics differ in reliability and in the properties they capture (Yeh et al., 2021), while reference-free approaches assess additional response properties (Mehri & Eskenazi, 2020).

These limitations motivate a deliberately narrow, paired outcome: at a given dialogue position, all conditions are compared with the same recorded reference. The measure identifies differences in fidelity to that continuation; it cannot determine which of several plausible replies a human would prefer.

## 3 Controlled Dialogue-Replay Protocol

### 3.1 Problem Formulation

Let a recorded two-party dialogue be represented as a sequence of alternating speaker runs, where a run groups consecutive utterances by the same speaker. Let $a_t$ denote the recorded interlocutor run preceding the $t$-th eligible target-speaker response position, and let $b_t$ denote the corresponding recorded human target-speaker run. Let $O$ denote an optional recorded target-speaker opening that precedes the first eligible response position; $O = \varnothing$ when no such opening exists.

At every eligible position, the evaluated language model generates

$$\hat{b}_t^{(c)} \sim p_\theta \left( \cdot \mid P, X, H_t^{(c)} \right), \tag{1}$$

where $c$ denotes the history condition, $P$ is the shared prompt configuration, $X$ contains any fixed task context, and $H_t^{(c)}$ is the dialogue history reconstructed under condition $c$. The model and generation parameters are held fixed across conditions.

The history used to predict position $t$ is

$$H_t^{(c)} = \left( O, a_1, \tilde{b}_1^{(c)}, \ldots, a_{t-1}, \tilde{b}_{t-1}^{(c)}, a_t \right), \tag{2}$$

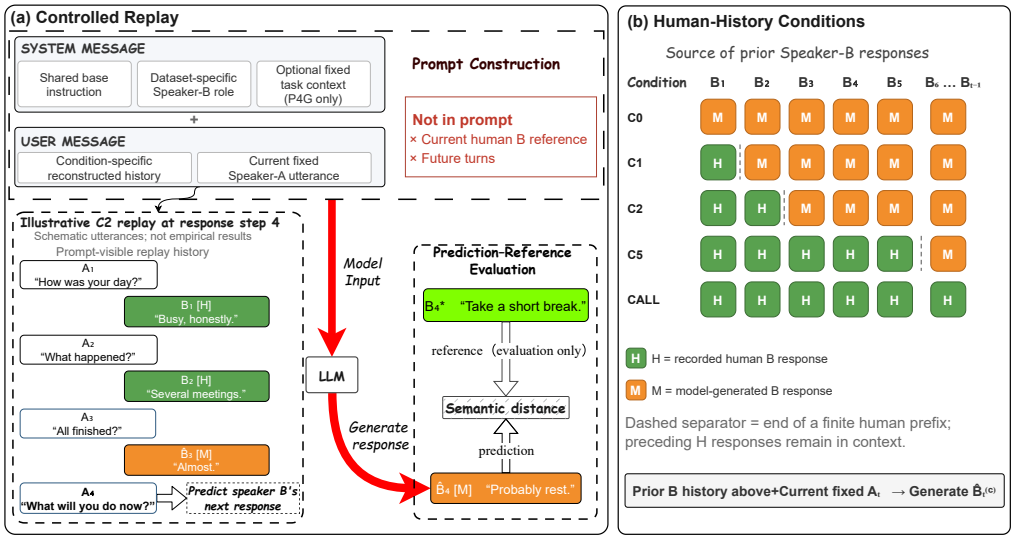



Figure 1: **Controlled dialogue replay and history conditions.** **(a)** At response position $t$, the model receives shared instructions, a dataset-specific target role, optional fixed task context, the condition-specific history, and the current fixed Speaker-A run $a_t$. The generated response $\hat{b}_t^{(c)}$ is compared offline with the hidden reference $b_t$. The example shows C2 at $t = 4$. **(b)** Sources of previous target-speaker responses. H denotes a recorded human response and M a model-generated response. Dashed separators mark the end of finite human prefixes; previously supplied human responses remain in context.

where an empty $O$ is omitted and $\tilde{b}_i^{(c)}$ is the response inserted at an earlier target-speaker position. Depending on the condition, this response is either the recorded human response $b_i$ or the response $\hat{b}_i^{(c)}$ previously generated under the same condition.

The current recorded response $b_t$ and all future utterances are excluded from $H_t^{(c)}$ and all other model inputs. The response $b_t$ is used only for offline evaluation of $\hat{b}_t^{(c)}$. Figure1 illustrates the prompt construction, generation and evaluation procedure, and five history conditions.

### 3.2 Human-History Conditions

The five conditions determine the source of earlier target-speaker responses inserted into the history (Figure 1b). C0 inserts model-generated responses throughout. C1, C2, and C5 insert the first one, two, or five recorded human responses, respectively, followed by responses generated under that condition. CALL inserts recorded human responses at every previous eligible position.

The model still generates and evaluates a prediction at every eligible position under every condition; the conditions determine which response enters subsequent histories. For example, under C2, predictions at positions 1 and 2 are evaluated, but the corresponding recorded human responses are inserted into later histories. Ending a finite prefix does not remove those earlier human responses from context; it stops only the insertion of new ones.

C0 does not imply an entirely model-generated dialogue: all interlocutor runs remain recorded and fixed. An optional recorded target-speaker opening $O$ is shared across conditions and is not counted toward C1, C2, or C5.

### 3.3 Controlled Comparison and Scope

For a given model, dialogue, and target position, conditions differ only in the source of previous target-speaker responses. The current Speaker-A utterance, prediction target, prompt components, and generation settings are otherwise held fixed.

The protocol is a controlled replay rather than a simulation of a freely evolving conversation. Speaker-A utterances remain fixed to their recorded sequence and do not adapt to model-generated Speaker-B responses. A fixed Speaker-A utterance may therefore be less compatible with a model-generated history than with the original human response history. The resulting estimand is the model's fidelity to recorded target-speaker continuations under a fixed interlocutor sequence. It does not directly measure performance in an interactive conversation with an adaptive human or user simulator.

# 4 Experimental Setup

## 4.1 Datasets and Target Roles

We evaluate three English human–human dialogue datasets. For MaiChat, an informal-conversation dataset, we treat the second speaker as the target (Dao et al., 2026). For ESConv, we generate the emotional supporter's responses (Liu et al., 2021). For PersuasionForGood, we generate the persuader's responses in donation-oriented dialogues (Wang et al., 2019). We use the 42-dialogue MaiChat release available to our pipeline and an expanded 1,300-dialogue release of ESConv. Table 1 summarizes dataset coverage and saved predictions. Further release, preprocessing, and participant details appear in Appendix A.1, including Table 3.

Table 1: Dataset coverage and saved predictions. Eligible positions precede model-specific runtime exclusions; saved predictions sum over five models and five history conditions before matched-window filtering.

| Dataset | Target speaker | Dialogues | Eligible positions | Saved predictions |
|---|---|---|---|---|
| MaiChat | Second speaker | 42 | 1,425 | 35,523 |
| ESConv | Supporter | 1,300 | 14,642 | 364,683 |
| PersuasionForGood | Persuader | 1,017 | 9,583 | 239,399 |

Consecutive messages from the same speaker are grouped into speaker runs in each dataset (Appendix A.1). An eligible position corresponds to a recorded target-speaker run preceded by an interlocutor run. Each saved prediction is one model-generated response for that position under one model and one history condition; it does not count an original corpus message.

## 4.2 Prompt Construction

Each prompt combines a shared base instruction, a dataset-specific target-role description, the condition-specific reconstructed history, and the current recorded Speaker-A run. PersuasionForGood additionally includes fixed task context. The current human reference, future utterances, and turn-level strategy annotations are excluded from the model input. Complete prompt templates and role blocks are provided in the appendix A.2.

## 4.3 Models and Generation

We evaluate GPT-4.1-nano (OpenAI, 2025), DeepSeek-V4-Flash (DeepSeek-AI et al., 2026), Qwen3.8-Flash (Alibaba Cloud, 2026), Gemini 3.5 Flash-Lite (Google, 2026), and Claude Haiku 4.5 (Anthropic, 2025). All runs use temperature zero. Optional reasoning is disabled for DeepSeek, Qwen, and Claude, while Gemini uses minimal thinking. API routes, and requested and returned model identifiers are reported in the appendix A.3.

## 4.4 Semantic-Distance Measure

For each eligible position, we encode the generated response $\hat{b}_t^{(c)}$ and recorded human response $b_t$ using `sentence-transformers/paraphrase-multilingual-mpnet-base-v2` (Reimers & Gurevych, 2019; Sentence Transformers, 2021). After normalizing the embeddings, we define semantic distance as

$$d_t^{(c)} = 1 - \cos\left(e\left(\hat{b}_t^{(c)}\right), e\left(b_t\right)\right), \tag{3}$$

where lower values indicate greater similarity to the recorded continuation. This measure captures fidelity to one recorded response, rather than general dialogue quality or task success.

