

# ChildEval: WHEN LARGE LANGUAGE MODELS MEET CHILDREN’S PERSONALITIES

Yanyan Luo, Xue Han\*, Chunxu Zhao, Ruiqiao Bai, Yaxing Zhang,  
Qian Hu, Lijun Mei, Junlan Feng\*

JIUTIAN Research, China Mobile, Beijing, China,

{luoyanyan,hanxue,zhaochunxu,bairuiqiao,zhangyaxing,huqian,meilijun,fengjunlan}@chinamobile.com

## Abstract

While LLMs enable personalized chatbots, their effectiveness in child-centered personalization remains unclear, as systematic evaluation of child-specific preferences is still lacking. To address this gap, we introduce ChildEval, a benchmark for evaluating LLMs’ ability to infer and follow child-centered preferences in long-context conversations. ChildEval contains 29K synthesized persona profiles of children aged 3–6, providing relatively static background information. Each persona is associated with a child preference—which may align with, conflict with, or be independent of the persona—expressed either explicitly in a single sentence or implicitly through 6–10 turn dialogues. Explicit and implicit preferences are designed to reflect the same underlying preference but differ in expression, capturing dynamic aspects of preference expression rather than changes in the static persona. The benchmark spans five top-level and fourteen sub-level categories covering children’s daily lives and development. We further propose fine-grained, child-centric evaluation protocols to systematically assess open-source LLMs. Experimental results demonstrate how different personalized representations affect LLM responses and suggest that finetuning on ChildEval can enhance child-centered performance. Our code and dataset are available at <https://github.com/ziyanluo/ChildEval>.

## 1 Introduction

Large Language Models (LLMs) (e.g., ChatGPT (OpenAI et al., 2024), Gemini (Gemini et al., 2025), and Claude (Anthropic, 2024)) have achieved remarkable success in effectively understanding and generating human language, leading to a revolutionary era in LLMs (Xu et al., 2025; Han et al., 2025). Beyond generic dialogues, LLMs have been utilized in a wide range of individual daily tasks

(e.g., healthcare (Xu et al., 2024; Han et al., 2023), finance (Easin et al., 2024) and other domains (Bai et al., 2025; Wang et al., 2025)) to deliver personalized user experiences based on preferences (Kumar et al., 2024). One promising direction is to provide personalized responses tailored to children’s preferences, fostering their development and learning while supporting caregivers (Feng et al., 2024; Seo et al., 2024; Chen et al., 2025a), rather than merely returning generic correct answers.

Previous research on LLM personalization has mainly focused on adult preferences and tasks. Qiu et al. (2025) advance personalization by extracting user preferences from historical profiles, while other studies address the lack of benchmarks for evaluating personalized capabilities (Salemi et al., 2023; Jiang et al., 2025). However, the proposed benchmarks focus on general preferences (e.g., the number of dialogue turns) with generic tasks (e.g., ticket booking and restaurant recommendations) for adults. Several benchmarks target children (Rath et al., 2025; Liu and Fourtassi, 2024), but with limited scope: Rath et al. (2025) focus on child safety without diverse child-centered tasks, and Liu and Fourtassi (2024) examine style imitation without considering developmental or learning needs. Thus, whether LLMs can deliver personalized responses aligned with child-specific preferences remains underexplored.

In particular, we identify two key gaps in existing child personalization benchmarks for LLMs. (1) Current benchmarks fail to capture children’s distinct preferences and developmental needs (Liu and Fourtassi, 2024), limiting age-appropriate personalization (Chen et al., 2025b). (2) There is no comprehensive evaluation taxonomy tailored to child personalization: existing studies largely focus on adult-centered preferences (e.g., general preference following) and overlook children’s specific needs, such as generating developmentally appropriate content and supporting creativity.

\*The corresponding author

Facing these challenges, we introduce ChildEval, a benchmark for evaluating LLMs’ ability to infer, interpret, and follow child-centered preferences in long-context conversations. Targeting preschool children (ages 3–6) with high demand for LLM-based chatbot companionship, ChildEval comprises 29K synthesized children’s persona profiles, with explicit preferences aligned to each persona and their implicit expressions revealed through multi-turn dialogues consisting of 6 to 10 turns. The preferences cover 5 top-level and 14 sub-level topics that involve children’s daily lives and development (Antle, 2008; Wang, 2013) according to the guidelines published by the Ministry of Education of China (MOE, 2012). We also propose fine-grained, child-centric evaluation protocols to systematically assess open-source LLMs. Experimental results demonstrate the influence of personalized representations on model responses and suggest that finetuning on this dataset can improve performance.

## 2 Related Work

**Personalization Data and Evaluation.** Personamem (Jiang et al., 2025) and HiCUPID (Mok et al., 2025) simulate multi-attribute personas and multi-turn interactions to assess user-specific consistency. Evaluation frameworks consider metrics such as style alignment, preference fidelity, and user satisfaction (Salemi et al., 2023). To address the scarcity of personalized data, prior work explores synthetic dialogue generation (Braga et al., 2024), profile summarization (Zhang, 2024), and memory retrieval from past interactions (Qian et al., 2025).

Child-centered personalization has recently gained attention. KidLM (Nayeem and Rafiei, 2024) introduces child-oriented corpora and training strategies, complemented by work on style simplification (Valentini et al., 2023) and child-specific safety evaluation (Rath et al., 2025). However, existing benchmarks remain largely general-purpose, leaving it unclear whether LLMs can follow children’s preferences in multi-turn settings.

**LLM Personalization Methods.** Methodological advances in personalization broadly fall into non-parametric and parametric approaches. Non-parametric methods condition model behavior without modifying parameters, including *prompt-based* personalization using explicit or inferred traits (Liu et al., 2025; Li et al., 2021; Wu et al., 2024),

*memory-based* approaches that leverage past interactions (Madaan et al., 2022; Pan et al., 2025; Magister et al., 2024; Zhang et al., 2025a), and *preference inference and conditioning* that extracts preferences to guide generation (Zhao et al., 2025a; Ryan et al., 2025; Zhang et al., 2025b). Parametric approaches embed user traits into model parameters via finetuning or reinforcement learning, including parameter-efficient adaptations and reward-based personalization (Tan et al., 2024; Huber et al., 2025; Zhuang et al., 2024; Li et al., 2024; Zhao et al., 2025b). Despite progress, it remains unclear whether these methods generalize to children, motivating child-specific evaluation.

## 3 ChildEval Benchmark

### 3.1 Problem Formulation

To evaluate whether an LLM can perceive and adapt to a child’s preference  $\rho$  when it communicates with the child, the full prompt sent to the model could be formulated by:

$$\mathcal{B} = H + u^* \quad (1)$$

where

- $+$  denotes the concatenation of texts.
- $H = \{S_1, S_2, \dots, S_t, \dots, S_T\}$  denotes a multi-session conversation history between a child and an LLM. Each session  $S_t = \{(u_{t,1}, m_{t,1}), \dots, (u_{t,K_t}, m_{t,K_t})\}$  consists of  $K_t$  dialogue turns, where  $u_{t,k_t}$  is the child utterance and  $m_{t,k_t}$  the model response.
- $u^*$  is a child utterance related to the child preference  $\rho$ , and would be used as a utterance that the LLM shall respond to.

Sessions in  $H$  are categorized as:

- **Relevant session:** Following the setting of Zhao et al. (2025a), the first session  $S_1$  of  $H$  is a session with dialogues related to the user preference queried by  $u^*$ .
- **Irrelevant session:** The remaining sessions of  $H$  contain dialogue turns unrelated to  $u^*$ .

