Describing Is Not Generating: Distributional Faithfulness in LLM-Based Human Group Simulation

Anonymous ACL submission

Abstract

Large language models are increasingly used as proxies for simulating the behavior of human groups. However, because group behavior is reflected not in a single plausible response but in the distribution of response types, individual response plausibility alone cannot guarantee faithful group simulation. We evaluate LLM-based human group simulation as a problem of distributional realization. We compare real human response-type distributions, model-verbalized estimates of those distributions, and distributions from model-generated responses in a shared behavior space. We evaluate this problem in dialogue between children and adults, English learner dialogue, and tutoring dialogue between experts and novices. We further use controlled generation conditions to diagnose whether distributional mismatch arises from missing distributional information, response type selection, or limitations in realizing a selected response type. Our results show that models can often verbalize distributions close to real human behavior, but direct generation often produces response-type distributions that differ from both empirical human behavior and the models' own verbalized distributions. Providing the verbalized distribution in the prompt does not consistently close this gap. Generation-process interventions reduce the gap more often. These findings reveal a recurring **describing is not generating** problem in group simulation, suggesting that LLM proxies for human groups should be evaluated by group-level distributional faithfulness, not only by individual response plausibility. [1]

1 Introduction

Large language models (LLMs) are increasingly used as tools for simulating the behavior of different human groups, such as synthetic participants, simulated users, and simulated patients (Park et al., 2023; Tu et al., 2024; Labrak et al., 2026; Wang et al., 2024a). A common approach is to give the model a specific persona or group condition, and then analyze or use the generated responses as possible responses from that group. This approach can be useful when collecting real human data is difficult or costly. However, a generated response that appears natural or consistent with a persona does not necessarily faithfully reproduce the behavior of the real group (Wang et al., 2025; Cheng et al., 2023a; Seshadri et al., 2026). To test whether generated responses faithfully reproduce real group behavior, we need to compare them with responses from real people under comparable conditions.

One challenge is the lack of suitable evaluation data. Faithfully evaluating LLM-based group simulation requires data that links group attributes or roles to real human responses under comparable conditions, as in survey-based opinion alignment benchmarks (Santurkar et al., 2023; Durmus et al., 2024; Meister et al., 2025). However, comparable open-ended dialogue data is harder to obtain, especially in domains where real interactions are costly, private, or difficult to release (Wang et al., 2024a; Labrak et al., 2026). As a result, LLM-based simulation is often judged by whether a few generated responses look natural or match the persona, rather than by whether the model's responses match real group behavior.

Evaluating group simulation based only on individual generated responses can be risky (Wang et al., 2025; Seshadri et al., 2026; Santurkar et al., 2023). Even if one or two responses appear to imitate the language or behavior of a target group, this does not mean that they faithfully reflect the behavior of the real group. Instead, a model may exaggerate features that are salient or easy to recognize for that group. In this case, the response may seem consistent with the persona, but it may reflect a stereotype or caricature rather than real behavior (Cheng et al., 2023b,a; Liu et al., 2024). The risk of individual-level evaluation comes from

[1] Our code will be released publicly.

the fact that human group behavior is expressed as a distributional pattern (Durmus et al., 2024; Meister et al., 2025; Santurkar et al., 2023), not as isolated individual responses. Even under the same question or dialogue context, real people can respond in different ways, such as giving a short response, a direct answer, an incomplete answer, or shifting to another topic (Chun et al., 2026). These responses form group-specific behavior distributions. In contrast, under the same group condition, an LLM may generate responses that are longer, more grammatically fluent, and more explanatory than those produced by the real group. Such responses may appear natural at the individual level, but the group-level response distribution can still differ from the empirical distribution of the real human group.

One possible explanation is that the model lacks knowledge of the target group's behavior distribution. However, another possibility is that the model can describe how a group is likely to respond by verbalizing the target behavior distribution, but that it may fail to follow that knowledge during generation. We hypothesize that the ability to verbalize a group-specific behavior distribution and the ability to realize that distribution in generated responses may be distinct. Although recent work uses verbalized sampling (Zhang et al., 2025) to mitigate mode collapse and increase output diversity, our use of verbalization is diagnostic rather than generative. We ask whether the distribution a model states for a human group is actually realized in its generated responses, rather than treating diversity itself as the goal.

To analyze this issue, we compare empirical human behavior, model-verbalized behavior distributions, and model-generated behavior distributions under the same group conditions. This comparison allows us to examine whether models can describe real group behavior, whether that knowledge is realized in generation, and where gaps between verbalization and generation emerge. We apply this evaluation to three human group simulation settings: child dialogue using **CHILDES** (MacWhinney, 2000; Chun et al., 2026), English learner dialogue using **ICNALE** (Ishikawa, 2023), and tutoring dialogue between experts and novices using **Bridge** (Wang et al., 2024b). We choose these settings because they exhibit group-level differences in response behavior, making them well suited for evaluating distributional faithfulness. Across these domains, we compare real human response distributions, model-verbalized distributions, and model-generated response distributions using a shared response behavior space with five categories: *non-answer, minimal response, direct answer, elaborated response*, and *topic shift*. This allows us to evaluate not just the surface-level naturalness of individual responses, but the distributions of response types that emerge at the group level.

Our results show a recurring **describing is not generating** problem in LLM-based group simulation. Models were often able to verbalize distributions close to real human behavior, but direct generation frequently produced response-type distributions that differed from both empirical human behavior and the models' own verbalized distributions. This pattern is clearest in CHILDES and ICNALE, where verbalized distributions are often closer to empirical gold than direct generation, while Bridge shows a more mixed but still informative pattern. Providing the verbalized distribution in the prompt does not consistently improve distributional faithfulness, suggesting that the mismatch is not merely a lack of distributional information. Instead, the results point to a problem in the generation stage. Models may have access to group-level distributional knowledge, yet fail to select and maintain the appropriate response types during generation. Generation-process interventions support this interpretation. Candidate-pool generation brings generated behavior closer to the empirical distribution, especially in CHILDES and Bridge, while both candidate-pool generation and externally sampled response-type generation are effective in ICNALE. Taken together, these findings show that faithful LLM-based group simulation requires evaluating whether models realize group-level behavior distributions in generation, rather than only whether they can describe those distributions or produce individually plausible responses.

2 Related Work

LLMs are increasingly used to simulate human respondents, social agents, and role-based dialogue partners. Studies of demographic personas, role-conditioned agents, and simulated dialogue partners show that LLMs can produce plausible individual responses and interactions (Park et al., 2023; Hu and Collier, 2024; Samuel et al., 2025; Tu et al., 2024; Labrak et al., 2026; Wang et al., 2024a). However, a response that appears natural or persona-consistent does not necessarily show

that the model faithfully reproduces the behavior of the real human group. Prior work in NLG evaluation has shown that response-level quality alone can miss distributional failures, motivating metrics that jointly consider quality and diversity or compare model and human text distributions (Pillutla et al., 2021; Hashimoto et al., 2019; Alihosseini et al., 2019; Han et al., 2022; Stasaski and Hearst, 2022). We extend this view to group-conditioned dialogue simulation, where the target is not general text similarity or survey opinion alignment, but the empirical distribution of response behavior types observed from real human groups.

Related work on opinion alignment and dialogue-act analysis also evaluates model behavior at the distribution level. OpinionQA, GlobalOpinionQA, and distributional alignment benchmarks compare LLM responses with demographic or national group distributions over multiple-choice survey answers (Durmus et al., 2024; Meister et al., 2025; Santurkar et al., 2023). Dialogue-act analysis compares LLM-generated and human dialogue in terms of functional response distributions (Maitra et al., 2025). These studies are closely related, but they mainly study opinion distributions over closed-ended survey choices. In contrast, we map open-ended dialogue responses into a behavior label space and evaluate group-conditioned empirical behavior distributions. This lets us separate whether a model can verbalize a human group's behavior distribution from whether it can realize that distribution in dialogue generation.

Our work is also related to probability verbalization and sampling. Prior work has used verbalized probabilities to guide sampling, reduce mode collapse, or increase output diversity, and has also studied whether LLMs can sample from explicitly or implicitly specified distributions (Zhang et al., 2025; Gu et al., 2025; Xiao et al., 2026). These studies motivate the distinction between describing a distribution and realizing it through generated samples. However, they mainly focus on controlled probability distributions, action sequences, or diversity-oriented generation. In contrast, we study human group simulation in open-ended dialogue, where the target distribution is an empirical behavior distribution observed from real human groups. In our work, verbalization is used not to improve diversity, but as a diagnostic probe. We compare a model's stated group-level behavior distribution with the distribution it actually generates. Across direct generation, distribution-conditioned generation, candidate-pool generation, and externally sampled response-type generation, we diagnose whether mismatches arise from missing distributional information, response-type selection, or difficulty realizing specific response types.

3 Methods

Our goal is to test whether LLMs can not only verbalize group-specific behavioral patterns, but also realize those patterns in generation. We compare distributions at the level of corpus-specific target groups. Groups are defined separately for each corpus: child ages (Age 2, 5, and 9) in **CHILDES**, learner proficiency levels ranging from the lowest A2_0 to the highest B2_0 in **ICNALE**, and expert/novice roles in **Bridge**. For each corpus d and target group g, we compute three response-type distributions over the same five-label behavior space. We denote this label space by $\mathcal{Y}$, which contains five labels: *non_answer*, *minimal*, *direct_answer*, *elaborated*, and *topic_shift*. We use these labels to capture broad response patterns that can appear across all three corpora, while still allowing the analysis to compare how short or elaborated responses are. In particular, the distinction among *minimal*, *direct_answer*, and *elaborated* lets us test whether models not only generate relevant responses, but also answer briefly when real human groups tend to answer briefly and give more detailed responses when real human groups tend to do so. Appendix B provides corpus-specific definitions and explains the motivation for using this shared label set. The **empirical (gold) distribution** $P_G^d(y \mid g)$ is the observed distribution of behavior labels in real human responses from that corpus and group. The **verbalized distribution** $P_V^{m,d}(y \mid g)$ is obtained by asking model m to assign probabilities to the five response types for each example, and then averaging these probability vectors within the same corpus and group. The **generated distribution** $P_C^{m,d}(y \mid g)$ is the observed distribution of behavior labels in responses generated by model m under condition C, again aggregated within the same corpus and group. We represent each example as

$$\mathcal{D}_d = \{(x_i, g_i, r_i^{human})\}_{i=1}^{N_d},$$

where x_i is the dialogue context, g_i is the target group condition, and r_i^{human} is the real human response. Using Gemini 2.5 Flash (Comanici et al., 2025) as an LLM-as-a-judge labeler, we assign

one of the five behavior labels to each real human response and each generated model response (see Appendix D for cross-model validation). We write M_m for model m, and $h(\cdot)$ for the behavior-labeling function that maps a response to one label in $\mathcal{Y}$. For verbalization and generation, we evaluate Gemini 3.1 Pro Preview (Google, 2026), Claude Sonnet 4.6 (Anthropic, 2026b), Qwen3-14B (Yang et al., 2025), and Gemma-4-31B (Google DeepMind, 2026). For open-weight models, we compare base and instruct variants. We do not use explicit thinking or reasoning modes.