### 4.5 Matched-Window Mixed-Effects Analysis

For the trajectory analysis, we form model-specific matched cohorts over response steps 1–10 for MaiChat and ESConv and steps 1–9 for PersuasionForGood. A separate MaiChat analysis over steps 1–20 is reported in Appendix B.6. A conversation is retained only if all five conditions have valid predictions at every position in the selected window. The cohort is therefore fixed across response steps within each analysis. Retained cohorts and runtime exclusions are reported in Appendix A.4.

We then fit a separate mixed-effects model for each dataset, model, and matched window. Response step is centered within the window, and the fixed-effects specification is

$$d \sim \text{condition} \times \left(\text{step}_c + \text{step}_c^2\right). \tag{4}$$

The quadratic term allows condition trajectories to vary nonlinearly over response positions.

Models include random intercepts for conversation and matched response position. A conversation-level step slope is retained when supported by convergence and singularity checks, and PersuasionForGood models may additionally include a target-speaker intercept. Complete random-effects specifications and diagnostics are reported in Appendix A.5.

We fit the models using `lme4` and `lmerTest` (Bates et al., 2015; Kuznetsova et al., 2017). Estimated marginal means, condition contrasts, and trend contrasts are computed with `emmeans` (Lenth & Piaskowski, 2026). Holm's procedure adjusts $p$-values within each predefined comparison family (Holm, 1979). At step 1, all conditions use the same input history, so any difference in their predictions cannot be attributed to the history intervention. We interpret condition contrasts only at steps where the compared histories can differ.

## 5 Results

Unless otherwise stated, results refer to the primary matched windows: steps 1–10 for MaiChat and ESConv and steps 1–9 for PersuasionForGood. Appendix A.4.1 (Table 7) reports the number of conversations with complete five-condition trajectories at these and longer response-step cutoffs.

We begin by asking whether reference distance rises even under fully human-written target-speaker history. We then address the central post-prefix question with two comparisons: C5 versus C0 for the carryover of a finite prefix, and CALL versus C5 for the additional benefit of continued human responses. Finally, we examine consistency across models and dialogue settings.

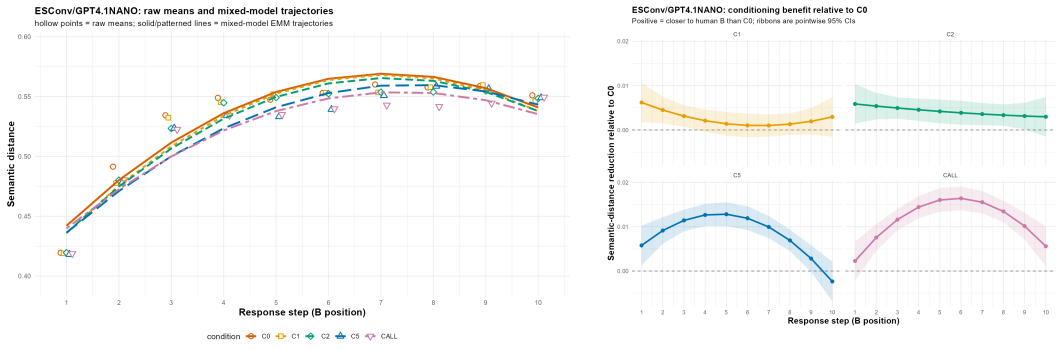
### 5.1 Reference Distance Rises Even Under Full Human History

Across the 15 model–dataset analyses and five history conditions, all 75 fitted window-center trends are positive; 74 remain detectable after Holm adjustment. The sole exception is C0 for Claude Haiku 4.5 in MaiChat, whose trend remains positive but does not reach the adjusted threshold ($p_{\text{adj}} = .068$). Because the fitted models include a quadratic step term, these positive centered trends do not imply a strictly monotonic increase at every adjacent response position.

Under C0, the fitted semantic distance at the final primary-window step exceeds that at step 1 in all 15 model–dataset analyses. The fitted endpoint increase ranges from 0.043 to 0.112 in MaiChat, from 0.099 to 0.140 in ESConv, and from 0.061 to 0.085 in PersuasionForGood.

CALL has a positive fitted window-averaged benefit relative to C0 in all 15 analyses (Table 10), yet its window-center trend is positive and Holm-detectable in all 15 (Table 9). The adjusted C0–CALL trend contrast is detectable in six analyses: five show a smaller trend under CALL, whereas PersuasionForGood with GPT-4.1-nano shows a larger one.

Figure 2 illustrates these patterns for GPT-4.1-nano in ESConv: the fitted C0 distance rises from 0.442 at step 1 to 0.541 at step 10, while the condition contrasts show the relative benefits of human history. Corresponding primary-window figures for all other model–dataset combinations appear in Appendix B.




**(a) Raw means and estimated trajectories** **(b) Benefit relative to C0**

Figure 2: **GPT-4.1-nano results in ESConv, response steps 1–10.** (a) Raw condition means and mixed-effects estimated trajectories. (b) Estimated reductions in semantic distance relative to C0; positive values indicate greater similarity to the recorded human response. Shaded ribbons show pointwise 95% confidence intervals; statistical detectability reported in the text is based on Holm-adjusted tests.

Taken together, these results show that reference distance can rise even when all previous target-speaker responses are human-written. Human history more consistently lowers window-averaged distance than flattens the response-step trajectory.

### 5.2 Benefits Scale with the Amount of Human History

We summarize fitted benefit magnitude separately from step-specific statistical detectability. For each model–dataset analysis, we average the fitted contrast $d_{C0} - d_c$ over its primary window. Table 2 reports the equal-weight mean of these model-specific summaries for each dataset, together with the number of models exhibiting at least one positive, Holm-adjusted post-step-1 contrast.

Table 2: **Magnitude and cross-model detectability of human-history benefits.** Each cell reports the equal-weight mean of the model-specific primary-window fitted benefit $d_{C0} - d_c$, followed in brackets by the number of models with at least one positive post-step-1 contrast surviving Holm adjustment. The overall row gives equal weight to all 15 model–dataset analyses. These means are descriptive rather than pooled estimates; a bracketed count of zero does not establish equivalence to C0.

| Dataset | C1 | C2 | C5 | CALL |
|---|---|---|---|---|
| MaiChat | 0.0050 [0/5] | 0.0110 [1/5] | 0.0152 [2/5] | 0.0225 [2/5] |
| ESConv | 0.0024 [0/5] | 0.0035 [3/5] | 0.0098 [5/5] | 0.0144 [5/5] |
| PersuasionForGood | 0.0040 [5/5] | 0.0075 [5/5] | 0.0148 [5/5] | 0.0191 [5/5] |
| Overall | 0.0038 [5/15] | 0.0073 [9/15] | 0.0133 [12/15] | 0.0187 [12/15] |

The descriptive magnitudes show the same ordering in the five-model mean for every dataset:

$$\text{CALL} > \text{C5} > \text{C2} > \text{C1}.$$

At the individual model–dataset level, the window-average benefit is positive in 13 of 15 analyses for C1, 15 of 15 for C2, 14 of 15 for C5, and 15 of 15 for CALL. The ordering is therefore a strong aggregate pattern rather than a rule that holds exactly for every individual analysis.

Step-specific contrasts show that finite human prefixes can carry beyond the positions at which human responses are supplied. For GPT-4.1-nano in ESConv, Figure 2(b) shows no adjusted C1 contrast, but positive contrasts at steps 2–4 for C2, steps 2–8 for C5, and steps 2–9 for CALL. This graded pattern is not restricted to GPT-4.1-nano.

Under C5, the model generates the response at step 6 and inserts that response into the history used from step 7 onward. Nevertheless, a positive C5 contrast remains detectable at step 7 for all five

models in both ESConv and PersuasionForGood. It remains detectable at step 8 for all five ESConv models and four of five PersuasionForGood models (Table 10). The finite prefix therefore does not lose its association with reference fidelity immediately after model-generated responses begin to accumulate, even though its advantage is not guaranteed to persist throughout the window.

To distinguish the carryover of a finite prefix from the benefit of continued human responses, we directly compare C5 with CALL after step 6. At step 7, C5 history includes its model-generated step-6 response, whereas CALL history includes the recorded human response. A positive, Holm-adjusted C5–CALL contrast is detectable at one or more post-prefix steps for all five PersuasionForGood models and four of five ESConv models, but for none of the MaiChat models (Table 11).

These comparisons answer two distinct persistence questions. C5 versus C0 shows that an early human prefix can remain beneficial after generated responses enter the history; CALL versus C5 shows that continuing to insert recorded human responses can provide a further detectable benefit in PersuasionForGood and ESConv. Thus, carryover from earlier human responses is not equivalent to continued human input. The absence of a detectable direct contrast in MaiChat does not establish equivalence.