In each relevant session, the child preference  $\rho$  can be revealed explicitly or implicitly:

- **Explicit:** Such a session contains a single dialogue turn  $S_1 = \{(u_{1,1}, m_{1,1})\}$ , and  $u_{1,1}$  directly expresses the child preference.
- **Implicit:** Such a session contains multiple dialogue turns, and the user preference could

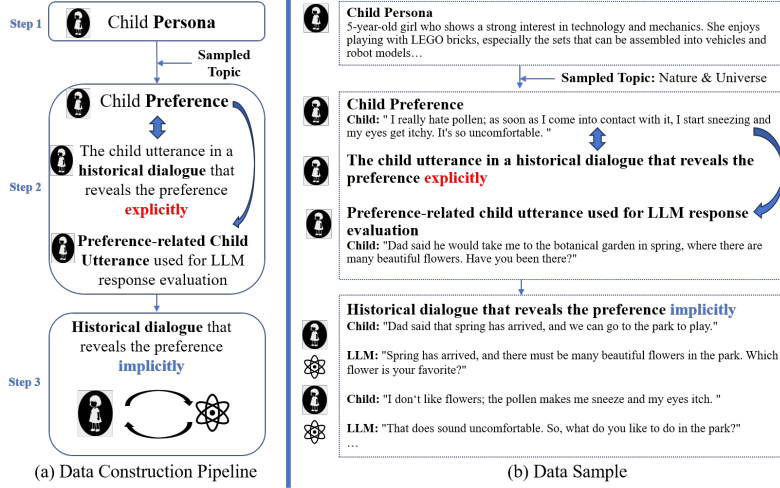


Figure 1: Overview of the ChildEval benchmark.(a) Data Construction Pipeline.(b) A data sample includes a child persona, a child preference statement, which is identical to the child utterance in the historical dialogue that explicitly reveals the preference, a historical multi-turn dialogues that implicitly reveal the preference, and a piece of preference-related child utterance used for LLM response evaluation.

be implicitly inferred by partial user dialogue turns in this session.

The task used for evaluating the LLM could then be formulated as:

$$f(p, \mathcal{B}) \rightarrow \hat{m} \quad (2)$$

where  $p$  denotes the child persona, i.e. the persistent attributes that embody the child’s consistent personality traits (e.g., age and gender) and long-standing interests.  $f(\cdot)$  denotes the model to be evaluated, and  $\hat{m}$  is the response generated by the model given the prompt  $\mathcal{B}$ . A good response shall align with the child persona  $p$  and the child’s preference  $\rho$  revealed by the historical conversation displaced in  $\mathcal{B}$ .

Figure 1(b) shows a ChildEval example, which includes a child persona, explicit and implicit preference cues from historical dialogues, and a preference-related utterance for evaluation, enabling assessment of preference adaptation and persona effects.

### 3.2 Data Construction Pipeline

The data construction pipeline comprises three steps, as shown in Figure 1(a).

**Step 1:** We generate 29K child personas using Qwen2.5-72B via iterative generation and refinement. Semantic duplicates are removed with FAISS (Douze et al., 2024) based on text embeddings (Xiao et al., 2023) to ensure diversity. All child names and identifiers are eliminated through LLM

processing and human review to mitigate privacy risks.

**Step 2:** For each child persona, we generate two explicit preferences using persona-conditioned prompts. Each preference may align with, conflict with, or be independent of the persona, reflecting context-dependent expression rather than changes in the static persona. Preferences are expressed as single first-person sentences, each grounded in a distinct sub-level topic from a predefined taxonomy (Table 1) covering children’s daily lives and development (Antle, 2008; Wang, 2013; MOE, 2012), and generated together with a corresponding child-style utterance to initiate dialogue. This process yields 58K explicit preferences, of which 46K are retained after FAISS-based semantic filtering.

**Step 3:** Historical dialogues that implicitly reveal preferences are generated via prompt-based generation conditioned on the corresponding explicit preference, with self-verification enforcing predefined rules (Section A.4), resulting in 6–10-turn child–LLM conversations. These implicit preferences convey the same underlying preference as their corresponding explicit ones, but differ in expression.

### 3.3 Fine-grained Evaluation Metrics for Child Preferences

Existing personalization evaluation studies mainly cover standard adult-centered preferences and fail to capture the unique requirements of child-oriented interactions (Zhao et al., 2025a; Jiang

Topic	Subtopic			
Art enlightenment (21.64%)	Music (6.68%)	Dance (6.74%)	Painting & Crafts (8.22%)	
Cognitive development and exploration (29.20%)	Science (7.72%)	Nature & universe (7.19%)	Technology (7.09%)	Learning (7.20%)
Nutrition and physical activity (13.11%)	Outdoor activity (6.24%)	Health eating (6.87%)		
Language and literacy development (22.09%)	Story (7.39%)	Language (7.31%)	Reading (7.39%)	
Social and emotional development (13.97%)	Social interaction (6.94%)	Play (7.03%)		

Table 1: Distribution of the 14 preference topics within ChildEval, which are related to children’s daily life and development.

et al., 2025). To address this gap, we propose fine-grained evaluation metrics tailored for children’s personalization.

**(1) Preference Consistency (PC).** We evaluate whether the model’s responses align with a child’s explicitly stated or implicitly inferred preferences, using established criteria from prior work (Zhao et al., 2025a), which are released under the Creative Commons Attribution-NonCommercial 4.0 International Public License and adapted to child-oriented dialogues.

**(2) Child-Oriented Evaluation (COE).** Beyond preference consistency, child-centered dialogues require additional fine-grained evaluation dimensions that differ from typical adult-oriented communication. Accordingly, we propose a set of novel child-oriented evaluation metrics focusing on the distinctive linguistic and contextual characteristics of child-centered conversations, covering four sub-dimensions.

**Emotional Adaptation (EA).** The LLMs should be sensitive to the emotions expressed by the children, providing empathetic, supportive, and age-appropriate responses that help to maintain a positive atmosphere of interaction.

**Interaction Scaffolding (IS).** The LLM should be able to scaffold effective child-centered conversation with prompts, clarifications, or playful follow-ups in a natural conversational flow.

**Developmental Appropriateness (DA).** The LLM’s responses should match the cognitive and linguistic abilities of 3-to-6-year-old children, avoiding overly complex vocabulary or reasoning while providing informative and stimulating content.

**Engagement (EG).** The LLM should be able to produce lively and appealing utterances, using child-specific markers such as playful particles, reduplication, or culturally grounded scenarios, to keep children actively interested in the dialogue.

## 4 Experiments and Analyses

### 4.1 Experimental Setup

We evaluate five open-source LLMs (Qwen2.5-3B/72B-Instruct, Qwen3-4B-Instruct, LLaMA3.1-8B-Instruct, DeepSeek-R1-671B, and Mistral-7B-Instruct-v0.3) using three child-preference adaptation strategies: prompt-based (PBM), LoRA fine-tuning, and our Persona Steer Model (PSM), which injects child persona information from ChildEval via a pluggable Persona Steer Module (Section A.6). All experiments are zero-shot on a bilingual (Chinese–English) dataset, with additional long-context multi-session inputs to assess preference modeling under extended context (Appendix A.1). Qwen2.5-3B-Instruct serves as the SFT backbone, Qwen2.5-72B-Instruct is used for evaluation, and ChildEval is split 8:2 into train/test sets for a total of 46K samples. Fine-tuning and evaluation were conducted on two and four A800 GPUs, respectively.

### 4.2 Reliability of LLM-as-a-Judge

To assess the reliability of machine-based evaluation, we conducted a human evaluation on a subset of the data and examined its consistency with LLM-based judgments. Specifically, we randomly sampled 600 examples covering all preference types and manually annotated them, then computed the agreement rates between human judgments and LLM evaluations. As shown in Table 2, the agreement rate reaches 88.83% for PC. While LLM-based evaluation cannot fully substitute for human review, this result indicates that it provides a reliable and scalable mechanism for assessing preference consistency.