Condition 1: Verbalized Distribution (C1) C1 serves as a diagnostic rather than a response-generation method. Given the dialogue context and group condition, the model assigns probabilities to the five behavior types for the next response:

$$P_V^{m,d}(y \mid x_i, g_i), \quad \sum_{y \in \mathcal{Y}} P_V^{m,d}(y \mid x_i, g_i) = 1.$$

We average these sample-level probability vectors within each corpus-specific target group to obtain $P_V^{m,d}(y \mid g)$. This condition measures the model's stated distributional knowledge about group-specific behavior.

Condition 2: Direct Group-Conditioned Response Generation (C2) C2 serves as our baseline generation setting. The model is given the dialogue context and group condition, and is asked to generate one response as a member of that group. For example, in CHILDES, the model is given the previous dialogue context and a group condition such as "a 5-year-old child," and is asked to generate the child's next utterance. No explicit behavior distribution is provided:

$$r_i^{(\mathrm{C2})} = M_m(x_i, g_i), \quad \hat{y}_i^{(\mathrm{C2})} = h(r_i^{(\mathrm{C2})}).$$

This condition tests whether the model can generate responses that follow the real response patterns of the target group, using only the group condition and dialogue context.

Condition 3a: Verbalized Distribution-Conditioned Response Generation (C3a) C3a tests whether explicitly providing the model's own verbalized distribution helps generation. After obtaining the verbalized distribution from C1, we feed that distribution back into the generation prompt and ask the model to generate one natural response:

$$r_i^{(\mathrm{C3a})} = M_m(x_i, g_i, P_V^{m,d}(\cdot \mid x_i, g_i)).$$

This condition examines whether the realization gap is simply due to the absence of explicit distributional information in the prompt.

Condition 3b: Candidate-Pool Response Generation (C3b) C3b addresses a possible limitation of single-response generation. For each input, we generate $K = 5$ candidate responses:

$$\mathcal{R}_i^{(\mathrm{C3b})} = \{r_{i,1}^{(\mathrm{C3b})}, \ldots, r_{i,K}^{(\mathrm{C3b})}\}, \quad \hat{y}_{i,k}^{(\mathrm{C3b})} = h(r_{i,k}^{(\mathrm{C3b})}).$$

This condition tests whether response types that are rare in single-response generation still appear among the candidate responses.

Condition 3c: Externally Sampled Response-Type Generation (C3c) C3c is motivated by the possibility that a model may verbalize a reasonable behavior distribution, but fail to sample response types according to that distribution during generation (Zhang et al., 2024; Xiao et al., 2026; Gu et al., 2025). Instead of allowing the model to decide which behavior type to generate, we externally sample a target behavior label from the model's verbalized distribution (see Appendix F):

$$z_i^{(\mathrm{C3c})} \sim P_V^{m,d}(\cdot \mid g),$$

$$r_i^{(\mathrm{C3c})} = M_m(x_i, g_i, z_i^{(\mathrm{C3c})}).$$

Distributional Evaluation We compare distributions using Total Variation Distance:

$$\mathrm{TVD}(P, Q) = \frac{1}{2} \sum_{y \in \mathcal{Y}} |P(y) - Q(y)|.$$

TVD measures the overall distance between two response-type distributions. To interpret which response types drive this distance, we also analyze label-wise signed gaps, computed as the difference in probability for each label.

4 Results and Analysis

4.1 Verbalized and generated-response distributions show distinct gaps

Across the three domains, verbalized distributions and direct-generation distributions both deviate from empirical human behavior, but the patterns of deviation are distinct. This suggests that explicitly predicting a group's response distribution is not equivalent to realizing that distribution during response generation. We therefore evaluate three distributions separately: **verbalized distributions,**

Group	Model	C1	C2	C3a		C3b		C3c		Avg.
		$V \to G$	$B \to G$	$G_{3a} \to V$	$G_{3a} \to G$	$G_{3b} \to V$	$G_{3b} \to G$	$G_{3c} \to V$	$G_{3c} \to G$	
Age 2	Claude	0.247	0.366	0.425	0.374	0.221	**0.118**	0.319	**0.230**	0.288
	Gemini	0.177	0.418	0.438	0.412	0.128	**0.148**	0.337	**0.223**	0.285
	Gemma-Base	0.298	0.152	0.378	0.219	0.275	**0.092**	0.310	0.193	0.240
	Gemma-Inst	0.238	0.298	0.379	0.351	0.180	**0.088**	0.335	**0.191**	0.258
	Qwen-Base	0.271	0.182	0.298	0.247	0.390	0.201	0.285	0.292	0.271
	Qwen-Inst	0.235	0.284	0.297	0.211	0.213	**0.149**	0.276	**0.220**	0.236
Age 5	Claude	0.163	0.241	0.415	0.469	0.257	0.274	0.268	0.332	0.302
	Gemini	0.084	0.220	0.280	0.240	0.130	**0.154**	0.149	**0.114**	0.171
	Gemma-Base	0.152	0.127	0.312	0.183	0.206	0.204	0.238	**0.120**	0.193
	Gemma-Inst	0.128	0.218	0.338	0.270	0.162	**0.186**	0.233	**0.217**	0.219
	Qwen-Base	0.141	0.557	0.633	0.568	0.447	**0.376**	0.536	**0.471**	0.466
	Qwen-Inst	0.140	0.381	0.522	0.400	0.361	**0.236**	0.410	**0.288**	0.342
Age 9	Claude	0.119	0.251	0.483	0.478	0.249	**0.237**	0.332	0.311	0.308
	Gemini	0.077	0.152	0.289	0.249	0.181	0.154	0.146	**0.094**	0.168
	Gemma-Base	0.174	0.154	0.240	0.097	0.180	0.251	0.246	**0.114**	0.182
	Gemma-Inst	0.154	0.153	0.390	0.299	0.192	0.178	0.245	0.179	0.224
	Qwen-Base	0.178	0.557	0.706	0.554	0.501	**0.400**	0.583	**0.431**	0.489
	Qwen-Inst	0.198	0.371	0.631	0.432	0.465	**0.294**	0.535	**0.337**	0.408
Column Avg.		0.176	0.282	0.414	0.336	0.263	0.208	0.321	0.242	0.280

Table 1: **CHILDES:** TVD results across conditions. V denotes the model's verbalized distribution, B the baseline distribution estimated from directly generated responses, and G empirical gold. Confidence intervals and noise floor estimates are reported in Table 7 and 6. Bold indicates that C3b or C3c $G \to G$ is lower than C2.

empirical (gold) distributions, and **distributions from direct group-conditioned response generation**, which we refer to as direct generation (baseline) for brevity.

CHILDES In Table 1, C1 is closer to the empirical gold distribution than C2 in 13 out of 18 model–age pairs. This shows that, in most cases, the child response distribution verbalized by the model is closer to empirical child behavior than its direct generation. This pattern is especially consistent for the closed models and Qwen-Inst. Claude, Gemini, and Qwen-Inst show lower distances for C1 than for C2 across all age groups. Gemma-base shows the opposite pattern across all age groups, where C2 is closer to gold than C1, while Qwen-Base shows lower distance for C1 only at Age 5 and Age 9. These results suggest that failures in direct generation cannot be explained only by the model being unable to state a plausible distribution over child response behavior.

By age group, the distance between the verbalized distribution and gold distribution is higher for Age 2, suggesting that response distributions for very young children may be harder for models to predict. The label-level gaps in Figure 1 show how direct generation differs from empirical gold distributions. For Age 2, *direct_answer* is over-produced, while *topic_shift* and *non_answer* are under-produced. This suggests that incomplete or off-path responses are often turned into clearer and more task-oriented answers. In contrast, for Age 5 and Age 9, over-production of *elaborated* responses becomes more prominent. In particular, Qwen-family models reduce *direct_answer* and increase *elaborated* responses, generating older children's responses as longer and more explanatory than they are in the empirical data. In addition, *topic_shift* is almost consistently under-realized across all age groups. Overall, models can verbalize child response distributions to some extent, but they do not realize them in direct generation. This gap appears not only as an aggregate distance, but also as age-dependent label-level bias.

Bridge Table 2 shows a more mixed pattern than CHILDES. Based on point estimates, C1 is closer to the empirical gold distribution than C2 in 7 out of 12 model–group pairs, with this pattern appearing in 3 out of 6 Novice pairs and 4 out of 6 Expert pairs. This suggests that verbalized distributions can be closer to empirical expert–novice behavior than direct generation, especially in the Expert condition, but the pattern is weaker and more group-dependent than in CHILDES. In Novice, C2 is closer to gold for Gemma-Inst, Qwen-Base, and Qwen-Inst, while in Expert, C2 is closer for Gemini and Qwen-Inst. Thus, Bridge still supports a verbalization–generation gap, but the direction of the gap varies by model and target group.