### 5.3 Consistency Across Models Depends on Dialogue Setting

The bracketed counts in Table 2 reveal substantial differences in cross-model reproducibility. PersuasionForGood provides the clearest replication of the amount–persistence gradient. Each human-history condition has a detectable benefit at one or more post-step-1 positions for all five models, but shorter prefixes are concentrated earlier in the window. C5 remains detectable through step 7 for all five models and through step 8 for four, while CALL remains detectable through step 8 for all five and through step 9 for three. PersuasionForGood therefore provides the strongest evidence that supplying more human target-speaker history is associated with benefits detectable over more subsequent response positions.

ESConv shows a clear separation between shorter and longer human-history conditions. C1 has no adjusted support for any model, and C2 has support for three of five. In contrast, C5 and CALL are detectable for all five models. C5 remains detectable at step 8 for all models and at step 9 for three; CALL remains detectable through step 9 for all models and through step 10 for four. In this setting, one or two human responses are less reproducible across models than a five-response prefix or continued human history. ESConv therefore most clearly distinguishes brief human exposure from longer or continuous human anchoring.

MaiChat is the least stable setting. In the primary step 1–10 window, Gemini 3.5 Flash-Lite exhibits detectable C2, C5, and CALL contrasts, while Claude Haiku 4.5 exhibits detectable C5 and CALL contrasts; the other three models have no adjusted post-step-1 contrasts. This lower detectability should not be conflated with uniformly smaller fitted benefits. MaiChat has the largest dataset-level descriptive means in Table 2, but also the smallest matched cohort and the widest uncertainty. Its defining pattern is therefore cross-model variability and imprecision rather than a uniformly absent human-history benefit.

In a separate MaiChat analysis over steps 1–20, the window-averaged CALL benefit remains positive for all five models, whereas finite-prefix results vary across models (Appendix B.6, Table 12). This longer window retains fewer conversations than the primary steps 1–10 analysis (Table 7), so we do not interpret it as a continuation of the same matched-cohort trajectory.

The same history condition can yield reproducible detectable benefits in one dataset and uncertain evidence in another; prefix length alone does not determine how consistently a benefit can be detected. MaiChat's less precise results do not establish that human history is ineffective there, and cross-dataset differences cannot be attributed to dialogue purpose alone. Model-specific figures and numerical summaries of fitted trends, window-averaged benefits, and Holm-detectable response steps are reported in Appendix B.

## 6 Discussion

**How far does human history carry?** A five-response human prefix can remain beneficial after a model-generated response enters the history. Yet carryover from that prefix is not equivalent to

continued human input: in PersuasionForGood and ESConv, CALL shows an additional detectable benefit over C5 at one or more post-prefix positions for most models. The earlier human responses remain in the prompt, so a weakening advantage cannot be explained by their explicit removal. Our protocol does not determine whether attenuation reflects recency, changes in the generated dialogue, or reduced use of earlier context. Nor does the last detectable step mark an exact expiration point. Full human history more consistently lowers reference distance, but does not generally prevent that distance from rising at later positions.

**Variation across settings.** The amount–persistence pattern is most reproducible across models in PersuasionForGood. ESConv distinguishes short prefixes from C5 and CALL, whereas MaiChat yields less precise, more model-dependent estimates. These differences cannot be attributed to dialogue purpose alone: dataset size, participants, collection procedures, and task context also differ. They do not establish a general ranking of domains or models.

**Implications for dialogue evaluation.** Evaluation under recorded human history and evaluation after model-generated responses accumulate answer different questions. Multi-turn dialogue studies should specify the source of historical responses and, when self-generated context matters, compare the resulting trajectories. A window-average score alone can obscure both carryover beyond a finite prefix and later attenuation. Matched dialogue windows help ensure that apparent temporal changes do not simply reflect changing conversation cohorts.

## 7 Limitations and Conclusion

**Limitations.** Fixing the recorded interlocutor's utterances may yield contexts less compatible with model-generated replies than with the original dialogue; replay fidelity therefore need not predict performance in adaptive conversations. Semantic distance to a single reference does not establish response quality, empathy, or task success. Dataset differences and model-specific matched cohorts limit attribution of cross-setting and cross-model differences. The C5–CALL contrast compares complete history-construction policies, not separate carryover and continued-input mechanisms. Finally, the last Holm-detectable step is not an effect's endpoint: within each matched window the cohort is fixed, and later non-detection may reflect a smaller contrast or greater uncertainty. Longer matched windows retain fewer dialogues (Table 7); establishing that a benefit has ceased would require more long dialogues and analyses that distinguish a negligible effect from low power.

**Conclusion.** Controlled dialogue replay asks not only whether human response history helps, but how far its benefit carries after new human responses stop being supplied. In the primary windows, C5 remains detectably beneficial relative to C0 after one model-generated response has entered the history for all five models in both ESConv and PersuasionForGood, and after two such responses for nine of these ten model–dataset analyses. This demonstrates post-prefix carryover. Continued human history shows an additional detectable advantage over C5 at one or more post-prefix positions for all five PersuasionForGood models and four ESConv models, while MaiChat provides less conclusive evidence. Reference distance also rises with response position under full human history, so this increase is not unique to self-generated context. Multi-turn evaluations should therefore report who produced historical responses and compare matched trajectories: fidelity under recorded history cannot be assumed to describe fidelity once a model's own responses enter that history.

## AI USE STATEMENT

We used a generative AI assistant (OpenAI Codex) to help refine the research question, discuss analysis choices, interpret statistical outputs, and draft, and edit manuscript text. It also assisted with analysis and figure-generation code, and checking proposed references. The synthetic data generated for this study consist of the saved dialogue continuations produced by the five evaluated models under the experimental conditions. The fixed interlocutor utterances and human reference responses came from the cited human–human datasets, not from generative AI.

Reported quantitative results are based on saved experimental outputs and the specified analysis pipeline, not on estimates supplied by the assistant. We reviewed AI-assisted material, made the final decisions about methods and interpretation, and take responsibility for the paper's code, citations, figures, results, and claims.

## ETHICS STATEMENT

This study reuses existing human–human dialogue datasets and does not recruit or interact with new participants. The ESConv dialogues may contain sensitive emotional-support content, and PersuasionForGood contains incentivized persuasion. Recorded dialogue text was included in prompts sent to external model APIs; structured participant identifiers, profiles, survey fields, and turn-level strategy annotations were not supplied. Free-text dialogue may nevertheless contain sensitive information, so excluding those fields does not eliminate privacy risk. We report aggregate results rather than participant-level assessments. Our semantic-distance measure evaluates similarity to a recorded reply; it does not assess whether a generated response is safe, clinically appropriate, or ethically persuasive. The models were evaluated through replay, not deployed in interactions with participants.

## REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe the dialogue-replay protocol, history conditions, generation setup, semantic-distance measure, matched-window rules, and statistical analysis. Appendix A documents dataset preprocessing, prompts, model identifiers, exclusions, matched cohorts, and model-fitting diagnostics; Appendix B provides model-specific figures and numerical results. The anonymous supplementary materials provide dataset-construction code and acquisition instructions (`canonical/`), as well as replay, metric, and analysis code, prompt assets, and the primary-window figure recipe (`code/`). Raw dialogues, saved predictions and metrics, and API credentials are not redistributed. Regeneration therefore requires obtaining the source datasets and accessing the hosted models. Exact generated responses may vary with provider-side changes; the reported model identifiers and filtering events document, but cannot eliminate, this source of variation.

## REFERENCES


Alibaba Cloud. Qwen3.8-Flash. Alibaba Cloud Model Studio documentation, 2026. URL https://docs.modelstudio.console.alibabacloud.com/en/model-studio/qwen3-8-flash. Accessed September 22, 2026.

Anthropic. Claude Haiku 4.5 System Card. Technical report, Anthropic, October 2025. URL https://www-cdn.anthropic.com/7aad69bf12627d42234e01ee7c36305dc2f6a970.pdf.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.401. URL https://aclanthology.org/2024.acl-long.401/.

Douglas Bates, Martin Mächler, Benjamin M. Bolker, and Steven C. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/e995f98d56967d946471af29d7bf99f1-Paper.pdf.

Mai Hoang Dao, Catherine Lai, and Peter Bell. MaiChat: A text-based dialogue corpus rich in conversational features. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 1585–1594, Palma de Mallorca, Spain, May 2026. ELRA Language Resource Association. doi: 10.63317/3kpp3zj47d6d. URL https://aclanthology.org/2026.lrec-1.123/.

DeepSeek-AI, Anyi Xu, et al. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. BotChat: Evaluating LLMs' capabilities of having multi-turn dialogues. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3184–3200. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.201. URL https://aclanthology.org/2024.findings-naacl.201/.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 82–89. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-2310. URL https://aclanthology.org/W19-2310/.