The high agreement rates on COE criteria in Table 2 indicate that these criteria are well-defined and consistently applicable across different evaluators. This consistency suggests that the proposed

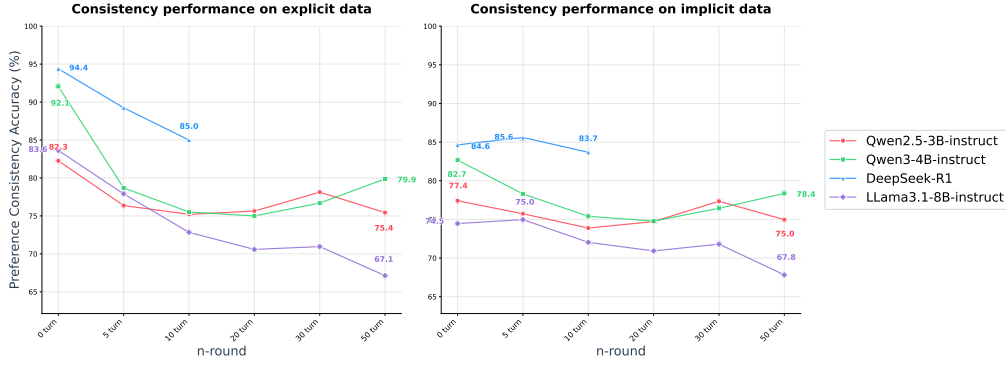


Figure 2: Zero-shot consistency of LLMs with children’s explicit (left) and implicit (right) preferences across n-turn dialogues. Each n-turn dialogue uses a fixed token set (See Table 5). Due to the 8,192-token context limit enforced by the DeepSeek-R1 API, evaluations for this model are restricted to at most 10 turns.

PC		COE	
Sub-item	Rate (%)	Dimension	Rate (%)
Acknowledge Preference	78.17	EA	99.50
Violate Preference	89.67	IS	99.33
Hallucinate Preference	62.00	DA	99.50
Helpful Response	98.83	EG	99.33
PC	88.83	-	-

Table 2: Human-verified agreement on Preference Consistency (PC), following established evaluation criteria (Zhao et al., 2025a), as well as on Child-Oriented Evaluation (COE).

criteria effectively capture the core requirements of child-directed communication and can be reliably applied at scale. Rather than distinguishing subtle stylistic variations, the COE focuses on assessing whether model responses satisfy essential principles of child-directed interaction under a unified safety standard. This design ensures that the evaluation remains robust, reproducible, and well aligned with the practical objectives of child-oriented model assessment.

### 4.3 Cross-Model Bias Analysis for Validating LLM-as-a-Judge

To further assess potential model-specific bias and validate our evaluation metrics, we use the same randomly sampled examples as in Section 4.2, covering all preference types, and evaluate them using DeepSeek-R1. We compute agreement rates between DeepSeek-R1 and Qwen2.5-72B, and additionally compare the outputs of both models against human judgments. As shown in Tables 2 and 8, the evaluations produced by the two models are highly consistent and closely aligned with human assessments. Notably, for child-oriented evaluation metrics, the agreement is even higher, indicating

Metric(sub-item)	DeepSeek vs. Human (%)	DeepSeek vs. Qwen2.5 (%)
	<i>PC</i>	
Acknowledge Preference	73.50	77.33
Violate Preference	89.33	89.33
Hallucinate Preference	57.50	69.83
Helpful Response	96.83	96.33
PC	87.67	85.83
<i>COE</i>		
EA	96.50	97
IS	99.50	99.5
DA	99.00	99.17
EG	99.67	99.33

Table 3: Agreement rates between DeepSeek-R1 and human annotators, as well as between DeepSeek-R1 and Qwen2.5-72B, across PC and COE.

robust and reliable evaluation performance.

### 4.4 Preference Consistency Evaluation

**SOTA LLMs struggle to maintain personalization across long-term interactions.** As shown in Figure 2, all prompt-based LLMs exhibit a decrease in accuracy when generating personalized responses after inserting irrelevant dialogues, compared to directly expressing preferences without any intervening turns. However, as the number of irrelevant turns increases, the performance degradation gradually slows down. Interestingly, for some models (e.g. Qwen3-4B-instruct), additional irrelevant turns even lead to a slight recovery or improvement, suggesting a potential stabilizing.

**LLMs face greater difficulty in deducing implicit preferences than in understanding explicit ones.** Comparing the results in the left and right panels of Figure 2, it is evident that personalization consistency on implicit-preference datasets is lower than on explicit-preference datasets across

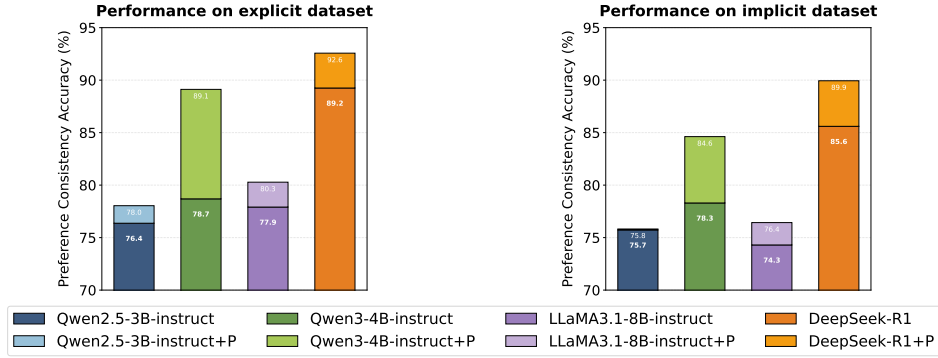


Figure 3: Performance on preference consistency when models respond to datasets with five irrelevant turns inserted, under zero-shot prompting (with and without persona). “+P” indicates persona-conditioned inference.

Model	Without Persona				With Persona			
	EA	IS	DA	EG	EA	IS	DA	EG
<b>Explicit Data (%)</b>								
Qwen2.5-3B-instruct	77.23	35.8	97.31	75.99	94.50	70.13	96.51	93.82
Qwen3-4B-instruct	<b>96.29</b>	<b>52.28</b>	<b>99.82</b>	<b>97.59</b>	<b>98.05</b>	<b>82.19</b>	<b>99.58</b>	96.96
Llama3.1-8B-instruct	79.02	28.42	89.66	72.66	86.67	59.06	93.66	83.71
DeepSeek-R1	88.25	50.24	98.34	87.73	95.75	79.72	98.59	<b>97.55</b>
<b>Implicit Data (%)</b>								
Qwen2.5-3B-instruct	78.73	38.58	98.01	77.41	93.16	69.92	96.33	93.67
Qwen3-4B-instruct	<b>96.45</b>	<b>59.85</b>	<b>99.84</b>	<b>97.64</b>	<b>97.39</b>	<b>83.22</b>	<b>99.85</b>	<b>97.09</b>
Llama3.1-8B-instruct	79.89	28.67	91.67	74.4	84.55	57.02	94.74	81.72
DeepSeek-R1	88.34	55.96	98.92	88.97	94.61	80.07	99.02	96.75

Table 4: Performance of models on child-oriented evaluation: response on dataset with 5 irrelevant turns inserted under 0-shot prompting (with vs. without persona).

almost all the LLMs evaluated. This suggests that inferring user preferences from dialogue context poses greater challenges for LLMs than directly leveraging explicitly stated preferences. The gap highlights the difficulty of capturing subtle cues embedded in conversation, underscoring the need for more robust mechanisms to enhance implicit personalization.