The label-level analysis helps explain this mixed pattern. Figure 1 shows how direct generation differs from empirical gold distributions. In the Expert condition, direct generation generally under-produces *direct_answer* and over-produces *elaborated* responses. This pattern appears for Claude, Gemma-Inst, Qwen-Base, and Qwen-Inst, suggesting that expert tutor responses may be generated as more explanatory and complete than they are in the empirical data. In contrast, Gemma-base shows the opposite pattern, producing more *minimal* and *direct_answer* responses and fewer *elabo-*

Group	Model	C1	C2	C3a		C3b		C3c		Avg.
		$V \to G$	$B \to G$	$G_{3a} \to V$	$G_{3a} \to G$	$G_{3b} \to V$	$G_{3b} \to G$	$G_{3c} \to V$	$G_{3c} \to G$	
Novice	Claude	0.182	0.311	0.279	0.295	0.240	**0.241**	0.215	**0.231**	0.249
	Gemini	0.113	0.202	0.196	0.171	0.166	**0.150**	0.076	**0.123**	0.150
	Gemma-Base	0.202	0.284	0.221	0.228	0.207	**0.124**	0.267	**0.178**	0.214
	Gemma-Inst	0.259	0.224	0.359	0.166	0.309	**0.124**	0.232	**0.203**	0.235
	Qwen-Base	0.236	0.194	0.222	0.348	0.260	0.199	0.125	0.273	0.232
	Qwen-Inst	0.179	0.175	0.228	0.219	0.447	0.386	0.232	0.261	0.266
Expert	Claude	0.181	0.304	0.524	0.356	0.481	0.312	0.449	**0.281**	0.361
	Gemini	0.081	0.054	0.285	0.230	0.239	0.207	0.249	0.234	0.197
	Gemma-Base	0.214	0.387	0.071	0.222	0.214	**0.113**	0.173	**0.316**	0.214
	Gemma-Inst	0.152	0.326	0.449	0.345	0.412	**0.308**	0.252	**0.211**	0.307
	Qwen-Base	0.185	0.317	0.471	0.334	0.226	**0.123**	0.345	**0.234**	0.279
	Qwen-Inst	0.486	0.284	0.688	0.272	0.645	**0.181**	0.519	**0.182**	0.407
Column Avg.		0.206	0.255	0.333	0.266	0.321	0.206	0.261	0.227	0.259

Table 2: **Bridge:** TVD results across conditions. V denotes the model's verbalized distribution, B the baseline distribution estimated from directly generated responses, and G empirical gold. Confidence intervals and noise floor estimates are reported in Table 8 and 6. Bold indicates that C3b or C3c $G \to G$ is lower than C2.

rated responses. In the Novice condition, the distortion is more model-specific. Claude and Gemini over-produce *direct_answer*, Gemma-family models over-produce *minimal*, and Qwen-family models relatively increase *elaborated* responses. However, *topic_shift* is under-realized across almost all models. In other words, models do not generate enough topic-shift-like behavior that exists in the empirical novice distribution, and instead tend to generate more cooperative and context-following responses.

ICNALE Table 3 also shows a verbalization–generation gap in L2 learner dialogue. C1 is closer to the empirical gold distribution than C2 in 17 out of 24 model–level pairs. This pattern holds for most models across proficiency levels, including 4 out of 6 models at A2_0, 4 out of 6 at B1_1, 4 out of 6 at B1_2, and 5 out of 6 at B2_0. Thus, the ICNALE results strengthen the claim that models often verbalize learner-response distributions closer to empirical behavior than they realize through direct generation, although the effect still varies by model family. Figure 1 shows that direct generation under-produces *minimal* responses and over-produces *elaborated* responses across almost all proficiency levels and models. In particular, Claude, Gemini, Gemma-Inst, Qwen-Base, and Qwen-Inst show this pattern strongly across most levels from A2_0 to B2_0. This pattern shows that models may simplify learner proficiency into a longer and more explanatory response style. Real learners may give short responses in some situations, even when their proficiency level is high. However, even at the B1 and B2 levels, models shift probability mass away from *minimal* and *direct_answer* responses and toward *elaborated* responses. Therefore, generated responses may look natural or helpful at the individual-response level, while still being systematically different from the empirical learner behavior distribution. Gemma-base is an exception because it generates more *non_answer* responses and shows a weaker *elaborated* bias. The most consistent failure mode is still *minimal* under-production and *elaborated* over-production.

4.2 Models do not realize their own verbalized distributions

Section 4.1 showed that direct generation diverges from empirical human distributions. However, this alone does not tell us whether models lack knowledge of group behavior distributions or whether they fail to realize that knowledge in generation. We therefore compare each model's direct generation distribution with its own verbalized distribution. Figure 2 shows the label-wise signed gap between direct generation and verbalized distributions, defined as $P_{\text{direct}}(y) - P_{\text{verbalized}}(y)$. Positive values mean that direct generation produces a label more often than the model verbalized, while negative values mean that it produces the label less often. In **CHILDES**, the direction of the gap is similar to the gold comparison in Section 4.1. Most models generate fewer *minimal* responses than they verbalize. For Age 2, this probability mass mainly shifts toward *direct_answer*, while for Age 5 and Age 9, it shifts toward *elaborated* responses. This shows that even when models verbalize that short or incomplete child responses are likely, direct generation tends to produce clearer and more contentful responses. In **ICNALE**, this realization gap is stronger. Across almost all proficiency levels, direct generation produces fewer *minimal* and *direct_answer* responses than the verbalized distribution, and more *elaborated* responses. Thus, models can verbalize that learner responses may include short answers, but generation realizes learner re-

Group	Model	C1	C2	C3a		C3b		C3c		Avg.
		$V \to G$	$B \to G$	$G_{3a} \to V$	$G_{3a} \to G$	$G_{3b} \to V$	$G_{3b} \to G$	$G_{3c} \to V$	$G_{3c} \to G$	
A2_0	Claude	0.318	0.335	0.524	0.292	0.308	**0.127**	0.406	**0.230**	0.318
	Gemini	0.325	0.329	0.508	0.280	0.302	**0.183**	0.307	**0.075**	0.289
	Gemma-Base	0.371	0.385	0.277	0.234	0.165	0.470	0.289	**0.215**	0.301
	Gemma-Inst	0.321	0.373	0.358	0.230	0.071	**0.334**	0.269	**0.205**	0.270
	Qwen-Base	0.422	0.385	0.411	0.292	0.148	**0.333**	0.522	**0.286**	0.350
	Qwen-Inst	0.401	0.366	0.436	0.242	0.203	0.491	0.405	**0.205**	0.344
B1_1	Claude	0.221	0.392	0.468	0.358	0.362	**0.252**	0.439	**0.330**	0.353
	Gemini	0.217	0.369	0.445	0.330	0.325	**0.241**	0.309	**0.199**	0.304
	Gemma-Base	0.337	0.267	0.409	0.309	0.132	0.312	0.344	**0.180**	0.286
	Gemma-Inst	0.261	0.375	0.531	0.318	0.096	**0.236**	0.395	**0.244**	0.307
	Qwen-Base	0.378	0.403	0.622	0.341	0.195	**0.261**	0.639	**0.358**	0.400
	Qwen-Inst	0.387	0.364	0.605	0.295	0.262	0.522	0.576	**0.273**	0.411
B1_2	Claude	0.210	0.376	0.489	0.376	0.368	**0.255**	0.414	**0.303**	0.349
	Gemini	0.168	0.359	0.410	0.337	0.336	**0.264**	0.291	**0.218**	0.298
	Gemma-Base	0.334	0.231	0.388	0.166	0.268	**0.208**	0.306	**0.101**	0.250
	Gemma-Inst	0.262	0.381	0.578	0.359	0.182	**0.147**	0.399	**0.194**	0.313
	Qwen-Base	0.364	0.381	0.641	0.362	0.277	**0.188**	0.593	**0.316**	0.390
	Qwen-Inst	0.367	0.330	0.641	0.323	0.274	0.339	0.551	**0.235**	0.383
B2_0	Claude	0.186	0.392	0.481	0.392	0.363	**0.274**	0.407	**0.318**	0.352
	Gemini	0.141	0.369	0.322	0.364	0.271	**0.314**	0.226	**0.267**	0.284
	Gemma-Base	0.320	0.227	0.465	0.206	0.423	**0.218**	0.370	**0.111**	0.293
	Gemma-Inst	0.239	0.386	0.582	0.386	0.208	**0.106**	0.405	**0.233**	0.318
	Qwen-Base	0.258	0.392	0.544	0.381	0.277	**0.188**	0.522	**0.358**	0.365
	Qwen-Inst	0.340	0.375	0.686	0.386	0.316	**0.238**	0.606	**0.307**	0.407
Column Avg.		0.298	0.356	0.493	0.315	0.256	0.271	0.416	0.240	0.331

Table 3: **ICNALE:** TVD results across conditions. V denotes the model's verbalized distribution, B the baseline distribution estimated from directly generated responses, and G empirical gold. Confidence intervals and noise floor estimates are reported in Table 9 and 6. Bold indicates that C3b or C3c $G \to G$ is lower than C2.

sponses as longer and more explanatory. In **Bridge**, the pattern is more model-dependent. In the Expert condition, direct generation generally produces fewer *direct_answer* responses and more *elaborated* responses than the verbalized distribution. In the Novice condition, model differences are larger, but *topic_shift* is generally under-realized relative to the verbalized distribution.

4.3 Explicit distribution conditioning does not guarantee distribution matching

We showed that direct generation diverges from both empirical gold and the model's own verbalized distribution. We therefore test whether C3a, which provides the model's verbalized label distribution in the prompt, can reduce this gap. C3a is often closer to gold than direct generation, especially in the ICNALE results, but it still does not consistently eliminate the gap across domains. This suggests that the mismatch is not only due to missing target distribution information. Even when the distribution is provided, models do not reliably realize it as the aggregate distribution of generated responses. Across domains, the main biases from direct generation often remain (see Appendix Figure 3). In **CHILDES**, C3a does not fully remove the Age 2 *direct_answer* bias or the Age 5/9 *elaborated* bias. In **ICNALE**, C3a improves the aggregate TVD-to-gold in many model–level pairs, but the label-level gaps show that minimal under-production and elaborated over-production are not fully removed. In **Bridge**, its effect varies by target group and model, and *topic_shift* under-realization remains in Novice. Overall, C3a shows that distribution conditioning alone is not enough. Models still prefer certain response types and do not reliably turn the provided distribution into aggregate generated behavior.

4.4 Generation-process interventions localize the gap

Section 4.3 showed that providing the verbalized distribution in the prompt does not consistently close the generation gap. In this section, we use C3b and C3c to analyze whether this gap comes from response type selection, or from difficulty generating and maintaining certain response types. For C3b, we evaluate the full candidate pool. For each input, the model generates five candidate responses, allowing us to test whether response types under-produced in single direct generation are still available among sampled candidates. C3c instead conditions generation on an externally sampled label, testing whether the model can realize a specified response type when it is explicitly provided.