Google. Gemini 3.5 Flash-Lite. Google AI for Developers, July 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash-lite. Accessed September 22, 2026.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 379–391. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-5944. URL https://aclanthology.org/W19-5944/.

Tianxing He, Jingzhao Zhang, Zhiming Zhou, and James Glass. Exposure bias versus self-recovery: Are distortions really incremental for autoregressive text generation? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5087–5102, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.415. URL https://aclanthology.org/2021.emnlp-main.415/.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. doi: 10.2307/4615733.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune Haubo Bojesen Christensen. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26, 2017. doi: 10.18637/jss.v082.i13.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20153–20177, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1124. URL https://aclanthology.org/2024.emnlp-main.1124/.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 54738–54778, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/59f6421e64707225fdf5b28840679a07-Paper-Conference.pdf.

Alex M. Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. Professor Forcing: A new algorithm for training recurrent networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/16026d60ff9b54410b3435b403afd226-Abstract.html.

Russell V. Lenth and Julia Piaskowski. *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2026. URL https://CRAN.R-project.org/package=emmeans. R package version 2.0.4.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2122–2132, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1230. URL https://aclanthology.org/D16-1230/.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. 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. 3469–3483, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.269. URL https://aclanthology.org/2021.acl-long.269/.

Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 681–707. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.64. URL https://aclanthology.org/2020.acl-main.64/.

OpenAI. Introducing GPT-4.1 in the API. OpenAI, April 2025. URL https://openai.com/index/gpt-4-1/. Published April 14, 2025; accessed September 22, 2026.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Florian Schmidt. Generalization in generation: A closer look at exposure bias. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pp. 157–167. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-5616. URL https://aclanthology.org/D19-5616/.

Sentence Transformers. paraphrase-multilingual-mpnet-base-v2. Hugging Face Model Hub, 2021. URL https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2. Model card; revision 4328cf26390c98c5e3c738b4460a05b95f4911f5; accessed September 22, 2026.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5635–5649, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1566. URL https://aclanthology.org/P19-1566/.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pp. 15–33, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eancs-1.3. URL https://aclanthology.org/2021.eancs-1.3/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

# A Experimental Details

## A.1 Dataset Releases, Preprocessing, and Participants

Table 3 distinguishes the published source statistics from the canonical data used in our experiments. The full MaiChat release contains 50 conversations between 100 participants and 4,975 messages; the version available to our pipeline contains 42 conversations, 84 conversation-local participant identifiers, and 4,177 messages. The original ESConv paper reports 1,053 qualified conversations and 31,410 utterances; we use a subsequently expanded release containing 1,300 conversations and 38,365 messages. PersuasionForGood contains 1,017 conversations, 20,932 utterances, and 1,285 participants.

Canonicalization preserves the original message text and groups consecutive messages from the same speaker into one speaker run. MaiChat keystroke logs are stored separately and are not used in model prompts. In ESConv, the help seeker is mapped to Speaker A and the supporter to Speaker B; strategy, feedback, and survey annotations are retained for provenance but excluded from generation inputs. Stable participant identifiers are unavailable for ESConv.

In PersuasionForGood, the persuadee is mapped to Speaker A and the persuader to Speaker B. Source worker identifiers are transformed into stable anonymized identifiers using salted hashing, allowing repeated participants across conversations to be represented without retaining the original identifiers. Donation outcomes and dialogue-act annotations are retained for auditing but are not supplied to the model. Conversations whose identifiers begin with `BAD_` are generated but excluded from the primary statistical analysis.

When a recorded dialogue begins with Speaker B, this opening run is included as shared context in all five conditions. It is not counted as an eligible response position or toward the C1, C2, or C5 human prefix.

Table 3: Dataset releases and canonical data used in the experiments. Speaker runs group consecutive messages from the same speaker. Eligible positions are target-speaker runs with a preceding interlocutor run and a recorded reference response.

| Dataset | Release used | Conversations | Messages | Speaker runs | Eligible positions | Participant identifiers |
|---|---|---|---|---|---|---|
| MaiChat | Local 42-dialogue release | 42 | 4,177 | 2,870 | 1,425 | 84, conversation-local |
| ESConv | Expanded release | 1,300 | 38,365 | 30,392 | 14,642 | Unavailable |
| PersuasionForGood | Full corpus | 1,017 | 20,932 | 20,932 | 9,583 | 1,285, stable and anonymized |

## A.2 Prompt Templates and Role Blocks

All datasets use the following base instruction:

```
You are a model that reproduces the response behavior
of the real interlocutor B. Read the given conversation
history and generate exactly one natural English utterance
that B is likely to say next.  Do not add explanations,
annotations, speaker labels, or quotation marks; return only
B's utterance text.  Do not invent personal information that
is not present in the conversation.
```

A dataset-specific role block is appended to this instruction. Table 4 lists the exact role block used for each dataset.

Table 4: Dataset-specific target-speaker role blocks.

| Dataset | Role block |
|---|---|
| MaiChat | You are the second speaker in an informal two-person conversation. |
| ESConv | You are the supporter. You respond with empathy and attempt to help the seeker with their problem. |
| PersuasionForGood | You are the persuader. Your goal is to encourage the other person to donate to the charity. |

PersuasionForGood additionally uses the following fixed task context, reconstructed from its published data-collection protocol and shared across all conditions:

```
Task context from the original study: A is the persuadee
and B is the persuader. B attempts to persuade A to donate
part of A's task earnings to Save the Children. B may also
choose to donate.
```

The user message serializes the condition-specific history using the following structure:

```
A: {recorded interlocutor run 1}
B: {human or model response selected by condition}
...
A: {current recorded interlocutor run}
```

If Speaker B opens the recorded dialogue, the shared opening is placed before the first Speaker-A run. The current human Speaker-B response, future turns, persona and profile fields, external knowledge, dialogue goals, and ESConv or PersuasionForGood strategy annotations are not included in the prompt.

### A.3 Model Access and Generation Provenance

Table 5 reports the access route and the requested and returned identifiers observed across the three dataset runs. Each model resolved to the same returned identifier across all saved predictions.

Table 5: Model access routes and observed identifiers.

| Model | Route and protocol | Requested identifier | Returned identifier |
|---|---|---|---|
| GPT-4.1-nano | Azure; OpenAI-compatible chat | `gpt-4.1-nano` | `gpt-4.1-nano-2025-04-14` |
| DeepSeek-V4-Flash | OpenRouter; OpenAI-compatible chat | `deepseek/deepseek-v4-flash` | `deepseek/deepseek-v4-flash` |
| Qwen3.8-Flash | OpenRouter; OpenAI-compatible chat | `qwen/qwen3.8-flash` | `qwen/qwen3.8-flash` |
| Gemini 3.5 Flash-Lite | GPTSAPI; Gemini-compatible GenerateContent API | `gemini-3.5-flash-lite` | `gemini-3.5-flash-lite` |
| Claude Haiku 4.5 | GPTSAPI; Anthropic-compatible Messages API | `claude-haiku-4-5-cc` | `claude-haiku-4-5-20251001` |

All runs use temperature zero. Optional reasoning was disabled for DeepSeek-V4-Flash, Qwen3.8-Flash, and Claude Haiku 4.5. Gemini 3.5 Flash-Lite used its minimal-thinking setting, while GPT-4.1-nano was run without an optional-reasoning configuration. Within each model, the generation settings were held fixed across datasets and history conditions.

### A.4 Matched Cohorts and Runtime Exclusions

A provider-side filtering event at any condition or response position causes the entire dialogue to be excluded for that model, preserving condition pairing. The primary analysis then retains only dialogues with all five conditions at every position in the specified window. Table 6 reports each resulting model-specific cohort.

The primary windows are steps 1–10 for MaiChat and ESConv and steps 1–9 for PersuasionForGood. The additional DeepSeek/ESConv exclusion is conversation `esconv:00767`. Four of its five independently generated step-1 responses included a `B:` speaker label despite the instruction to return only the response text. We conservatively excluded the conversation from the matched analysis as a response-format violation; this was not a provider content-filter event.

The cohort is fixed across response positions within each model-specific analysis. Cohorts are not forced to be identical across models because provider-side filtering differs by access route. Consequently, cross-model comparisons are interpreted descriptively rather than as estimates from a single pooled common cohort.