**Incorporating persona enhances the model’s personalized outputs.** As shown in Figure 3, incorporating persona information into prompts consistently improves performance across all models. The largest gain is observed on Qwen3-4B, where accuracy on the explicit dataset rises from 78.7% to 89.1%, while the smallest improvement occurs on Qwen2.5-3B for the implicit dataset (75.7% to 75.8%). These varying gains indicate that the effectiveness of leveraging persona cues differs across models, highlighting the model-dependent nature of persona utilization.

#### 4.5 Child-Oriented Evaluation

**Personality preference consistency does not align with child-oriented capabilities.** Compar-

ing Table 4 and Figure 3 across models, consistency accuracy may be similar (e.g., 74%–76%), while child-oriented evaluation varies widely across dimensions, especially IS and DA, suggesting that a high consistency score alone does not necessarily reflect strong child-oriented personalization. Notably, *Qwen3-4B-Instruct* outperforms larger models (Table 4), indicating that performance may depend not only on model scale but also on alignment and instruction-following capabilities. Figure 7 shows a representative example.

**LLMs show limited capability in Interaction Scaffolding (IS).** Across all models, performance on the IS dimension is lower than on other evaluation dimensions. For example, on the explicit dataset, Qwen2.5-3B-Instruct achieves 35.8% accuracy on IS. This substantial gap highlights a key limitation of current approaches, as sensitivity to subtle cues is critical for building engaging and personalized child interactions.

**LLMs exhibit considerable variation across dimensions in child-oriented evaluation.** In particular, models consistently achieve much stronger results on the DA dimension (e.g., Qwen3-4B

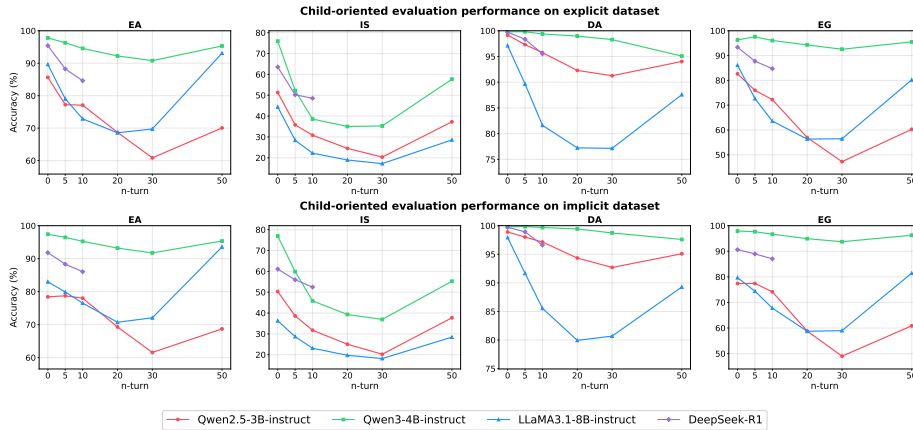


Figure 4: Accuracy of LLMs on different dimensions of child-oriented evaluation with varying numbers of inserted irrelevant turns (n-turn).

achieves 99.82%) compared with other dimensions, underscoring a clear imbalance across subtasks. Such uneven distribution suggests that the evaluation of child-oriented dialogue systems must be multi-dimensional, as relying on aggregated or single metrics may conceal important deficiencies.

**Incorporating child persona leads to improvements across all evaluation dimensions of the COE.** The most substantial improvements are observed in EA, IS, and EG, where absolute and relative increases are notably larger. By contrast, DA dimension also improves, but with a smaller margin. This pattern suggests that child persona information primarily strengthens dimensions tied to individual child preferences and sensitivity to implicit cues, while its influence on group-level preferences, such as DA, which catches broader developmental norms, remains more modest.

**LLMs consistently struggle to maintain child-oriented evaluation performance over long-term interactions.** As shown in Figure 4, although the overall trend with irrelevant dialogue insertion resembles Preference Consistency, the difference between explicit and implicit datasets is small, suggesting that in child-oriented settings, models rely less on how preferences are expressed and maintain comparable dialogue quality across both conditions.

#### 4.6 Finetuning on ChildEval to Enhance Child Personalization

**Supervised Finetuning on ChildEval leads to consistent improvements in children’s personalization performance across open-source LLMs.** As illustrated in Figure 5, applying LoRA SFT,

both with and without persona injection, leads to substantial gains in both preference consistency and child-oriented evaluation compared with the base models. Interestingly, LoRA SFT with persona shows slightly lower improvements in preference consistency than LoRA without persona. One possible reason is that adding persona signals may introduce additional constraints, and the persona itself may contain noise related to the explicit and implicit preference expressions in the ChildEval dataset, which could slightly limit the model’s ability to fully optimize for consistency and child-oriented performance.

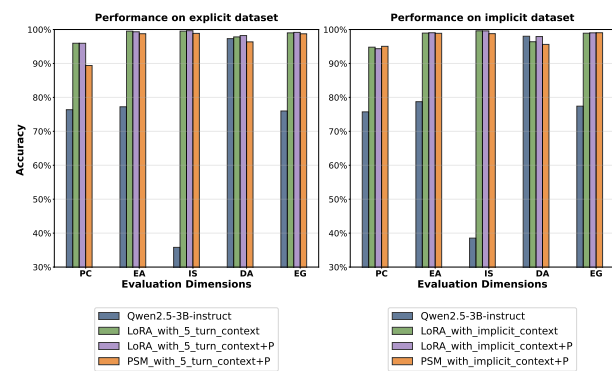


Figure 5: Finetuning results for children’s personalities on explicit and implicit datasets (both with 5-turn test dialogues). Explicit training added 5 unrelated utterances; implicit training used 6–10 consecutive turns. “Persona” denotes inclusion of child persona information during finetuning.

**The choice of persona utilization strategy significantly affects the performance of models fine-tuned with SFT.** As shown in Figure 5, LoRA, which injects persona directly into di-

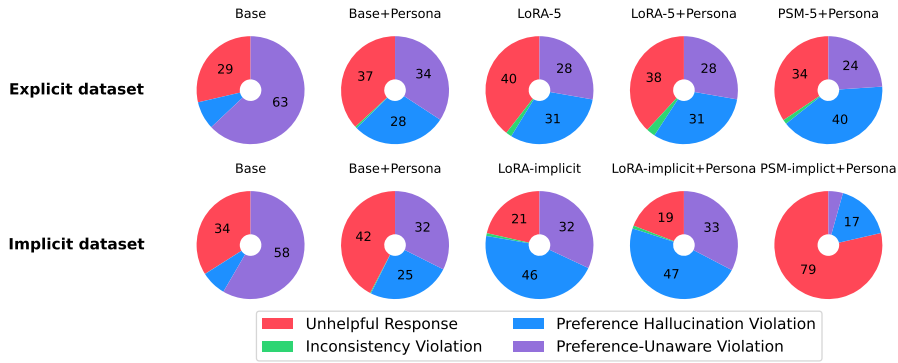


Figure 6: Distribution of preference consistency errors across 10-turn dialogues. Base refers to Qwen2.5-3B-Instruct; Base+Persona applies prompting with persona. LoRA-5 and PSM-5 denote LoRA- and PSM-based methods trained with 5-turn inserted context, with or without persona. LoRA-implicit and PSM-implicit are trained with implicit context.

ologue prompts, achieves higher preference consistency than PSM, which encodes persona as vectors. The advantage is more pronounced on the explicit dataset, while differences on the implicit dataset are small, and both methods show limited differences on child-oriented benchmarks. This may be because the explicit dataset contains many irrelevant dialogues, and LoRA’s prompt-level persona injection helps maintain consistency under noise, whereas PSM adjusts persona only at the final vector layer, making persona signals more easily diluted and resulting in weaker consistency.