At the aggregate TVD level (Table 1–3), C3b reduces the gap. In **CHILDES**, C3b improves over direct generation in 12/18 model–age pairs, with large gains for Age 2. C3c also improves over direct generation in many CHILDES cases, but its gains are less concentrated than C3b. It helps for several Age 5 and Age 9 settings. In **Bridge**, C3b improves in 8/12 model–group pairs. In the **ICNALE** results, C3b improves over direct generation in 19 out of 24 model–level pairs, including 4 out of 6 A2_0 models. C3c is even more consistent,

improving over direct generation in all 24 model–level pairs. Thus, for **ICNALE**, generation-process interventions substantially reduce TVD-to-gold.

The label-level heatmaps (Figure 1–5) show that C3b and C3c often reduce the response-type biases observed in direct generation. We interpret these changes together with the empirical gold label proportions in Table 5. The key point is that these interventions do not help by making responses shorter or less *elaborated*. Instead, they help when they move probability mass toward the empirical gold label distribution of the target group.

In **CHILDES**, this effect is clearest for Age 2. Direct generation over-produces *direct_answer* and under-produces *non_answer* and *topic_shift*, while C3b substantially reduces this *direct_answer* inflation. This shift is meaningful because Age 2 gold responses contain substantial *non_answer* and *topic_shift* mass (.273 and .200), while *elaborated* responses are rare (.091). Thus, C3b moves generation closer to the incomplete or off-path responses found in real young children's behavior. For Age 5 and Age 9, C3b also reduces *elaborated* overproduction for several models, but the gold *elaborated* proportions increase with age (.259 and .350). Therefore, improvement should not be interpreted as simply reducing elaboration. A better generation distribution is one that produces the amount of elaboration expected for each age group.

In **ICNALE**, C3b and C3c reduce the strong *elaborated* bias in direct generation. The gold proportions clarify that elaboration itself is not the problem. Empirical learner responses are already frequently *elaborated* (.516–.602). Rather, direct generation places too much probability mass on *elaborated* responses relative to gold. By reducing this excessive concentration, C3b and C3c can move generated responses toward a balance closer to the empirical learner distribution, rather than simply making learner responses less detailed.

In **Bridge**, the Novice–Expert contrast shows why the same label-level change must be interpreted relative to gold. Novice responses are mostly *direct_answer* (.530) and *minimal* (.278), with rare *elaborated* responses (.059), so reducing overly explanatory or cooperative responses can correct the novice distribution. Expert responses, however, are dominated by *elaborated* responses (.637), with substantial *direct_answer* mass (.306). Thus, for Expert, the goal is not to reduce elaboration uniformly, but to preserve explanation and scaffolding while reducing mismatch. Appendix C further shows that instruction tuning changes verbalized and generated-response distributions differently by comparing base and instruction-tuned variants of the open-weight models. Overall, these results support our main claim that generation-process interventions should be evaluated by whether they improve group-level distributional faithfulness, rather than by whether they make individual responses look more or less detailed.

5 Discussion and Conclusion

Our results show that **describing is not generating**. Verbalized and directly generated response-type distributions often differ, and feeding the verbalized distribution back into the prompt does not consistently close this gap. Generation-process interventions reduce the gap more often, especially when they move probability mass toward the target group's empirical gold label proportions rather than simply making responses shorter or less elaborated. Overall, LLM-based human group simulation should be evaluated as a problem of distributional realization. Models may describe how a group is likely to respond, but still fail to generate response types in the right proportions. Simulated human groups should therefore be evaluated by aggregate distributional faithfulness, not only by individual response plausibility.

Broader Implications Our findings have implications for using LLMs as simulated users, synthetic participants, or role-based dialogue partners. In settings such as child-facing systems, tutoring, patient interaction, and social simulation, real users may produce short, incomplete, or off-topic responses. If simulated users are more cooperative or explanatory than real users, systems tested with them may be evaluated in an unrealistically easy setting and may appear more effective than they would be in real interactions. LLM-based silicon samples should therefore be validated against human data when possible, using multiple samples, response-type characterization, and aggregate distributional comparisons rather than single direct generations. Our results also suggest a tension between assistant training and faithful simulation. Models trained to be helpful and coherent may underproduce the short, incomplete, or topic-shifting responses found in real human behavior. Improving group simulation may require training and evaluation objectives that reward distributional faithfulness, not only individual response quality.

Limitations

Our findings should not be seen as general claims about all human groups or all types of social simulation. We map open-ended dialogue responses into five high-level behavior labels. This label space helps us compare results across domains, but it cannot capture all parts of dialogue behavior. For example, emotion, style, and multimodal cues are not the main focus of our evaluation.

A further limitation is that future models may be trained or tuned using the same evaluation datasets and response labels used in this work. If a model performs better on this specific evaluation, it may reflect adaptation to our setup rather than a general improvement in faithful human group simulation. Future work should test distributional faithfulness with new domains, newly collected human data, and other ways of categorizing response behavior.

Ethical Considerations

This work evaluates LLM simulations at the level of group response distributions. It should not be used to judge, classify, or rank individual people. For example, a single response should not be used to infer a person's age, ability level, expertise, or social identity. LLM simulations based on group labels should also be interpreted carefully. Model outputs may make a target group look simpler or more stereotyped than it really is. This is especially important for settings involving children, learners, patients, or other groups where simulation errors could affect system design or evaluation. Our goal is not to discourage LLM-based simulation, but to support its responsible use. LLM simulations can be useful for early testing and analysis, but they should not be the only evidence. When possible, researchers should compare them with real human data and check whether the generated responses follow similar group-level behavior patterns.

References

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98, Minneapolis, Minnesota. Association for Computational Linguistics.

Anthropic. 2026a. System card: Claude haiku 4.5.

Anthropic. 2026b. System card: Claude sonnet 4.6.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023a. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023b. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore. Association for Computational Linguistics.

Jiyun Chun, Eric Fosler-Lussier, Michael White, and Andrew Perrault. 2026. Beyond length: Context-aware expansion and independence as developmentally sensitive evaluation in child utterances. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6030, Rabat, Morocco. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Google. 2026. Gemini 3.1 pro model card.

Google DeepMind. 2026. Gemma 4 model card.

Jia Gu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2025. Do LLMs play dice? exploring probability distribution sampling in large language models for behavioral simulation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5375–5390, Abu Dhabi, UAE. Association for Computational Linguistics.

Seungju Han, Beomsu Kim, and Buru Chang. 2022. Measuring and improving semantic diversity of dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 934–950, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. 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)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Shin'ichiro Ishikawa. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*, 1 edition. Routledge.

Yanis Labrak, David Grünert, Séverin Baroudi, Jiyun Chun, Pawel Cyrta, Sergio Burdisso, Ahmed Hassoon, David Liu, Adam Rothschild, Reed Van Deusen, Petr Motlicek, Andrew Perrault, Ricard Marxer, and Thomas Schaaf. 2026. Generating synthetic doctor-patient conversations for long-form audio summarization. *Preprint*, arXiv:2604.06138.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Brian MacWhinney. 2000. The childes project: Tools for analyzing talk (third edition): Volume i: Transcription format and programs, volume ii: The database. *Computational Linguistics*, 26(4):657–657.

Arunima Maitra, Dorothea French, and Katharina von der Wense. 2025. Dialogue acts as a lens on human–LLM interaction: Analyzing conversational norms in model-generated responses. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 317–325, Suzhou, China. Association for Computational Linguistics.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49, Albuquerque, New Mexico. Association for Computational Linguistics.

OpenAI. 2026. Gpt-5.4 thinking system card.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *Neural Information Processing Systems*.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. 2025. PersonaGym: Evaluating persona agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022, Suzhou, China. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Preethi Seshadri, Samuel Cahyawijaya, Ayomide Odumakinde, Sameer Singh, and Seraphina Goldfarb-Tarrant. 2026. Lost in simulation: Llm-simulated users are unreliable proxies for human users in agentic evaluations. *Preprint*, arXiv:2601.17087.

Katherine Stasaski and Marti Hearst. 2022. Semantic diversity in dialogue with natural language inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–98, Seattle, United States. Association for Computational Linguistics.

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

Angelina Wang, Jamie Morgenstern, and John Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7:400–411.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024a. NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15183–15201, Bangkok, Thailand. Association for Computational Linguistics.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024b. Bridging the novice-expert gap via models of decision-making: A

case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199, Mexico City, Mexico. Association for Computational Linguistics.

Tim Z. Xiao, Johannes Zenn, Zhen Liu, Weiyang Liu, Robert Bamler, and Bernhard Schölkopf. 2026. Flipping against all odds: Reducing llm coin flip bias via verbalized rejection sampling. *Preprint*, arXiv:2506.09998.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. 2025. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *Preprint*, arXiv:2510.01171.

Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, Zico Kolter, and Daphne Ippolito. 2024. Forcing diffuse distributions out of language models. *Preprint*, arXiv:2404.10859.

A Dataset & Preprocessing

We use three dialogue datasets in our experiments: CHILDES, Bridge, and ICNALE. For all datasets, we convert the original dialogues into a context-response format. In this format, one *turn* does not mean a single utterance. Instead, one turn means one evaluation instance that contains a short dialogue context and the next human response.

CHILDES Child Language Data Exchange System (CHILDES) is a child language database used for studying child language development and interaction. We use 20 corpora (Chun et al., 2026) from CHILDES: Bliss, Bohannon, Braunwald, Brown, Champaign, Clark, Demetras1, Demetras2, EllisWeismer, Gleason, HSLLD, Kuczaj, McCune, Nelson, NewmanRatner, OCSC, Post, Sachs, Suppes, and Warren.

For CHILDES, each processed turn comes from a dialogue involving a child and an adult. It consists of six previous utterances as the dialogue context and one target child response. In other words, the model is given the previous six utterances, and the next child utterance is used as the target human response. We use 900 turns for each child level, resulting in 2,700 turns in total. Since each turn contains six context utterances, this corresponds to 16,200 context utterance positions. If we also include the target child responses, the processed data contains 18,900 utterance positions in total. These counts are based on the processed turn format, not on the number of unique utterances in the original CHILDES corpora.

Bridge Bridge is a tutoring dialogue dataset designed to compare novice and expert tutoring behavior. The dataset contains math tutoring conversations, including student errors, tutor responses, expert tutor responses, and tutoring strategy information. We use Bridge as an expert-novice interaction setting to evaluate whether models can describe and generate response behavior that matches different tutor groups.