#### A.4.1 Response-Step Window Coverage

Table 7 reports conversation counts at selected response-step cutoffs. "Canonical" counts preprocessed conversations with at least $N$ eligible target-speaker positions, irrespective of whether model

Table 6: Primary matched cohorts. Runtime exclusions are whole-dialogue exclusions caused by provider-side filtering. "Other" comprises the 27 `BAD_` PersuasionForGood dialogues or, for DeepSeek/ESConv, one dialogue with output-format violations. Analysis rows equal retained dialogues × response positions × five conditions.

| Dataset | Model | Runtime excluded | Other excluded | Incomplete for window | Retained dialogues | Analysis rows |
|---|---|---|---|---|---|---|
| MaiChat | GPT-4.1-nano | 0 | 0 | 0 | 42 | 2,100 |
| MaiChat | DeepSeek-V4-Flash | 0 | 0 | 0 | 42 | 2,100 |
| MaiChat | Qwen3.8-Flash | 0 | 0 | 0 | 42 | 2,100 |
| MaiChat | Gemini 3.5 Flash-Lite | 1 | 0 | 0 | 41 | 2,050 |
| MaiChat | Claude Haiku 4.5 | 0 | 0 | 0 | 42 | 2,100 |
| ESConv | GPT-4.1-nano | 30 | 0 | 354 | 916 | 45,800 |
| ESConv | DeepSeek-V4-Flash | 2 | 1 | 360 | 937 | 46,850 |
| ESConv | Qwen3.8-Flash | 0 | 0 | 361 | 939 | 46,950 |
| ESConv | Gemini 3.5 Flash-Lite | 3 | 0 | 360 | 937 | 46,850 |
| ESConv | Claude Haiku 4.5 | 0 | 0 | 361 | 939 | 46,950 |
| PersuasionForGood | GPT-4.1-nano | 2 | 27 | 0 | 988 | 44,460 |
| PersuasionForGood | DeepSeek-V4-Flash | 0 | 27 | 0 | 990 | 44,550 |
| PersuasionForGood | Qwen3.8-Flash | 0 | 27 | 0 | 990 | 44,550 |
| PersuasionForGood | Gemini 3.5 Flash-Lite | 6 | 27 | 0 | 984 | 44,280 |
| PersuasionForGood | Claude Haiku 4.5 | 0 | 27 | 0 | 990 | 44,550 |

predictions are complete. Each model row additionally requires valid predictions in all five conditions at every step from 1 through $N$, after runtime and format exclusions. In the PersuasionForGood panel, entries give counts before/after excluding `BAD_` conversations; the latter counts are used in the primary analysis.

Table 7: Conversations with complete five-condition trajectories through selected response-step cutoffs.

| **MaiChat** Model | ≥10 | ≥15 | ≥20 | ≥30 | ≥40 |
|---|---|---|---|---|---|
| Canonical | 42 | 41 | 39 | 22 | 12 |
| GPT-4.1-nano | 42 | 41 | 39 | 22 | 12 |
| DeepSeek-V4-Flash | 42 | 41 | 39 | 22 | 12 |
| Qwen3.8-Flash | 42 | 41 | 39 | 22 | 12 |
| Gemini 3.5 Flash-Lite | 41 | 40 | 38 | 21 | 11 |
| Claude Haiku 4.5 | 42 | 41 | 39 | 22 | 12 |

| **ESConv** Model | ≥10 | ≥12 | ≥15 | ≥20 |
|---|---|---|---|---|
| Canonical | 939 | 445 | 163 | 36 |
| GPT-4.1-nano | 916 | 432 | 158 | 35 |
| DeepSeek-V4-Flash | 937 | 444 | 163 | 36 |
| Qwen3.8-Flash | 939 | 445 | 163 | 36 |
| Gemini 3.5 Flash-Lite | 937 | 445 | 163 | 36 |
| Claude Haiku 4.5 | 939 | 445 | 163 | 36 |

| **PersuasionForGood: including/excluding `BAD_` conversations** Model | ≥9 | ≥10 | ≥11 | ≥14 |
|---|---|---|---|---|
| Canonical | 1016/990 | 363/361 | 54/54 | 1/1 |
| GPT-4.1-nano | 1014/988 | 362/360 | 54/54 | 1/1 |
| DeepSeek-V4-Flash | 1016/990 | 363/361 | 54/54 | 1/1 |
| Qwen3.8-Flash | 1016/990 | 363/361 | 54/54 | 1/1 |
| Gemini 3.5 Flash-Lite | 1010/984 | 359/357 | 53/53 | 1/1 |
| Claude Haiku 4.5 | 1016/990 | 363/361 | 54/54 | 1/1 |

The longest complete trajectories contain 71 response steps in MaiChat, 43 in ESConv, and 14 in PersuasionForGood. Among the 27 `BAD_` conversations in PersuasionForGood, 26 reach step 9, two reach step 10, and none reaches step 11. The ESConv DeepSeek counts include the previously specified exclusion of `esconv:00767`.

### A.5 Random-Effects Specifications and Diagnostics

For each model, dataset, and window, we first fit the richest supported random-effects structure. Candidate structures are considered in the following fixed order:

1. conversation-level random slope and intercept, target-speaker intercept when stable cross-conversation identifiers are available, and matched-position intercept;
2. conversation-level random slope and intercept plus matched-position intercept;
3. conversation and matched-position random intercepts.

Here, a matched position identifies the five condition-specific predictions for the same conversation and response step. A fitted structure is retained only when the optimizer converges, the relative gradient is below $0.001$ when available, and the fit is not singular. Table 8 reports the final structures for the primary analyses.

Table 8: Final random-effects structures in the primary analyses. All listed fits converged and were non-singular.

| Dataset | Model(s) | Final random-effects structure | Converged | Singular |
|---|---|---|---|---|
| MaiChat | GPT-4.1-nano | $(1 + \text{step}_c \mid \text{conversation}) + (1 \mid \text{matched position})$ | Yes | No |
| MaiChat | DeepSeek, Qwen, Gemini, Claude | $(1 \mid \text{conversation}) + (1 \mid \text{matched position})$ | Yes | No |
| ESConv | All five models | $(1 + \text{step}_c \mid \text{conversation}) + (1 \mid \text{matched position})$ | Yes | No |
| PersuasionForGood | GPT-4.1-nano, DeepSeek, Qwen, Gemini | $(1 + \text{step}_c \mid \text{conversation}) + (1 \mid \text{target speaker}) + (1 \mid \text{matched position})$ | Yes | No |
| PersuasionForGood | Claude Haiku 4.5 | $(1 + \text{step}_c \mid \text{conversation}) + (1 \mid \text{matched position})$ | Yes | No |

MaiChat participant identifiers are conversation-local and therefore do not support a separate cross-conversation speaker effect. ESConv does not provide stable participant identifiers. PersuasionForGood provides stable anonymized identifiers, allowing a target-speaker intercept to be considered. For Claude Haiku 4.5 on PersuasionForGood, the richer speaker-intercept specification did not pass the convergence and singularity criteria, so the predefined fallback without that intercept was used.

# B Detailed Model-Specific Results

This section provides the model-specific evidence underlying the main-text findings. Primary matched windows cover response steps 1–10 for MaiChat and ESConv and steps 1–9 for PersuasionForGood; each analysis uses a model-specific matched conversation cohort. The figures pair raw means and mixed-model estimated trajectories (left) with estimated benefits relative to C0 (right). Their ribbons show pointwise 95% confidence intervals.

The accompanying tables report primary-window response-step trends, window-averaged benefits and Holm-detectable steps, and direct post-prefix C5–CALL contrasts. A separate MaiChat analysis over steps 1–20 appears at the end of this section. Contrast directions and Holm-adjustment families are specified in the respective table captions.

## B.1 MaiChat

Figures 3–7 show the five model-specific analyses over response steps 1–10.

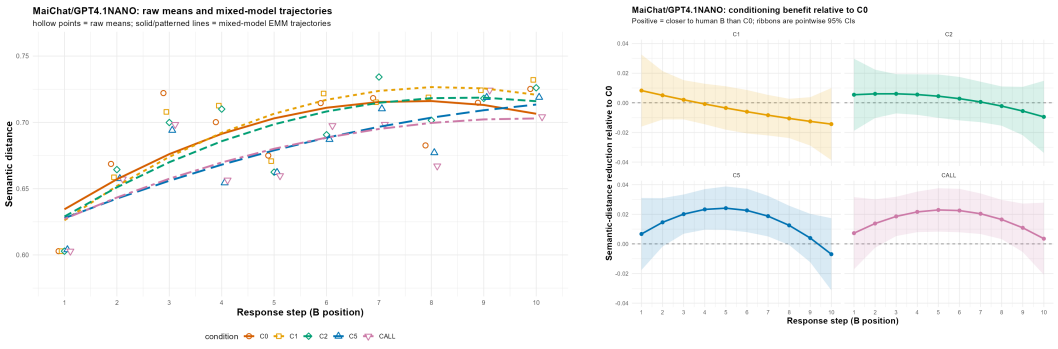



Figure 3: **MaiChat: GPT-4.1-nano.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