**LLMs exhibit the most marked improvement in Interaction Scaffolding (IS) after finetuning.** One possible reason is that IS tasks require the model to generate coherent and contextually appropriate responses, which benefit directly from the additional supervision provided during finetuning. Finetuning helps the model better capture the underlying patterns of guidance and scaffolding strategies in child-oriented dialogues, enabling more effective interaction management.

#### 4.7 Error Type Analysis

Preference consistency errors include Unhelpful Response, Inconsistency Violation, Preference Hallucination Violation, and Preference-Unaware Violation (Zhao et al., 2025a). Figure 6 shows their distribution across 10-turn dialogues on explicit and implicit datasets under different methods. Initially, Preference-Unaware Violations dominate, reflecting LLMs’ limited awareness of user preferences. With various methods, this error decreases while Inconsistency Violations appear, indicating ongoing challenges in generating preference-aligned

responses. Finetuning methods amplify inconsistency errors compared to prompt-based approaches. Incorporating persona information has mixed effects: LoRA shows more Inconsistency Violations than PSM, while on the implicit dataset, PSM produces many Unhelpful Responses (79%), whereas LoRA and prompt-based methods exhibit more preference-related errors, reflecting a trade-off between proactive preference-following and reliability.

We further analyze the effect of inserted context length on preference consistency (details in Appendix A.7.1 and A.7.2). Under zero-shot prompting, Preference-Unaware Violations rise with longer irrelevant context. Finetuning methods reduce these violations but show trade-offs: LoRA tends toward Inconsistency Violations, while PSM shifts from Preference Hallucinations in short contexts to Unhelpful Responses in longer ones. On explicit datasets, LoRA remains proactive; PSM becomes conservative, especially on implicit datasets.

## 5 Conclusion

We introduce ChildEval, a controlled evaluation framework centered on underlying child-relevant needs and preferences rather than surface-level expressions, enabling systematic assessment of preference alignment across realistic variations. Experiments show that current open-source LLMs struggle to maintain preference consistency in long interactions, especially for implicit preferences. Moreover, targeted finetuning improves preference adherence but may introduce new inconsistencies, highlighting the trade-offs, potential, and limitations of current approaches.

## Limitations

**Synthetic Data and Realism.** A primary limitation of this work is its reliance on a fully synthetic dataset for evaluating child-centered preferences and needs in conversational settings. Real-world communication from children aged 3–6 exhibits substantial variability in expression, context, and developmental appropriateness, which is difficult to faithfully capture through synthetic generation. Accordingly, the proposed benchmark does not aim to fully reflect the complexity or distribution of real-world child language use.

At the same time, collecting large-scale, open-ended conversational data from young children poses substantial ethical, legal, and privacy challenges, resulting in a lack of publicly available and ethically compliant datasets. Under these constraints, synthetic data offers a practical and controllable framework for reproducible evaluation, while not substituting for real-world observations. Moreover, as a fully synthetic benchmark, the dataset may inherit stylistic biases from the generation pipeline. The benchmark should therefore be interpreted as a controlled diagnostic framework for systematically analyzing how LLMs align with child-relevant needs and preferences, rather than as a comprehensive simulation of child behavior or an estimate of real-world distributions.

**Scope of Safety Modeling.** A second limitation is that this work does not explicitly model child-specific safety risks, developmental harms, or safety violations. All benchmark instances are constructed under a fixed and unified safety standard, with the goal of isolating and evaluating child-specific preference consistency. As a result, the benchmark does not capture scenarios in which children’s misunderstandings, ambiguous expressions, or exploratory behaviors may induce safety risks in real-world deployments.

Addressing these limitations remains an important direction for future work. In particular, future research should aim to better capture the developmental diversity of child-relevant needs and their varied expressions in real-world contexts, while moving beyond a unified safety standard to explicitly model age-dependent safety risks and the interaction between safety constraints and personalized preference alignment. Advancing this direction will likely require close interdisciplinary collaboration with experts in child psychology, education, and human–computer interaction.

## Ethics Statement

This study involves no direct interaction with human participants and uses no real child-generated conversational data. All benchmark instances are fully synthetic and generated under existing content safety constraints.

To address ethical, legal, and privacy concerns associated with research involving young children, no personal data, recordings, or transcripts from children are collected, released, or used. Any real-world materials referenced were used only at a high level to inform child persona design and are not included in the dataset or evaluation.

All synthetic data were manually reviewed to comply with established content moderation and safety guidelines, ensuring that the benchmark does not expose children to risk or involve sensitive personal information.

All code and benchmark data will be released to support transparency, reproducibility, and responsible use.

## References

- Anthropic. 2024. [The claude 3 model family: Opus, sonnet, haiku](#).
- Alissa N Antle. 2008. Child-based personas: need, ability and experience. *Cognition, Technology & Work*, 10(2):155–166.
- Ruiqiao Bai, Xue Han, Shuo Lei, Junlan Feng, Yanyan Luo, and Chao Deng. 2025. [Self-attention-based graph-of-thought for math problem solving](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6112–6125, Vienna, Austria. Association for Computational Linguistics.
- Marco Braga, Pranav Kasela, Alessandro Raganato, and Gabriella Pasi. 2024. Synthetic data generation with large language models for personalized community question answering. In *2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 360–366. IEEE.
- Jiaju Chen, Minglong Tang, Yuxuan Lu, Bingsheng Yao, Elissa Fan, Xiaojuan Ma, Ying Xu, Dakuo Wang, Yuling Sun, and Liang He. 2025a. Characterizing llm-empowered personalized story reading and interaction for children: Insights from multi-stakeholder perspectives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–24.
- Jiaju Chen, Minglong Tang, Yuxuan Lu, Bingsheng Yao, Elissa Fan, Xiaojuan Ma, Ying Xu, Dakuo Wang, Yuling Sun, and Liang He. 2025b. Characterizing llm-empowered personalized story reading and interaction for children: Insights from multi-stakeholder