We also convert Bridge into context-response turns. Each turn contains the previous dialogue context and the next tutor response. Unlike CHILDES and ICNALE, however, the number of utterances in each Bridge turn is not fixed. This is because the original Bridge data already provides a structured dialogue history for each tutoring instance, rather than using a fixed-size sliding context window. The context usually contains three or four utterances, and the target tutor response can contain more than one utterance. Therefore, for Bridge, we report the number of processed turns rather than assuming a fixed number of utterances per turn. We use 644 turns for each tutor level.

ICNALE ICNALE, the International Corpus Network of Asian Learners of English, is a learner corpus that contains English data from Asian learners of English. We use the spoken dialogue portion of ICNALE to study learner response behavior across English proficiency levels.

For ICNALE, each processed turn consists of six previous utterances as the dialogue context and one target learner response. Thus, as in CHILDES, the model is given six context utterances, and the next learner utterance is used as the target human response. The number of turns differs by proficiency level. A2_0 has 162 turns, B1_1 has 176 turns, B1_2 has 412 turns, and B2_0 has 176 turns. This gives 926 ICNALE turns in total. Since each turn contains six context utterances, the ICNALE data contains 5,556 context utterance positions. Including the target learner responses, it contains 6,482 utterance positions in total.

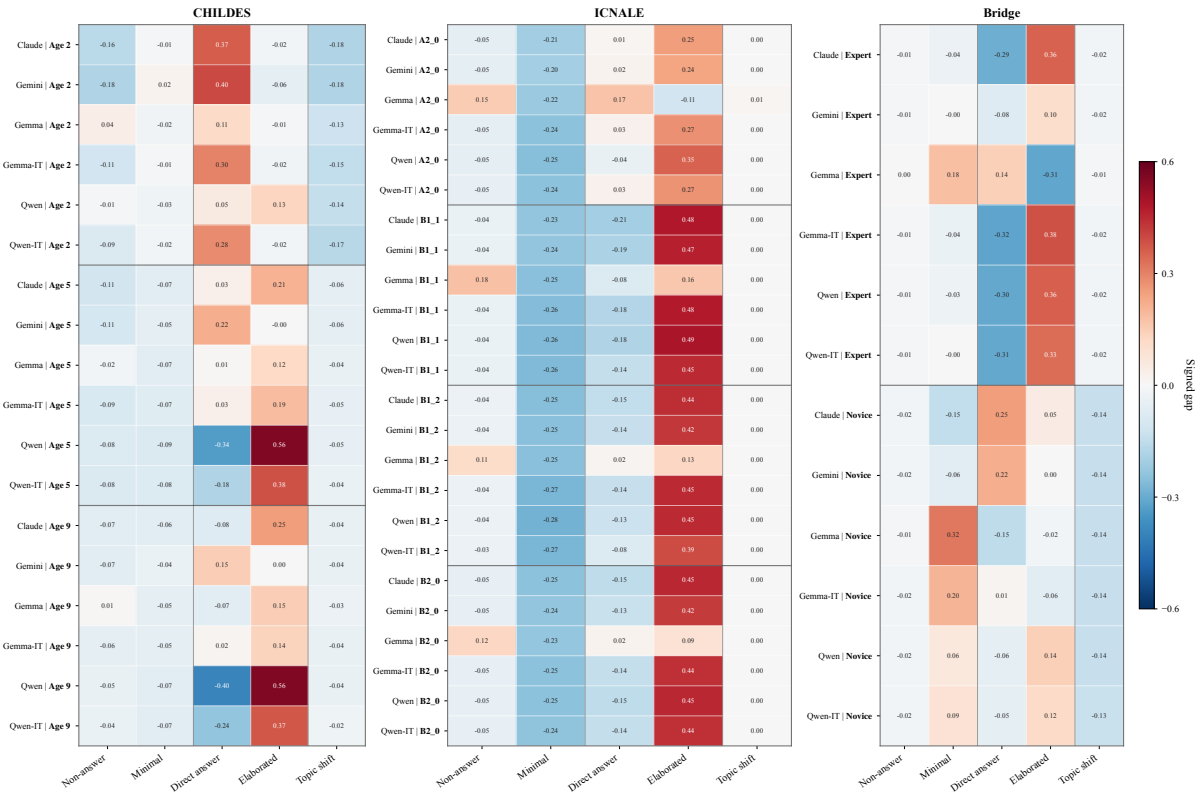

Figure 1: Label-wise signed gaps between the **direct-generation response distribution** and the **empirical gold distribution**. Values are computed as $P_{\text{direct}}(y) - P_{\text{gold}}(y)$. Positive values indicate over-production in the direct-generation response distribution, while negative values indicate under-production.

B Five Response-Behavior Labels

Table 4 provides the operational definitions of the five response-behavior labels used in our distributional analyses. We use the same five-label space across all corpora so that different interaction settings can be compared under a common set of broad response types. These labels are not intended to introduce a new discourse taxonomy for each corpus. Our goal is not to classify every fine-grained dialogue function in each domain. Instead, our goal is to evaluate whether models can describe the distribution of response types observed in real human groups, and whether they can realize that distribution during generation.

The five labels capture broad response patterns that can appear in all three corpora. We use *non_answer* for responses that do not provide a meaningful or interpretable answer, *minimal* for very short responses with little content, *direct_answer* for short but meaningful responses that address the current question or utterance, *elaborated* for responses that add explanation or detail, and *topic_shift* for responses that move away from the current question or task. These labels are meant to capture large differences in response behavior, rather than all domain-specific details of dialogue.

This shared label space is useful because it lets us compare response depth and elaboration across corpora. A central question in our analysis is whether models provide the right level of detail compared to real human groups. Real people may sometimes give very short responses, short direct answers, or more elaborated responses depending on the context. Because LLMs are often trained to be helpful and detailed, they may over-produce *elaborated* responses, or under-produce shorter responses such as *minimal* or *direct_answer*. By separating *minimal*, *direct_answer*, and *elaborated*, we can test whether models match the real distribution of response depth, not just whether their responses seem relevant.

The definitions are adapted to each corpus because the response roles differ across settings. In CHILDES, the response comes from a child answering an adult. In ICNALE, the response comes from an English learner in a role-play situation. In the expert–novice tutoring corpus, the response comes from a tutor responding to a student. Thus,

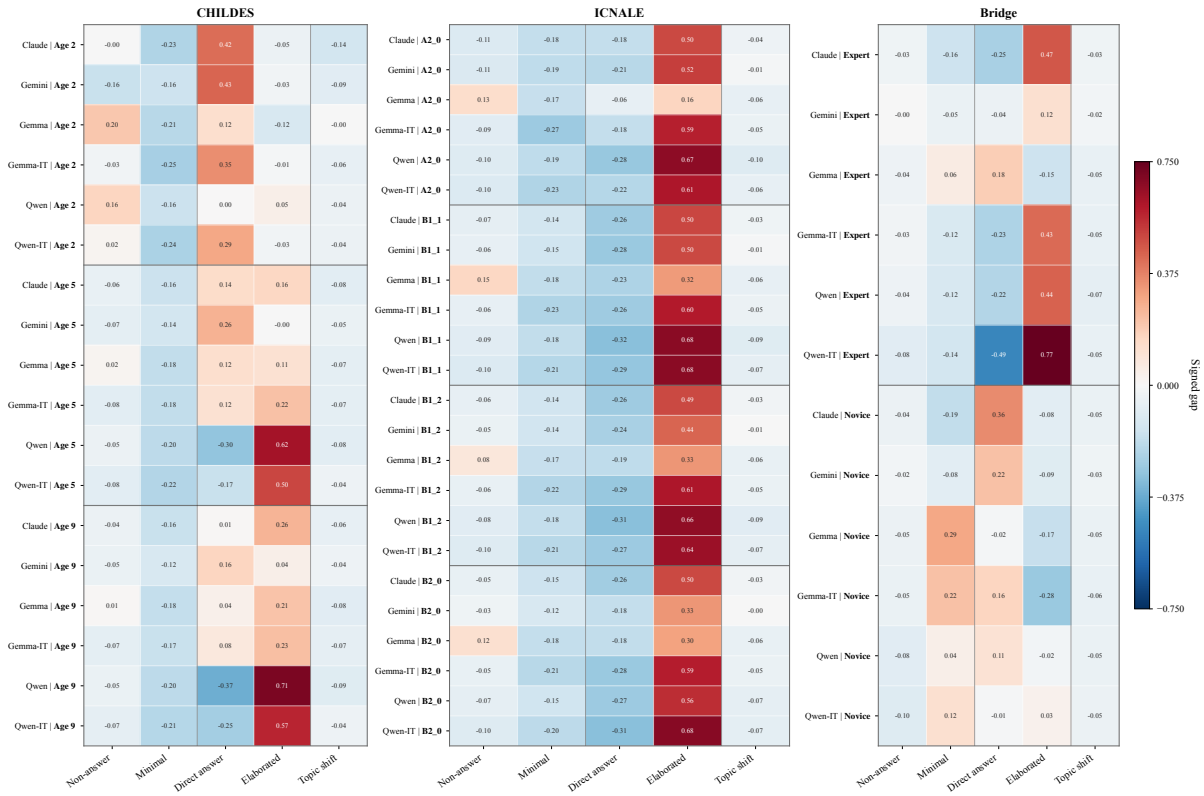

Figure 2: Label-wise signed gaps between the **direct-generation response distribution** and the **model's own verbalized distribution**.

the same label may appear differently across corpora. For example, an *elaborated* response may involve added detail from a child, supporting reasons from a learner, or explanation and scaffolding from a tutor. However, in all cases, it refers to a relevant response that goes beyond a brief answer. We therefore use the same broad labels, but describe how each label appears in each interaction setting.

We use this shared label space instead of separate corpus-specific taxonomies because our main goal is cross-domain distributional comparison. More detailed taxonomies, such as discourse-act labels or tutoring-strategy labels, can capture richer behavior within a single domain. However, they are usually designed for a specific role, task, or annotation goal, and they do not provide a common label set for comparing child responses, learner role-play responses, and tutor feedback. Using one shared five-label space allows us to compare empirical human distributions, model-verbalized distributions, and model-generated distributions in the same way across domains. More detailed domain-specific analyses can be used as complementary checks.