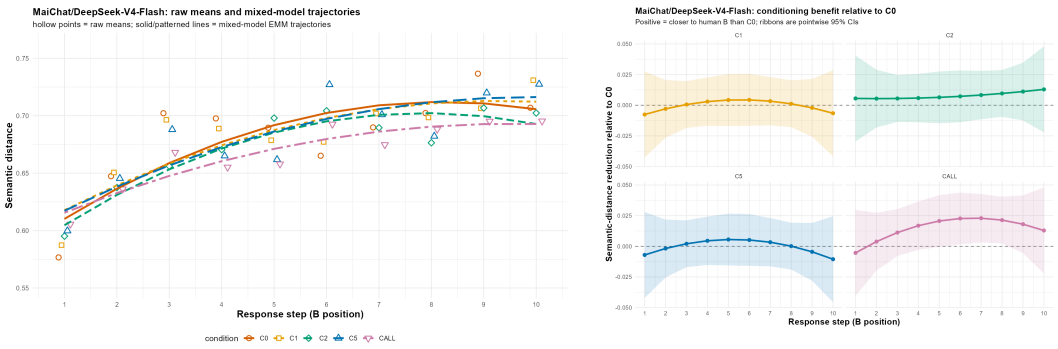



Figure 4: **MaiChat: DeepSeek-V4-Flash.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

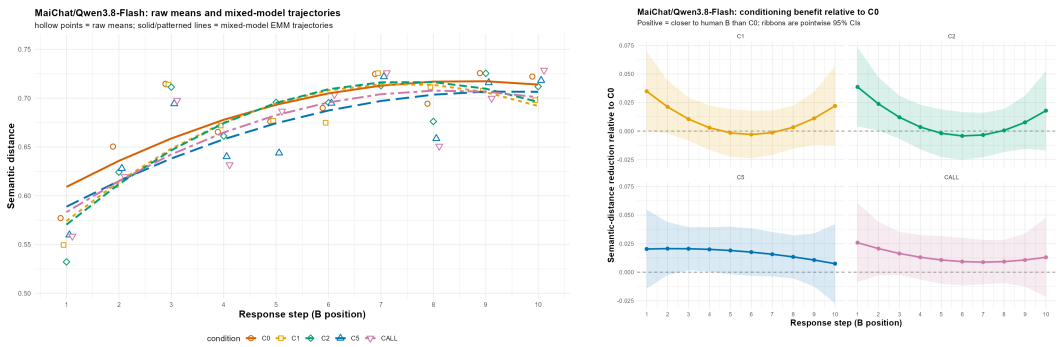



Figure 5: **MaiChat: Qwen3.8-Flash.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

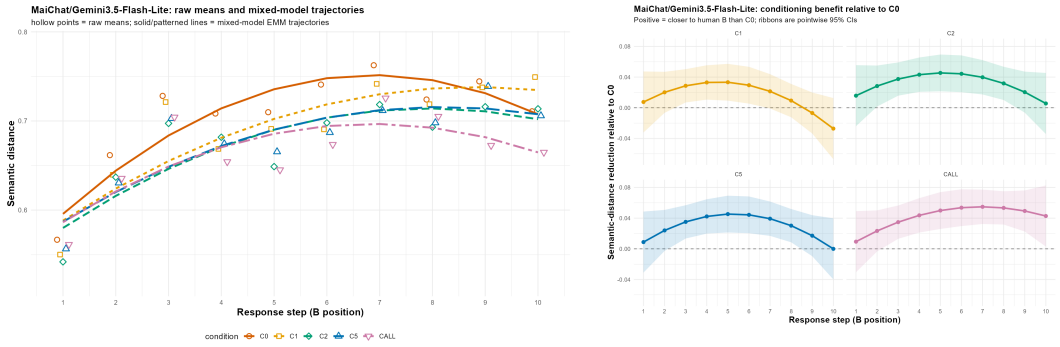



Figure 6: **MaiChat: Gemini 3.5 Flash-Lite.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

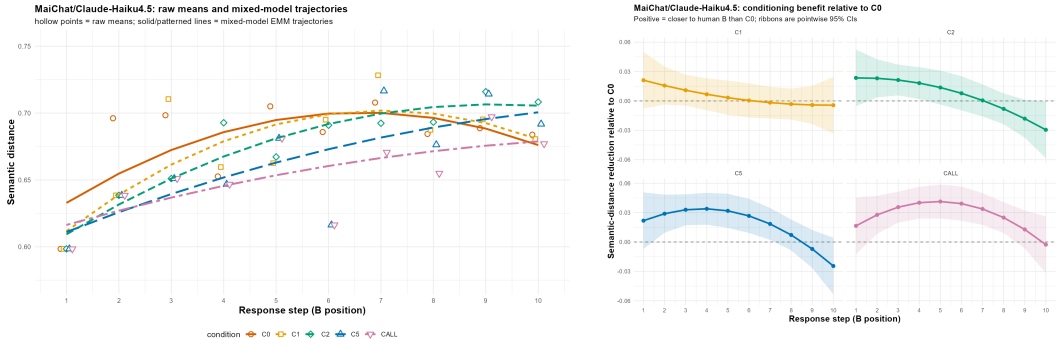



Figure 7: **MaiChat: Claude Haiku 4.5.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

### B.2 ESConv

Figure 2 presents GPT-4.1-nano in the main text. Figures 8–11 show the other four models over response steps 1–10.

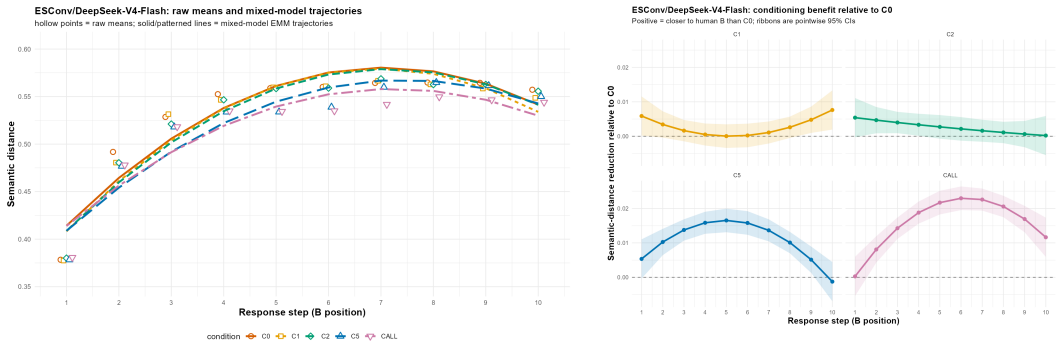



Figure 8: **ESConv: DeepSeek-V4-Flash.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

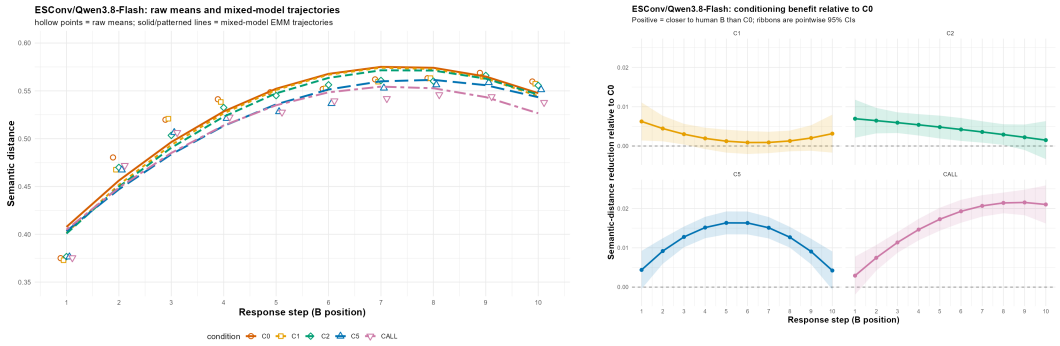



Figure 9: **ESConv: Qwen3.8-Flash.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

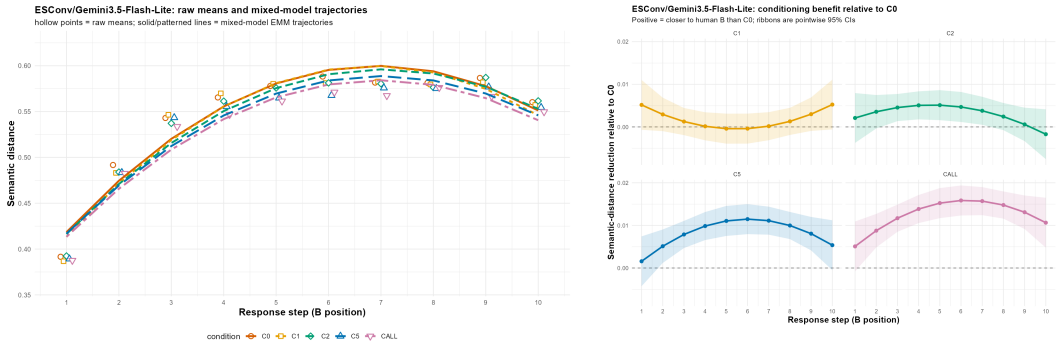



Figure 10: **ESConv: Gemini 3.5 Flash-Lite.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