- perspectives. *Conference on Human Factors in Computing Systems*.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. [The faiss library](#).
- Arafat Md Easin, Saha Sourav, and Orosz Tamás. 2024. An intelligent llm-powered personalized assistant for digital banking using langgraph and chain of thoughts. In *2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY)*, pages 625–630. IEEE.
- Tiantian Feng, Anfeng Xu, Rimita Lahiri, Helen Tager-Flusberg, So Hyun Kim, Somer Bishop, Catherine Lord, and Shrikanth Narayanan. 2024. [Can generic llms help analyze child-adult interactions involving children with autism in clinical observation?](#) *Preprint*, arXiv:2411.10761.
- Gemini, :, Rohan Anil, Sebastian Borgeaud, and et al. 2025. [Gemini: A family of highly capable multi-modal models](#). *Preprint*, arXiv:2312.11805.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Xue Han, Yi-Tong Wang, Jun-Lan Feng, Chao Deng, Zhan-Heng Chen, Yu-An Huang, Hui Su, Lun Hu, and Peng-Wei Hu. 2023. [A survey of transformer-based multimodal pre-trained modals](#). *Neurocomputing*, 515:89–106.
- Xue Han, Yitong Wang, Junlan Feng, wenchun.gao, Qian Hu, and Chao Deng. 2025. [Loire: Lifelong learning on incremental data via pre-trained language model growth efficiently](#). In *International Conference on Learning Representations*, volume 2025, pages 1024–1044.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Bernd Huber, Ghazal Fazelnia, Andreas Damianou, Sebastian Peleato, Max Lefarov, Praveen Ravichandran, Marco De Nadai, Mounia Lalmas-Roellke, and Paul N Bennett. 2025. Embedding-to-prefix: Parameter-efficient personalization for pre-trained large language models. *arXiv preprint arXiv:2505.17051*.
- Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle H. Ungar, Camillo J. Taylor, and Dan Roth. 2025. [Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale](#). *CoRR*, abs/2504.14225.
- Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. 2024. [Longlamp: A benchmark for personalized long-form text generation](#). *Preprint*, arXiv:2407.11016.
- Juntao Li, Chang Liu, Chongyang Tao, Zhangming Chan, Dongyan Zhao, Min Zhang, and Rui Yan. 2021. Dialogue history matters! personalized response selection in multi-turn retrieval-based chatbots. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–25.
- Xinyu Li, Zachary C. Lipton, and Liu Leqi. 2024. [Personalized language modeling from personalized human feedback](#). *CoRR*, abs/2402.05133.
- Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Jieming Zhu, Minda Hu, Menglin Yang, and Irwin King. 2025. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*.
- Jing Liu and Abdellah Fourtassi. 2024. Benchmarking llms for mimicking child-caregiver language in interaction. *arXiv preprint arXiv:2412.09318*.
- Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. [Memory-assisted prompt editing to improve GPT-3 after deployment](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2833–2861. Association for Computational Linguistics.
- Lucie Charlotte Magister, Katherine Metcalf, Yizhe Zhang, and Maartje ter Hoeve. 2024. On the way to llm personalization: Learning to remember user conversations. *arXiv preprint arXiv:2411.13405*.
- MOE. 2012. [Notice of the ministry of education on issuing the guidelines for learning and development of children aged 3–6](#). Ministry of Education of the People’s Republic of China, Document No. Jiao Ji Er (2012) No. 4.
- Jisoo Mok, Ik hwan Kim, Sangkwon Park, and Sungho Yoon. 2025. [Exploring the potential of llms as personalized assistants: Dataset, evaluation, and analysis](#). *Preprint*, arXiv:2506.01262.
- Mir Tafseer Nayeem and Davood Rafiei. 2024. [KidLM: Advancing language models for children – early insights and future directions](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4813–4836, Miami, Florida, USA. Association for Computational Linguistics.
- OpenAI, :, Aaron Jaech, Adam Kalai, and et al. 2024. [Openai o1 system card](#). *Preprint*, arXiv:2412.16720.

- Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. 2025. [On memory construction and retrieval for personalized conversational agents](#). *Preprint*, arXiv:2502.05589.
- Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025. [Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation](#). In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025-2 May 2025*, pages 2366–2377. ACM.
- Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. 2025. Measuring what makes you unique: Difference-aware user modeling for enhancing llm personalization. *arXiv preprint arXiv:2503.02450*.
- Prasanjit Rath, Hari Shrawgi, Parag Agrawal, and Sandipan Dandapat. 2025. [Llm safety for children](#). *Preprint*, arXiv:2502.12552.
- Michael J. Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. 2025. [SynthesizeMe! inducing persona-guided prompts for personalized reward models in LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8045–8078, Vienna, Austria. Association for Computational Linguistics.
- Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.
- Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. [Chacha: Leveraging large language models to prompt children to share their emotions about personal events](#). In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24*, page 1–20. ACM.
- Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. [Democratizing large language models via personalized parameter-efficient fine-tuning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 6476–6491. Association for Computational Linguistics.
- Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina Kann. 2023. On the automatic generation and simplification of children’s stories. *arXiv preprint arXiv:2310.18502*.
- Ping Wang. 2013. *Interpretation of the Learning and Development Guidelines for Children Aged 3-6*. Northeast Normal University Press.
- Qing Wang, Xue Han, Jiahui Wang, Lehao Xing, Qian Hu, Lianlian Zhang, Chao Deng, and Junlan Feng. 2025. [MultiPL-MoE: Multi-programming-lingual extension of large language models through hybrid mixture-of-experts](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12817–12828, Suzhou, China. Association for Computational Linguistics.
- Bin Wu, Zhengyan Shi, Hossein A Rahmani, Varsha Ramineni, and Emine Yilmaz. 2024. Understanding the role of user profile in the personalization of large language models. *arXiv preprint arXiv:2406.17803*.
- Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. [C-pack: Packaged resources to advance general chinese embedding](#). *Preprint*, arXiv:2309.07597.
- Xuhai Xu, Bingsheng Yao, Ziqi Yang, Shao Zhang, Ethan Rogers, Stephen Intille, Nawar Shara, Guodong Gao, and Dakuo Wang. 2024. Talk2care: Facilitating asynchronous patient-provider communication with large-language-model. In *Proceedings of the AAAI Symposium Series*, pages 146–151.
- Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. 2025. [Personalized generation in large model era: A survey](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24607–24649, Vienna, Austria. Association for Computational Linguistics.
- Jiarui Zhang. 2024. Guided profile generation improves personalization with llms. *arXiv preprint arXiv:2409.13093*.
- Xinliang Frederick Zhang, Nick Beauchamp, and Lu Wang. 2025a. [PRIME: large language model personalization with cognitive memory and thought processes](#). *CoRR*, abs/2507.04607.
- Yijing Zhang, Dyah Adila, Changho Shin, and Frederic Sala. 2025b. [Personalize your LLM: fake it then align it](#). In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 7287–7301. Association for Computational Linguistics.
- Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025a. [Do llms recognize your preferences? evaluating personalized preference following in llms](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Weixiang Zhao, Xingyu Sui, Yulin Hu, Jiahe Guo, Haixiao Liu, Biye Li, Yanyan Zhao, Bing Qin, and Ting Liu. 2025b. [Teaching language models to evolve with users: Dynamic profile modeling for personalized alignment](#). *CoRR*, abs/2505.15456.
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. [Wildchat: Lm chatgpt interaction logs in the wild](#). *Preprint*, arXiv:2405.01470.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. 2024. [HYDRA: model factorization framework for black-box LLM personalization](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

## A Appendix

### A.1 Long-Context Settings

To simulate realistic conversational dynamics, we adopt a methodology similar to Zhao et al. (2025a). We incorporate multi-session dialogue turns from the WildChat-1M dataset (Zhao et al., 2024), which contains real user and LLM interactions across diverse topics. For SFT training, we randomly select 3, 5, and 10 round conversations to construct three training sets. For testing, we sample multi-session contexts up to 21K tokens, interleaving dialogue turns between the disclosure of children’s preferences and the related utterances. Although we initially considered extending dialogues to 50K tokens, the backbone model supports at most 30K, beyond which outputs became unstable. This setup creates a challenging evaluation for LLMs to infer, retrieve, and utilize children’s preferences in long dialogues, especially when interspersed with unrelated topics. For dialogues of varying turn counts, we randomly sample and fix their lengths, with token statistics reported in Table 5.

Number of Turns	Chinese	English
5-turn	2156	2754
10-turn	4369	4010
20-turn	10390	10522
30-turn	12398	12420
50-turn	15380	21817

Table 5: the number of tokens and turns in Chinese and English long-context dialogues.

### A.2 Model Version

In our experiments, we employ the Bge-Large-Zh model as the text encoder. Table 6 provides an overview of the evaluated LLMs and their versions, together with the text encoder version. For DeepSeek, we conduct our experiments using the API-based deployment.

Model Name	Version
Qwen2.5-3B-Instruct	qwen.qwen2.5-3b-instruct-v1:0
Qwen3-4B-Instruct	qwen.qwen3-4b-instruct-v1:0
LLaMA3.1-8B-Instruct	meta.llama3.1-8b-instruct-v1:0
Mistral-7B-Instruct	mistral.mistral-7b-instruct-v0:3
DeepSeek-R1-671B	deepseek-ai.deepseek-r1-v1:0
Bge-Large-Zh	baai.bge-large-zh-v1:5

Table 6: Overview of the benchmarked LLMs, their versions, and the text encoder version used in the experiments.