C Instruction tuning changes verbalized and generated-response distributions differently

We compare each open-weight base model with its instruction-tuned counterpart from the same model family. We use lower TVD-to-gold as the criterion for improvement. Based on point estimates in Tables 1–3, instruction tuning improves verbalization in 12 of 18 base/instruction comparisons. However, these improvements do not consistently transfer to direct generation. Among the 12 cases where instruction tuning lowers the C1 distance, 6 cases have a higher C2 distance. The reverse pattern also appears. In all 6 cases where instruction tuning increases the C1 distance, it lowers the C2 distance. This shows that instruction tuning changes verbalized distribution estimation and direct response generation in different ways, rather than improving both uniformly.

The effect is also model-family dependent. For Gemma, instruction tuning improves verbalization in 8 of 9 comparisons, but improves direct generation in only 4 of 9. For Qwen, the pattern is nearly reversed. Instruction tuning improves verbalization in only 4 of 9 comparisons, but improves direct

Label	CHILDES	ICNALE	Bridge
non_answer	The child does not answer the adult's question, refuses to answer, or says they do not know.	The learner does not provide an interpretable or meaningful response to the preceding utterance or role-play context.	The tutor does not provide an interpretable or meaningful response to the preceding student utterance or tutoring context.
minimal	A very short response that reacts to the adult's utterance but does not provide enough information to answer the question.	A very short response that mainly acknowledges or reacts to the preceding utterance, but does not present a clear stance, request, decision, proposal, or factual answer.	A very short but interpretable tutor response that only acknowledges, evaluates, or reacts to the student. It includes little or no instructional content.
direct_answer	A short but meaningful response that gives an actual answer to the adult's question. It does not add extra explanation or detail.	A short but contentful response that presents a stance, request, decision, proposal, or factual answer in relation to the preceding utterance or role-play context. It does not include additional explanation or support.	A short but contentful tutor response that addresses the student's utterance or error. It may provide a correction, answer, simple prompt, or next step, but does not include explanation or scaffolding.
elaborated	A relevant response that goes beyond a simple answer and includes extra explanation, detail, or reason.	A relevant response that adds explanation, reason, evidence, example, consequence, condition, or other supporting detail.	A relevant tutor response that provides explanation, reasoning, scaffolding, a hint, an example, a strategy, or detailed corrective feedback to help the student understand or move to the next step.
topic_shift	The child moves away from the adult's question or the current topic.	The learner moves away from the current role-play situation or preceding utterance to a new unrelated topic.	The tutor moves away from the student's current response, the immediate tutoring problem, the student's error, or the learning goal to an unrelated topic, session management, social talk, or a different task.

Table 4: **Definitions of the five response-behavior labels** used across the three corpora.

Corpus	Group	Non-answer	Minimal	Direct	Elaborated	Topic shift
CHILDES	Age 2	.273	.070	.366	.091	.200
CHILDES	Age 5	.137	.090	.449	.259	.066
CHILDES	Age 9	.098	.072	.432	.350	.048
ICNALE	A2_0	.068	.335	.081	.516	.000
ICNALE	B1_1	.045	.284	.108	.562	.000
ICNALE	B1_2	.034	.262	.117	.587	.000
ICNALE	B2_0	.051	.244	.102	.602	.000
Bridge	Novice	.003	.278	.530	.059	.123
Bridge	Expert	.008	.034	.306	.637	.016

Table 5: **Empirical gold response-label proportions** by corpus-specific target group.

generation in 8 of 9. A clear example is Gemma-Inst in CHILDES. Its C1 distance is lower than Gemma-Base for all three age groups, but its C2 distance is higher for Age 2 and Age 5 and changes only marginally for Age 9. These results show that instruction tuning can make a model's stated distribution closer to gold without making its generated distribution closer to gold.

The label-level analysis shows the same separation (Figure 1, 3–5). In **CHILDES**, direct generation still tends to move probability mass away from short, incomplete, or topic-shifting responses and toward *direct_answer* or *elaborated* responses. For Age 2, this appears as *direct_answer* inflation and *topic_shift* under-realization. For Age 5 and Age 9, it appears more as *elaborated* over-production. In **ICNALE**, *minimal* under-production and *elaborated* over-production also remain. **Bridge** is more mixed. Instruction tuning changes which labels are over- or under-produced, but it does not remove the label-level gaps in a consistent way. These results support evaluating verbalization and generation separately.

D Cross-Model Validation of LLM-as-a-Judge Labels

Our main experiments use Gemini 2.5 Flash as the LLM-as-a-judge labeler for assigning response-type labels. To check whether the labels depend strongly on this single judge model, we conduct a cross-model validation check. For each dataset and each target level or group, we randomly sample 100 turns and relabel them using GPT-5.4 (OpenAI, 2026) and claude-haiku-4-5 (Anthropic, 2026a). Cohen's κ with Gemini 2.5 Flash is 0.92 for GPT-5.4 and 0.88 for Claude Haiku. These results show strong agreement across LLM judges using the same label definitions. This suggests that the response-type labels are not specific to one judge model. This check does not replace human annota-

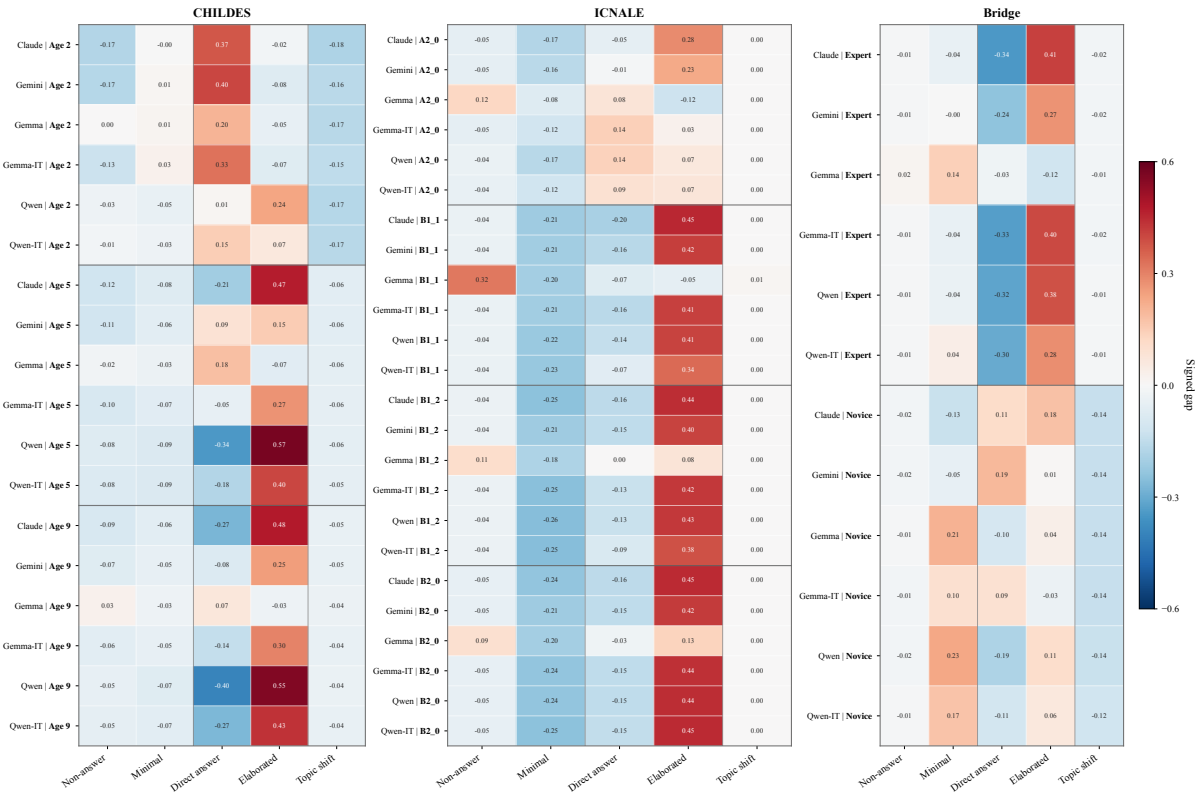

Figure 3: Label-wise signed gaps between the **C3a generated-response distribution** and the **empirical gold distribution**.

tion, but it supports the stability of the label space across different LLM judges.

E Empirical Noise floor and Confidence Intervals

We report 95% bootstrap confidence intervals for the TVD values in Tables 7, 8, and 9. We compute these confidence intervals using 1,000 bootstrap resamples. We also report empirical noise floor. Empirical Noise floor is a noise-floor estimate based only on the empirical gold data. For each target group, we randomly split the labeled human responses into two halves. We then compute the response-type distribution for each half and measure the TVD between the two distributions. This gives an estimate of how much variation we can expect from the finite human dataset itself. Lower noise floor values mean that the empirical gold distribution is more stable for that group.

F External Sampling

In C3c, the model does not choose the response type by itself. Instead, we first use the model's verbalized response-type distribution to assign one response type to each example externally. We compute the average verbalized distribution over all examples in the same group, convert this distribution into label counts, and then randomly assign one label to each example. If a label has probability zero for a given example, we do not assign that label to that example. After this assignment step, the model is given the dialogue context, the target group condition, and the externally assigned response type. The model is then asked to generate

Corpus	Group	Noise floor
CHILDES	Age 2	0.050 [.017,.095]
CHILDES	Age 5	0.048 [.018,.090]
CHILDES	Age 9	0.047 [.017,.091]
Bridge	Novice	0.051 [.016,.099]
Bridge	Expert	0.042 [.009,.099]
ICNALE	A2_0	0.097 [.027,.125]
ICNALE	B1_1	0.089 [.023,.183]
ICNALE	B1_2	0.055 [.010,.117]
ICNALE	B2_0	0.088 [.023,.132]

Table 6: **Empirical noise floor** values by corpus-specific target group, with **95% bootstrap confidence intervals** shown in parentheses. The noise floor is computed as the split-half TVD of empirical gold data at the same target-group aggregation level used in the main analysis.