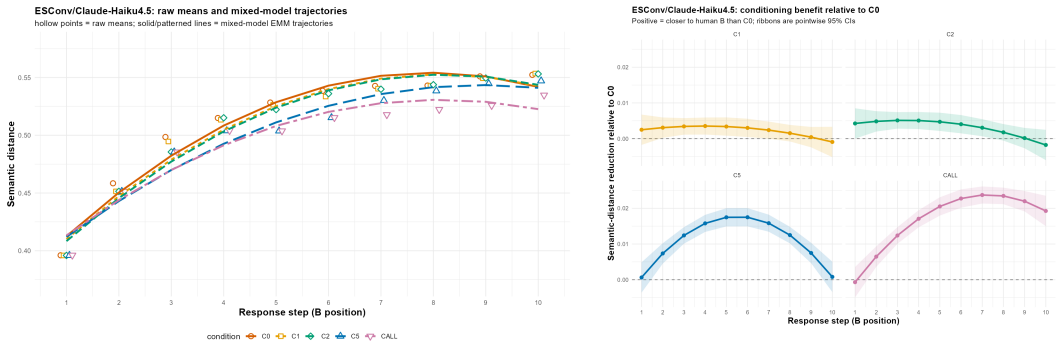



Figure 11: **ESConv: Claude Haiku 4.5.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

### B.3 PERSUASIONFORGOOD

Figures 12–16 show the five model-specific analyses over response steps 1–9. Conversations marked `BAD_` are excluded from these analyses.

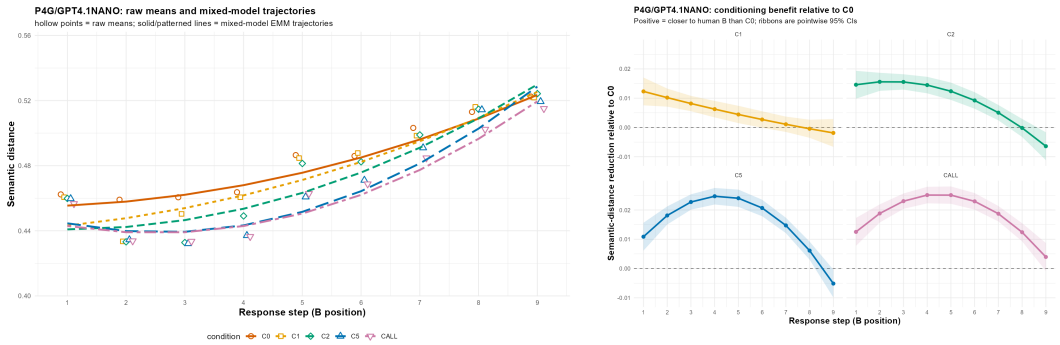



Figure 12: **PersuasionForGood: GPT-4.1-nano.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

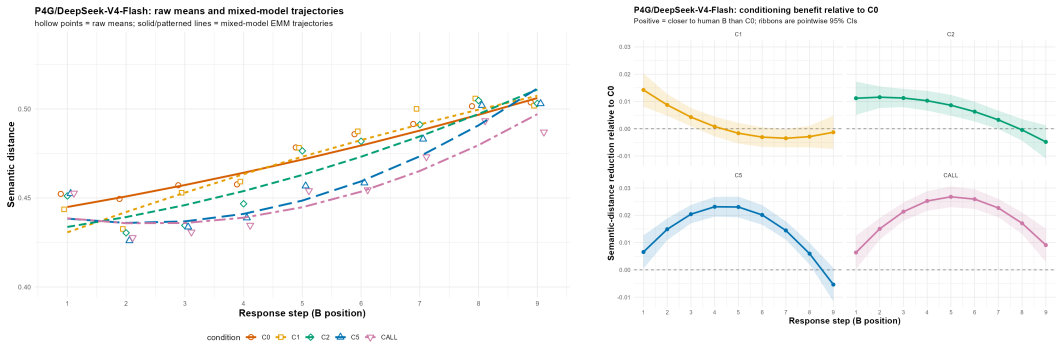



Figure 13: **PersuasionForGood: DeepSeek-V4-Flash.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

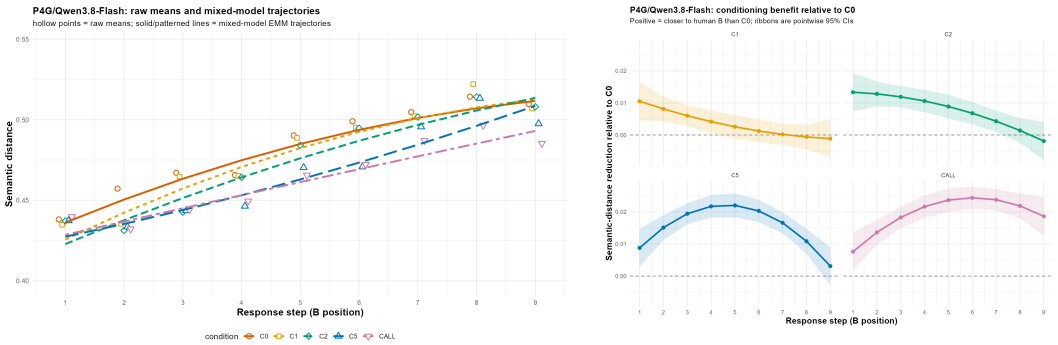



Figure 14: **PersuasionForGood: Qwen3.8-Flash.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

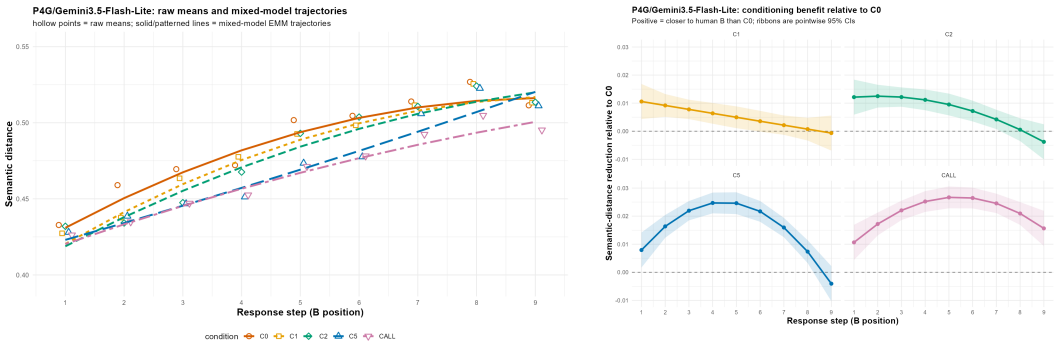



Figure 15: **PersuasionForGood: Gemini 3.5 Flash-Lite.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

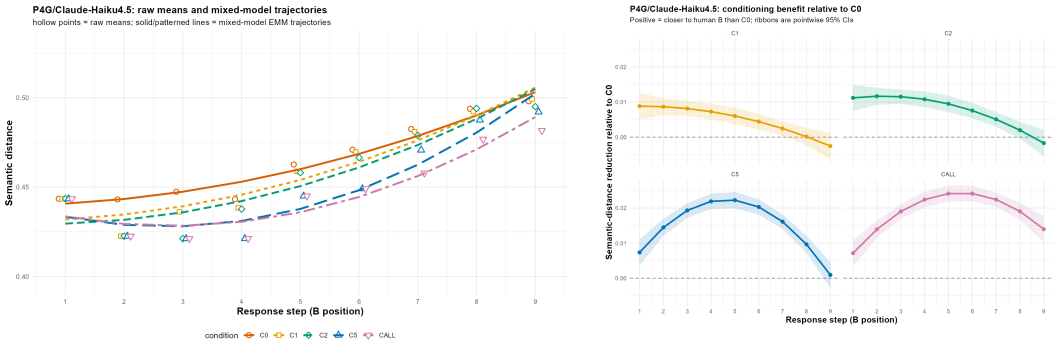



Figure 16: **PersuasionForGood: Claude Haiku 4.5.** Left: raw means and mixed-model trajectories. Right: estimated benefit relative to C0.

### B.4 Primary-Window Trend Estimates

Table 9 reports the fitted trend at the center of each primary matched window. It also reports the fitted change in C0 distance between the first and final response steps.