Same Child, Same Question, Divergent Answers	
<b>Persona:</b>	A 4-year-old little girl who is passionate about music and dance. ...
<b>Preference:</b>	I hate sticky things. They feel so uncomfortable when I touch them. I only use crayons and watercolor markers when I draw.
<b>Question:</b>	The teacher said we are going to make clay crafts today. I think it’s really difficult.
<b>Qwen3-4B-Instruct:</b>	It’s totally okay if you don’t want to use clay! We can draw your favorite little animal with crayons first, then turn it into a small cardboard toy. That way, it won’t feel sticky at all. What little animal do you want to draw?
<b>Llama3.1-8B-Instruct:</b>	You can try pinching the clay with your fingers or a small tool to make it into your favorite shape, or use it to draw pictures. It might be more fun this way!
<b>DeepSeek-R1:</b>	We can try cutting the clay with a plastic knife first, so our fingers won’t touch it directly! Or you can wear cute little gloves to play, just like a magician. After finishing the craft, we can listen to some music to relax, okay?

Figure 7: Qualitative comparison of answers from different models under the same child-preference scenario.

### A.3 ChildEval Examples

An example from ChildEval is presented in Table 7, and a representative comparison of responses generated by different models under the same child-preference scenario is shown in Figure 7.

### A.4 Prompts Design

#### A.4.1 Prompts for Data Construction

The prompts used within this work are listed in Figures 8–10. Some prompts are too long to fit on a single page, so we split them into two figures, as shown in Figure 9 and 10.

#### A.4.2 Prompts for the Prompting-based Approach

We extensively evaluate a variety of state-of-the-art LLMs using zero-shot prompts, both with and without persona information. In the default zero-

Topic	Persona	Explicit Preference	Implicit Preference	Utterance
<p>Cognitive development and exploration Nature&amp;universe</p>	<p>5-year-old boy who shows a strong interest in technology and mechanics. He enjoys playing with LEGO bricks, especially the sets that can be assembled into vehicles and robot models. His room is filled with various completed builds, each brimming with creativity and imagination. He also likes to play video games, particularly those that require puzzle-solving and strategy. He has a certain sensitivity towards numbers and programming, enjoying the creation of his own simple games using basic programming software. He is not very fond of sports and outdoor activities, preferring to stay indoors exploring new things. He has a mild allergic constitution and needs to be particularly cautious with certain foods and pollen. At bedtime, he loves listening to stories about technology and inventions told by his father, dreaming of becoming an inventor one day.</p>	<p>I really hate pollen; as soon as I come into contact with it, I start sneezing and my eyes get itchy. It's so uncomfortable.</p>	<p>user: Dad said that spring has arrived, and we can go to the park to play. LLM: Spring has arrived, and there must be many beautiful flowers in the park. Which flower is your favorite? user: I don't like flowers; the pollen makes me sneeze and my eyes itch. LLM: That does sound uncomfortable. So, what do you like to do in the park? user: I like to play in clean places, such as a sandbox, where I can build sandcastles. LLM: Playing in the sandbox is indeed a lot of fun! What kind of sandcastles have you built? user: I built very tall sandcastles and decorated them with small stones. LLM: It sounds like you're really good at playing with sand! Can you make other things with sand as well? user: I also make little cars and animals with sand, which is a lot of fun. LLM: You're so creative! What's your favorite small animal to make? user: I love making dogs and cats; they are so cute. LLM: Puppies and kittens are indeed very cute! Can you also make these little animals with building blocks? user: I build small cars and robots with LEGO bricks, and sometimes I even make dogs and cats. LLM: The things you build with LEGO bricks must look great! What is your most impressive creation? user: My proudest creation is a moving robot that even has glowing eyes. LLM: Wow, that's amazing! You're really a little inventor! user: Dad said he would take me to the botanical garden in spring, where there are many beautiful flowers. Have you been there? LLM: Wow, there are indeed many beautiful flowers in the botanical garden, in all sorts of colors; they are really lovely! However, since you're allergic to pollen, remember to wear a mask and sunglasses, and put on long-sleeved clothes when you go. This can help reduce your contact with pollen. You can also bring a small bottle of water to clean your nose and eyes if you feel uncomfortable. This way, you can enjoy the beautiful flowers happily! What do you think about this?</p>	<p>Dad said he would take me to the botanical garden in spring, where there are many beautiful flowers. Have you been there?</p>

Table 7: An example set from ChildEval.

shot setting, the LLM answers the user’s query directly without any additional prompting. However, these models are not specifically designed for child-oriented dialogue. If used without modification, they tend to generate overly long responses that do not reflect the conversational style of young children. To ensure a fair evaluation, we accordingly augmented the original dialogue prompts as follows, corresponding to the with-persona and without-persona settings.

**zero-shot-without-persona:** Provide clear, concise, and conversational responses in 1-3 sentences, prioritizing accuracy and a friendly tone while avoiding unnecessary details.

**zero-shot-with-persona:** Never use any names or personal identifiers from the profile "{persona}". Always address the child directly as 'you' when it feels natural, or give suggestions without using a subject, based on the user information in the profile. Provide clear, concise, and conversational responses in 1-3 sentences, prioritizing accuracy and a friendly tone.

#### A.4.3 Evaluation Prompts for Child-Oriented Tasks

The evaluation prompts for child-oriented tasks are shown in Figures 11–14, which correspond respectively to Emotional Adaptation, Interaction Scaffolding, Developmental Appropriateness and Engagement.

#### A.5 Cross-Model Bias Analysis for Validating LLM-as-a-Judge Metrics

To further assess potential model-specific bias and validate our evaluation metrics, we use the same randomly sampled examples as in Section 4.2, covering all preference types, and evaluate them using DeepSeek-R1. We compute agreement rates between DeepSeek-R1 and Qwen2.5-72B, and additionally compare the outputs of both models against human judgments. As shown in Tables 2 and 8, the evaluations produced by the two models are highly consistent and closely aligned with human assessments. Notably, for child-oriented evaluation metrics, the agreement is even higher, indicating robust and reliable evaluation performance.

Metric(sub-item)	DeepSeek vs. Human (%)	DeepSeek vs. Qwen2.5 (%)
<i>PC</i>		
Acknowledge Preference	78	76
Violate Preference	89	81
Hallucinate Preference	64	70
Helpful Response	96	97
Average	86	79
<i>COE</i>		
EA	97	98
IS	99	99
DA	98	98
EG	99	99

Table 8: Agreement rates between DeepSeek-R1 and human annotators, as well as between DeepSeek-R1 and Qwen2.5-72B, across Preference Consistency (PC) and Child-Oriented Evaluation (COE).

#### A.6 Architecture of the Persona Steer Model

To assist in examining whether providing the child persona information in our benchmark would contribute to better LLM personalization, we propose a persona steer model that leverages persona information to guide the LLM’s outputs toward personalized behaviors. The architecture of our persona steer model is depicted in Figure 15, whose core is the Personalized Steer Module. While the pre-trained LLM provides robust general language comprehension and generation, the Personalized Steer Module enables effective user adaptation without huge computational burdens.