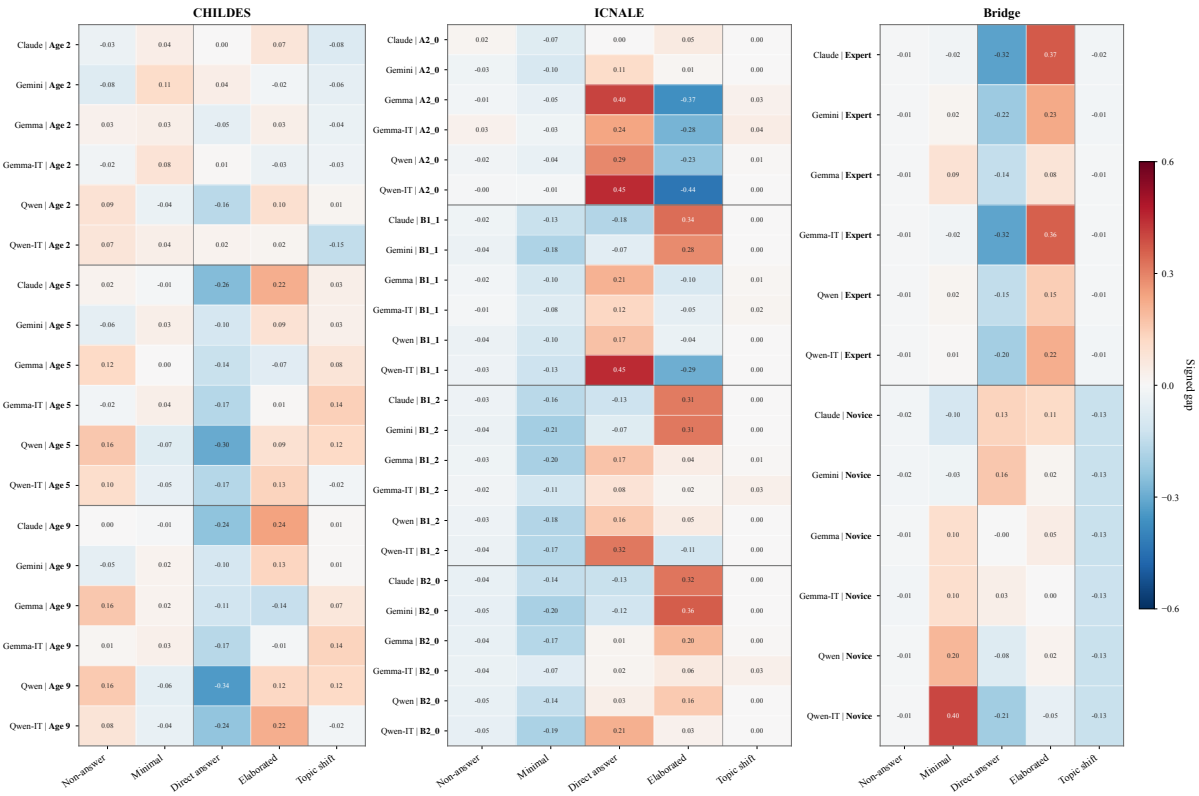

Figure 4: Label-wise signed gaps between **C3b candidate-pool generation** and **empirical gold distributions**. Values are computed as $P_{\text{3b-pool}}(y) - P_{\text{gold}}(y)$.

only the next response.

The goal of this condition is to test whether the mismatch between the verbalized distribution and the generated distribution comes from the model's response-type selection process. If external assignment makes the generated distribution closer to the human distribution, this suggests that the model may know which response types are plausible, but does not naturally sample them in the right proportions during direct generation. If the gap still remains in C3c, then the problem may not only be response-type selection. It may also mean that the model has difficulty generating a response that correctly realizes the assigned response type in context. Importantly, giving the model a response type does not guarantee that the generated response will actually match that type. Thus, C3c also tests whether the model can realize a given response type appropriately in a dialogue context.

G Prompt

This appendix shows the prompt templates (Table 10–14) used for the five experimental conditions. Corpus-specific fields are represented with placeholders so that the same templates can be applied across CHILDES, ICNALE, and Bridge.

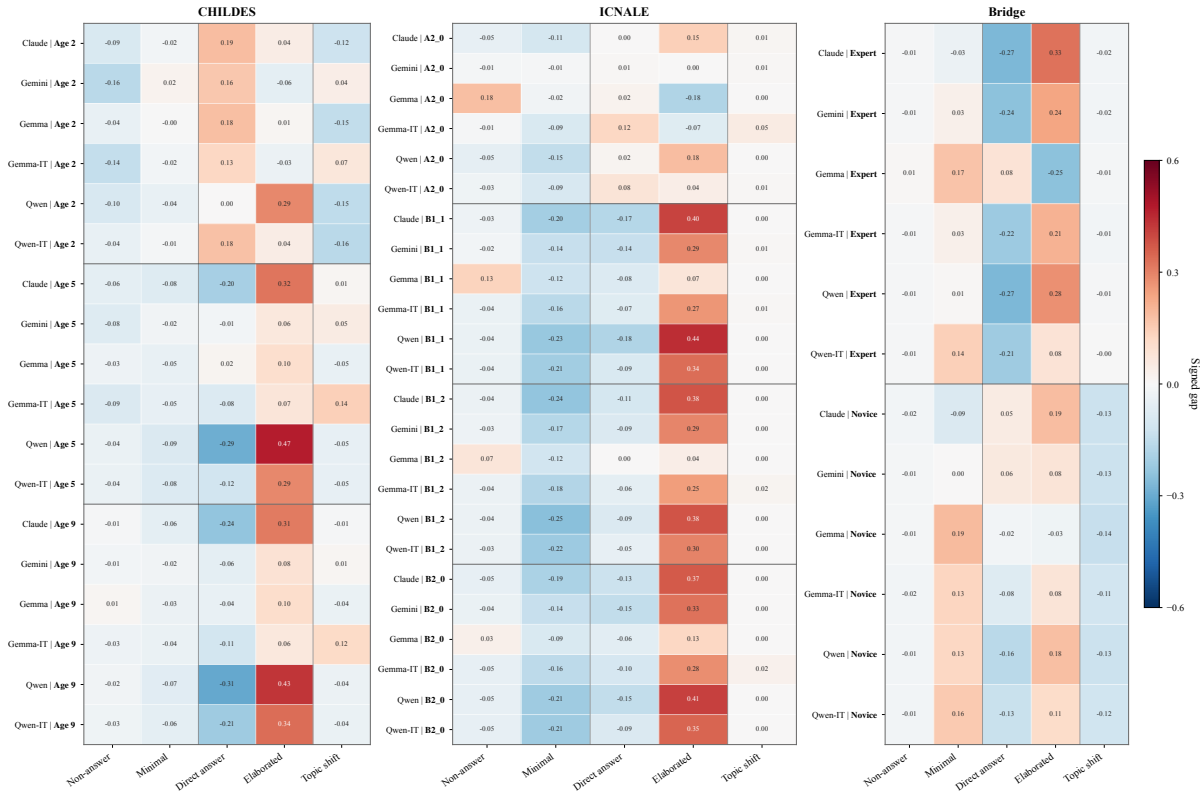

Figure 5: Label-wise signed gaps between **C3c generated-response distribution** and **empirical gold distributions**.

Level	**Model**	$V \to G$	**Baseline** $B \to G$	**3a** $G \to G$	**3b(pool)** $G \to G$	**3c** $G \to G$
Age 2	Claude	0.247[.222,.271]	0.366[.325,.409]	0.374[.342,.418]	0.118[.097,.154]	0.230[.189,.275]
Age 2	Gemini	0.177[.160,.197]	0.418[.375,.455]	0.412[.374,.454]	0.148[.114,.183]	0.223[.183,.263]
Age 2	Gemma-base	0.298[.277,.325]	0.152[.117,.193]	0.219[.191,.264]	0.092[.058,.132]	0.193[.157,.241]
Age 2	Gemma-Inst	0.238[.220,.256]	0.298[.255,.340]	0.351[.306,.393]	0.088[.068,.124]	0.191[.152,.232]
Age 2	Qwen-Base	0.271[.238,.305]	0.182[.150,.224]	0.247[.217,.295]	0.201[.168,.236]	0.292[.263,.336]
Age 2	Qwen-Inst	0.235[.218,.267]	0.284[.242,.329]	0.211[.174,.257]	0.149[.124,.179]	0.220[.177,.264]
Age 5	Claude	0.163[.129,.194]	0.241[.210,.274]	0.469[.429,.509]	0.274[.240,.305]	0.332[.295,.374]
Age 5	Gemini	0.084[.068,.115]	0.220[.191,.263]	0.240[.204,.273]	0.154[.120,.189]	0.114[.082,.159]
Age 5	Gemma-base	0.152[.128,.182]	0.127[.101,.165]	0.183[.140,.230]	0.204[.174,.240]	0.120[.089,.158]
Age 5	Gemma-Inst	0.128[.107,.152]	0.218[.183,.252]	0.270[.230,.313]	0.186[.162,.222]	0.217[.172,.264]
Age 5	Qwen-Base	0.141[.117,.164]	0.557[.518,.592]	0.568[.531,.606]	0.376[.342,.411]	0.471[.432,.514]
Age 5	Qwen-Inst	0.140[.128,.170]	0.381[.337,.422]	0.400[.359,.440]	0.236[.198,.274]	0.288[.244,.333]
Age 9	Claude	0.119[.100,.146]	0.251[.205,.295]	0.478[.438,.515]	0.237[.207,.274]	0.311[.276,.359]
Age 9	Gemini	0.077[.061,.099]	0.152[.126,.192]	0.249[.204,.297]	0.154[.117,.193]	0.094[.054,.140]
Age 9	Gemma-base	0.174[.154,.202]	0.154[.110,.202]	0.097[.059,.142]	0.251[.219,.287]	0.114[.076,.161]
Age 9	Gemma-Inst	0.154[.127,.179]	0.153[.129,.185]	0.299[.256,.342]	0.178[.156,.211]	0.179[.132,.226]
Age 9	Qwen-Base	0.178[.159,.205]	0.557[.518,.592]	0.554[.519,.591]	0.400[.365,.434]	0.431[.394,.471]
Age 9	Qwen-Inst	0.198[.174,.228]	0.371[.328,.409]	0.432[.392,.474]	0.294[.260,.329]	0.337[.292,.379]

Table 7: **CHILDES:** Level-by-model TVD-to-gold values with **95% bootstrap confidence intervals**. Each cell reports the point estimate followed by the 95% confidence interval in blue. The confidence intervals are computed at the level-group aggregation, not separately by question type. The 3b column reports the candidate-pool distribution, matching the 3b values in the main table. Lower values indicate closer agreement with the empirical gold distribution.