Table 9: **Primary-window response-step trends.** $\Delta d_{\mathrm{C0}}$ is the fitted C0 distance at the final step minus that at step 1. Each $\beta_c$ is the fitted window-center response-step trend for condition $c$; an asterisk denotes a positive trend with Holm-adjusted $p < .05$ among the five condition trends. $\Delta\beta = \beta_{\mathrm{C0}} - \beta_{\mathrm{CALL}}$; $p^H_{\Delta\beta}$ is Holm-adjusted across all pairwise condition trend contrasts within that model–dataset analysis. GPT, DeepSeek, Qwen, Gemini, and Claude abbreviate the five models named in Section 4.3.

| Dataset | Model | $\Delta d_{\mathrm{C0}}$ | $\beta_{\mathrm{C0}}$ | $\beta_{\mathrm{C1}}$ | $\beta_{\mathrm{C2}}$ | $\beta_{\mathrm{C5}}$ | $\beta_{\mathrm{CALL}}$ | $\Delta\beta$ | $p^H_{\Delta\beta}$ |
|---|---|---|---|---|---|---|---|---|---|
| MaiChat | GPT | 0.0720 | 0.0080* | 0.0105* | 0.0096* | 0.0095* | 0.0084* | -0.00042 | 1.000 |
| | DeepSeek | 0.0955 | 0.0106* | 0.0105* | 0.0098* | 0.0110* | 0.0086* | 0.00202 | 1.000 |
| | Qwen | 0.1048 | 0.0116* | 0.0131* | 0.0139* | 0.0131* | 0.0131* | -0.00142 | 1.000 |
| | Gemini | 0.1118 | 0.0124* | 0.0163* | 0.0135* | 0.0134* | 0.0087* | 0.00370 | 1.000 |
| | Claude | 0.0433 | 0.0048 | 0.0076* | 0.0107* | 0.0099* | 0.0069* | -0.00215 | 1.000 |
| ESConv | GPT | 0.0988 | 0.0110* | 0.0113* | 0.0113* | 0.0119* | 0.0106* | 0.00036 | 1.000 |
| | DeepSeek | 0.1274 | 0.0141* | 0.0139* | 0.0147* | 0.0149* | 0.0129* | 0.00125 | 0.016 |
| | Qwen | 0.1399 | 0.0155* | 0.0158* | 0.0161* | 0.0155* | 0.0135* | 0.00201 | $< .001$ |
| | Gemini | 0.1327 | 0.0147* | 0.0147* | 0.0151* | 0.0143* | 0.0141* | 0.00061 | 1.000 |
| | Claude | 0.1293 | 0.0143* | 0.0147* | 0.0150* | 0.0143* | 0.0121* | 0.00221 | $< .001$ |
| P4G | GPT | 0.0680 | 0.0085* | 0.0103* | 0.0111* | 0.0105* | 0.0096* | -0.00108 | 0.030 |
| | DeepSeek | 0.0612 | 0.0077* | 0.0096* | 0.0097* | 0.0092* | 0.0073* | 0.00033 | 1.000 |
| | Qwen | 0.0755 | 0.0094* | 0.0109* | 0.0113* | 0.0101* | 0.0081* | 0.00137 | 0.022 |
| | Gemini | 0.0853 | 0.0106* | 0.0121* | 0.0126* | 0.0122* | 0.0100* | 0.00062 | 0.889 |
| | Claude | 0.0624 | 0.0078* | 0.0092* | 0.0094* | 0.0086* | 0.0070* | 0.00085 | 0.025 |

### B.5 Model-Specific Benefits and Detectable Steps

Table 10 gives the model-specific results underlying the cross-model summaries in Table 2.

Table 10: **Primary-window fitted benefits and Holm-detectable response steps.** Each condition cell gives the window-average fitted benefit $d_{\mathrm{C0}} - d_c$, followed in parentheses by post-step-1 positions where that contrast is positive with Holm-adjusted $p < .05$. A dash means no such position; it does not establish equivalence to C0. Adjustment covers all pairwise condition contrasts across the response steps of the corresponding model–dataset analysis. $n$ is the number of conversations with complete five-condition trajectories in that window. P4G excludes dialogues marked `BAD_`.

| Dataset | Model | $n$ | C1 | C2 | C5 | CALL |
|---|---|---|---|---|---|---|
| MaiChat | GPT | 42 | -0.0041 (–) | 0.0014 (–) | 0.0140 (–) | 0.0158 (–) |
| | DeepSeek | 42 | -0.0003 (–) | 0.0078 (–) | -0.0003 (–) | 0.0145 (–) |
| | Qwen | 42 | 0.0100 (–) | 0.0095 (–) | 0.0165 (–) | 0.0137 (–) |
| | Gemini | 41 | 0.0149 (–) | 0.0312 (4–7) | 0.0286 (4–7) | 0.0415 (4–9) |
| | Claude | 42 | 0.0045 (–) | 0.0053 (–) | 0.0170 (3–5) | 0.0269 (3–7) |
| ESConv | GPT | 916 | 0.0026 (–) | 0.0042 (2–4) | 0.0081 (2–8) | 0.0113 (2–9) |
| | DeepSeek | 937 | 0.0028 (–) | 0.0026 (–) | 0.0105 (2–8) | 0.0158 (2–10) |
| | Qwen | 939 | 0.0025 (–) | 0.0044 (2–4) | 0.0115 (2–9) | 0.0158 (2–10) |
| | Gemini | 937 | 0.0018 (–) | 0.0030 (–) | 0.0081 (3–9) | 0.0125 (2–10) |
| | Claude | 939 | 0.0022 (–) | 0.0031 (2–5) | 0.0108 (2–9) | 0.0167 (2–10) |
| P4G | GPT | 988 | 0.0048 (2–4) | 0.0089 (2–7) | 0.0152 (2–8) | 0.0181 (2–8) |
| | DeepSeek | 990 | 0.0018 (2) | 0.0064 (2–6) | 0.0137 (2–7) | 0.0188 (2–8) |
| | Qwen | 990 | 0.0035 (2–3) | 0.0076 (2–6) | 0.0153 (2–8) | 0.0193 (2–9) |
| | Gemini | 984 | 0.0050 (2–4) | 0.0073 (2–6) | 0.0152 (2–8) | 0.0210 (2–9) |
| | Claude | 990 | 0.0048 (2–6) | 0.0075 (2–7) | 0.0147 (2–8) | 0.0185 (2–9) |

Table 11: **Continued human history versus a five-response prefix.** The mean is the average fitted contrast $d_{C5} - d_{CALL}$ over response steps 7–10 for MaiChat and ESConv or 7–9 for P4G; positive values favor CALL. Parentheses list post-prefix steps with a positive direct contrast and Holm-adjusted $p < .05$. Holm adjustment uses the full family of pairwise condition-by-step contrasts for each model–dataset analysis. A dash does not establish equivalence.

| Dataset | Model | Mean contrast | Detectable steps |
|---|---|---|---|
| MaiChat | GPT | 0.00575 | – |
| | DeepSeek | 0.02169 | – |
| | Qwen | -0.00136 | – |
| | Gemini | 0.02842 | – |
| | Claude | 0.01866 | – |
| ESConv | GPT | 0.00685 | 7–10 |
| | DeepSeek | 0.01102 | 7–10 |
| | Qwen | 0.01090 | 7–10 |
| | Gemini | 0.00493 | – |
| | Claude | 0.01296 | 7–10 |
| P4G | GPT | 0.00646 | 8–9 |
| | DeepSeek | 0.01128 | 7–9 |
| | Qwen | 0.01126 | 7–9 |
| | Gemini | 0.01396 | 7–9 |
| | Claude | 0.00959 | 7–9 |

Table 11 directly compares CALL with C5 after the five-response prefix ends, complementing the C0-relative comparisons in Table 10.

### B.6 Longer MaiChat Window

Table 12 reports a separate analysis of MaiChat response steps 1–20. Its matched cohorts differ from those of the primary steps 1–10 analysis, so the two windows should not be read as a single continuous fitted trajectory.

Table 12: **MaiChat results over response steps 1–20.** Cells have the same meaning as in Table 10: window-average fitted benefit, followed by positive post-step-1 response positions with Holm-adjusted $p < .05$. A dash indicates no such position. The longer window uses a different matched cohort and a larger step-specific comparison family than the primary window.

| Model | $n$ | C1 | C2 | C5 | CALL |
|---|---|---|---|---|---|
| GPT | 39 | 0.0046 (–) | 0.0098 (–) | 0.0201 (14–19) | 0.0196 (14–16) |
| DeepSeek | 39 | 0.0031 (–) | 0.0042 (–) | -0.0029 (–) | 0.0228 (14–17) |
| Qwen | 39 | 0.0085 (–) | 0.0070 (–) | 0.0087 (–) | 0.0260 (14–18) |
| Gemini | 38 | 0.0104 (–) | 0.0176 (5–6) | 0.0197 (6–8) | 0.0437 (4–17) |
| Claude | 39 | 0.0008 (–) | 0.0033 (–) | 0.0107 (–) | 0.0227 (3–11) |