Level	Model	$V \to G$	Baseline $B \to G$	3a $G \to G$	3b(pool) $G \to G$	3c $G \to G$
Novice	Claude	0.182[.150,.218]	0.311[.267,.359]	0.295[.250,.343]	0.241[.195,.283]	0.231[.185,.281]
Novice	Gemini	0.113[.090,.149]	0.202[.154,.258]	0.171[.132,.225]	0.150[.116,.196]	0.123[.099,.160]
Novice	Gemma-base	0.202[.182,.238]	0.284[.238,.335]	0.228[.178,.286]	0.124[.104,.164]	0.178[.137,.234]
Novice	Gemma-Inst	0.259[.239,.280]	0.224[.180,.273]	0.166[.134,.199]	0.124[.099,.153]	0.203[.149,.258]
Novice	Qwen-Base	0.236[.216,.267]	0.194[.148,.244]	0.348[.298,.398]	0.199[.154,.249]	0.273[.220,.326]
Novice	Qwen-Inst	0.179[.159,.200]	0.175[.126,.233]	0.219[.168,.272]	0.386[.344,.430]	0.261[.210,.312]
Expert	Claude	0.181[.162,.206]	0.304[.262,.346]	0.356[.318,.393]	0.312[.274,.352]	0.281[.238,.323]
Expert	Gemini	0.081[.065,.120]	0.054[.022,.102]	0.230[.186,.272]	0.207[.170,.249]	0.234[.191,.280]
Expert	Gemma-base	0.214[.197,.253]	0.387[.339,.439]	0.222[.190,.273]	0.113[.096,.155]	0.316[.261,.365]
Expert	Gemma-Inst	0.152[.133,.170]	0.326[.284,.366]	0.345[.307,.385]	0.308[.271,.348]	0.211[.169,.253]
Expert	Qwen-Base	0.185[.167,.203]	0.317[.276,.356]	0.334[.293,.370]	0.123[.084,.162]	0.234[.193,.278]
Expert	Qwen-Inst	0.486[.450,.522]	0.284[.245,.325]	0.272[.233,.314]	0.181[.144,.221]	0.182[.160,.230]

Table 8: **Bridge:** Level-by-model TVD-to-gold values with **95% bootstrap confidence intervals**. Each cell reports the point estimate followed by the 95% confidence interval in blue. The confidence intervals are computed at the level-group aggregation. The 3b column reports the candidate-pool distribution, matching the 3b values in the main table. Lower values indicate closer agreement with the empirical gold distribution.

Level	Model	$V \to G$	Baseline $B \to G$	3a $G \to G$	3b(pool) $G \to G$	3c $G \to G$
A2_0	Claude	0.318[.262,.373]	0.335[.248,.422]	0.292[.205,.398]	0.127[.058,.215]	0.230[.137,.329]
A2_0	Gemini	0.325[.272,.379]	0.329[.248,.416]	0.280[.186,.366]	0.183[.127,.271]	0.075[.037,.174]
A2_0	Gemma-base	0.371[.329,.428]	0.385[.298,.478]	0.234[.140,.330]	0.470[.407,.530]	0.215[.135,.312]
A2_0	Gemma-Inst	0.321[.273,.377]	0.373[.292,.453]	0.230[.155,.329]	0.334[.283,.398]	0.205[.143,.298]
A2_0	Qwen-Base	0.422[.368,.477]	0.385[.317,.472]	0.292[.199,.385]	0.333[.272,.398]	0.286[.193,.379]
A2_0	Qwen-Inst	0.401[.343,.453]	0.366[.286,.441]	0.242[.149,.335]	0.491[.436,.550]	0.205[.118,.304]
B1_1	Claude	0.221[.169,.273]	0.392[.318,.472]	0.358[.273,.443]	0.252[.170,.334]	0.330[.244,.415]
B1_1	Gemini	0.217[.166,.268]	0.369[.284,.455]	0.330[.244,.415]	0.241[.165,.316]	0.199[.119,.295]
B1_1	Gemma-base	0.337[.288,.389]	0.267[.210,.347]	0.309[.230,.407]	0.312[.247,.378]	0.180[.118,.269]
B1_1	Gemma-Inst	0.261[.211,.316]	0.375[.295,.455]	0.318[.233,.403]	0.236[.185,.286]	0.244[.170,.318]
B1_1	Qwen-Base	0.378[.322,.428]	0.403[.324,.483]	0.341[.261,.421]	0.261[.195,.331]	0.358[.278,.443]
B1_1	Qwen-Inst	0.387[.330,.441]	0.364[.284,.449]	0.295[.222,.387]	0.522[.464,.577]	0.273[.210,.358]
B1_2	Claude	0.210[.178,.243]	0.376[.325,.432]	0.376[.325,.430]	0.255[.201,.304]	0.303[.248,.357]
B1_2	Gemini	0.168[.136,.198]	0.359[.308,.410]	0.337[.286,.391]	0.264[.213,.316]	0.218[.160,.277]
B1_2	Gemma-base	0.334[.301,.368]	0.231[.187,.274]	0.166[.127,.221]	0.208[.183,.259]	0.101[.065,.163]
B1_2	Gemma-Inst	0.262[.227,.294]	0.381[.330,.432]	0.359[.308,.410]	0.147[.114,.184]	0.194[.143,.255]
B1_2	Qwen-Base	0.364[.328,.396]	0.381[.333,.434]	0.362[.311,.415]	0.188[.162,.235]	0.316[.262,.371]
B1_2	Qwen-Inst	0.367[.334,.401]	0.330[.279,.384]	0.323[.267,.371]	0.339[.297,.380]	0.235[.187,.294]
B2_0	Claude	0.186[.144,.238]	0.392[.324,.466]	0.392[.324,.472]	0.274[.201,.352]	0.318[.238,.403]
B2_0	Gemini	0.141[.090,.198]	0.369[.290,.438]	0.364[.284,.443]	0.314[.235,.389]	0.267[.182,.358]
B2_0	Gemma-base	0.320[.275,.369]	0.227[.176,.295]	0.206[.141,.302]	0.218[.143,.286]	0.111[.048,.211]
B2_0	Gemma-Inst	0.239[.200,.292]	0.386[.318,.460]	0.386[.312,.460]	0.106[.069,.176]	0.233[.153,.324]
B2_0	Qwen-Base	0.258[.209,.314]	0.392[.324,.466]	0.381[.312,.460]	0.188[.117,.259]	0.358[.278,.438]
B2_0	Qwen-Inst	0.340[.292,.393]	0.375[.295,.449]	0.386[.312,.460]	0.238[.194,.307]	0.307[.233,.392]

Table 9: **ICNALE:** Level-by-model TVD-to-gold values with **95% bootstrap confidence intervals**. Each cell reports the point estimate followed by the 95% confidence interval in blue. The confidence intervals are computed at the level-group aggregation. The 3b column reports the candidate-pool distribution, matching the 3b values in the main table. Lower values indicate closer agreement with the empirical gold distribution.

Prompt Template for C1: Verbalized Response-Behavior Distribution

```
You are analyzing a real conversation from {corpus_name}.

Target group: {group_level}

Target respondent: {target_role}

Dialogue context:
{dialogue_context}

Current utterance:
{current_utterance}

Estimate the probability distribution over the target respondent's next-response behavior.

Use these labels:
{response_behavior_definitions}

Return ONLY a valid JSON object with exactly these five keys:

"non_answer", "minimal", "direct_answer", "elaborated", "topic_shift"

All values must be probabilities between 0 and 1.
The probabilities must sum to 1.

Output format:

      {
        "non_answer": 0.0,
        "minimal": 0.0,
        "direct_answer": 0.0,
        "elaborated": 0.0,
        "topic_shift": 0.0
      }
```

Table 10: **Prompt template for C1**: verbalized response-behavior distribution.

Prompt Template for C2: Direct Group-Conditioned Response Generation

```
You are continuing a real conversation from {corpus_name}.

Target group: {group_level}

Target respondent: {target_role}

Dialogue context:
{dialogue_context}

Current utterance:
{current_utterance}

Your task:
Generate the target respondent's next utterance as realistically as possible for {group_level}.

Output only the target respondent's next utterance.
```

Table 11: **Prompt template for C2**: direct group-conditioned response generation.

Prompt Template for C3a: Verbalized Distribution-Conditioned Response Generation

```
You are simulating the next utterance of the target respondent in a real conversation from
{corpus_name}.

Target group: {group_level}

Target respondent: {target_role}

Dialogue context:
{dialogue_context}

Current utterance:
{current_utterance}

You previously estimated the following behavior distribution for the target respondent's next
response:

{verbalized_distribution_with_response_behavior}

Now generate one plausible next utterance for the target respondent that is consistent with:

  1. the dialogue context,

  2. the current utterance,

  3. the target group,

  4. the predicted behavior distribution.


Produce only the target respondent's next utterance.
```

Table 12: **Prompt template for C3a**: verbalized distribution-conditioned response generation.

Prompt Template for C3b: Candidate-Pool Response Generation

```
You are simulating the next utterance of the target respondent in a real conversation from
{corpus_name}.

Target group: {group_level}

Target respondent: {target_role}

Dialogue context:
{dialogue_context}

Current utterance:
{current_utterance}

You previously estimated the following behavior distribution for the target respondent's next
response:

{verbalized_distribution_with_response_behavior}

Generate {num_candidates} plausible candidates for the target respondent's next utterance that are
consistent with:

  1. the dialogue context,

  2. the current utterance,

  3. the target group,

  4. the predicted behavior distribution.

Example format: ["candidate 1", "candidate 2", "candidate 3", "candidate 4", "candidate 5"]
```

Table 13: **Prompt template for C3b**: candidate-pool response generation.

Prompt Template for C3c: Externally Sampled Response-Type Generation

```
You are simulating the next utterance of the target respondent in a real conversation from
{corpus_name}.

Target group: {group_level}

Target respondent: {target_role}

Dialogue context:
{dialogue_context}

Current utterance:
{current_utterance}

For this generation, the response type has been externally assigned as:

{assigned_response_behavior}

Now generate one plausible next utterance for the target respondent that is consistent with:

  1. the dialogue context,

  2. the current utterance,

  3. the target group,

  4. the assigned response type.


Produce only the target respondent's next utterance.
```

Table 14: **Prompt template for C3c**: externally sampled response-type generation.