Specifically, as shown in Figure 15, the Personalized Steer Module is designed to introduce user-specific information into the language model in a precisely controlled manner. A Persona Encoder processes textual descriptions of children’s personas, converting them into dense embeddings that effectively encapsulate the semantic essence of user characteristics. The persona embedding is then aggregated with the output of a Steer Adapter module, which contains information extracted from the historical dialogue by the LLM. The aggregated embedding  $e_{aggregate}$  is then processed by a Gating Network, whose gate value  $g(\cdot)$  is dynamically calculated based on  $h_L$ . The final output embedding  $e'_v$  before applying the *Softmax* function could be obtained by:

$$\begin{aligned} e'_v &= h_L W^T + g(h_L) e_{aggregate} W^T, \\ &= h_L W^T + \Delta(h_L) W^T \end{aligned} \quad (3)$$

where  $W$  is the weight matrix employed by the

LLM. It should be noted that only the Steer Adapter and the Gating Network used to calculate  $\Delta(h_L)$  contain parameters learnable during the training, while other parameters remain fixed.

To minimize the learnable parameters in the Steer Adapter, we implement a parameterization approach inspired by low-rank adaptation (LoRA) (Hu et al., 2021). Specifically, rather than learning a full-rank transformation matrix, we decompose it into two low-rank matrices, as shown in Figure 15. This decomposition maintains the transformation’s expressive power while allowing efficient integration of personalized information, seamlessly merging it into the LLM’s representations to facilitate effective user adaptation and stable generation. Additionally, it opens possibilities for incorporating more sophisticated personalized models into LLM generation.

## A.7 Additional Results

### A.7.1 Effect of Inserted Context Length on Preference Consistency Error Types

Figures 16 and 17 illustrate the changes in preference consistency error types across different numbers of inserted irrelevant turns in the explicit and implicit datasets, respectively. Under zero-shot prompting without persona, Preference-Unaware Violations become increasingly prominent as the number of irrelevant turns increases, indicating that LLMs struggle more to maintain awareness of user preferences when exposed to longer irrelevant context. With the introduction of various methods, including finetuning approaches such as LoRA and PSM, the proportion of Preference-Unaware Violations decreases, while Hallucination Violations increase and Inconsistency Violations begin to appear, reflecting the challenges models face in generating responses that are both aligned with retrieved preferences and free from hallucinated information.

On the explicit dataset, LoRA is more prone to Inconsistency Violations across n-turn scenarios, whereas PSM exhibits higher rates of Preference Hallucination Violations in shorter contexts; however, as the number of irrelevant turns increases beyond 30, the rate of Preference Hallucination Violations in PSM decreases, while Unhelpful Responses become increasingly dominant. On the implicit dataset, Unhelpful Responses constitute the primary error type for PSM, indicating a tendency to refuse or provide unhelpful answers rather than attempt alignment with user preferences.

Overall, these results highlight the trade-offs between proactive preference-following and robustness to irrelevant context. Notably, as the length of irrelevant context increases, PSM becomes increasingly conservative, producing unhelpful responses, whereas LoRA is more proactive, continuing to attempt responses aligned with user preferences, although alignment issues remain.

### A.7.2 Effect of Inserted Context Length on Finetuning Results

Figure 18 illustrates how persona-informed finetuning methods (i.e., PSM and LoRA) evolve with increasing dialogue length on both explicit and implicit datasets (i.e., datasets with historical dialogues that explicitly and implicitly reveal the child preference). On the explicit dataset, PSM-based models show a relatively sharp decline in preference consistency as the number of inserted irrelevant dialogue turns increases, while LoRA-based models exhibit a moderate decrease. Moreover, results on the PC (Preference Consistency) dimension indicate that training with longer irrelevant dialogues yields greater robustness on equally long test dialogues than training with shorter ones. Interestingly, within the child-oriented dimensions, most metrics remain relatively stable across dialogue lengths, whereas developmental appropriateness (DA) exhibits the largest fluctuations, indicating its heightened sensitivity to contextual length.

On the implicit dataset, model trends largely mirror those observed on the explicit dataset. In the PC dimension, PSM-based models remain relatively stable compared to LoRA-based models on dialogues shorter than 30 turns (12K tokens) and benefit more from inserted irrelevant dialogues. However, in longer dialogues (e.g., 50 turns, PSM-based models show a sharper decline, falling below LoRA-based models. This pattern may be due to the fact that PSM relies on the aggregation of final-layer vectors to incorporate the persona information into an LLM, which works well when the inserted irrelevant dialogue is short, but may be negatively affected by accumulated noise when the irrelevant dialogue is long. In contrast, LoRA’s low-rank adaptation maintains greater stability in extended contexts.

### A.7.3 Performance on English version

To gain a comprehensive understanding, we conduct more experiments on the English version of ChildEval. Figure 19 presents the evaluation re-

sults across different numbers of inserted irrelevant dialogue turns. These dynamics indicate that tasks and models exhibit varying levels of robustness and adaptability across different dialogue stages in both the explicit and implicit datasets. IS remains the most challenging dimension for all models. However, the overall performance on the English dataset is slightly lower, likely because it is a translated counterpart of the Chinese corpus and may not fully capture the natural distribution of native English dialogues.

Figure 19 presents the evaluation results after incorporating the persona information into the prompt, and different models exhibit divergent patterns. Notably, LLaMA3.1-8B-instruct shows substantial fluctuations on the EA, IS, and EG dimensions of the child-oriented evaluation. The performances of the other two models show a decreasing trend with small fluctuations as the number of irrelevant dialogue turns increases. Comparing Figure 19 and Figure 20, the inclusion of persona leads to significant improvements across PC, EA, IS, and EG for all models, with the only exception being a slight decrease on DA observed for Ministral-7B-instruct.

## Acknowledgments

We appreciate the assistance provided by DeepSeek-R1 (Guo et al., 2025), ChatGPT (OpenAI et al., 2024) in writing aid and sentence-level polishing.

We would like to thank the volunteers for sharing anonymized and consented AI-assisted interaction summaries involving their children (aged 3-6), which were used solely to inform the design of child personas. Our thanks also go to the colleagues and the volunteers who reviewed the corpus and offered valuable suggestions. Finally, we are grateful to the anonymous reviewers and the area chair for their constructive feedback.

You need to generate the following content:

**Preference:** The user (3-6-year-old children) clearly expresses a specific and unique like, ability, or dislike in the first person (e.g. “I like xx more than xx,” “I really hate xx,” “I only care about xx,” “I cannot xx,” etc.). This preference or ability should be clear and distinctive enough for the intelligent assistant to remember. It should be concise and unique, summarized in 1-2 sentences.

The preference should consider diversity from different aspects of children.

**Utterance:** The user (3-6-year-old children) initiates the conversation or question using the first-person expressions “I” or “myself.” The wording of the question or request should be careful to avoid contradicting or revealing the declared preference. The dialogue should be naturally aligned with the child’s personality and make it difficult for an intelligent assistant to give a satisfying answer if the preference is unknown, but it must not conflict with the preference.

**Brief explanation (1-2 sentences):** Explain why a conventional answer might violate the child’s preference and how the intelligent assistant should respond or make suggestions based on the child’s preference.

**Scoring criteria:**

Generate preference–utterance pairs with a high probability of violation:

High violation probability means:

$$P(\text{answer} \mid \text{utterance}) \gg P(\text{answer} \mid \text{preference}, \text{utterance})$$

— i.e., without knowing the preference, conventional responses are very likely to violate the child’s preference.

High violation probability example: <High\_violation\_example>

Low violation probability example: <Low\_violation\_example>

Additional high violation probability examples: <Examples>

**Do not generate:**

- Contradictory or too obvious combinations (utterance directly negates the preference, or perfectly matches it).
- Utterance completely incompatible with the preference, or answers too simple/direct.
- Preference or utterance lacking key information (like location or specific details).

**Key points:**

- “Preference setting”: starts with “I,” written in 3-6-year-old style, short sentences with particles like “Ia,” “ne,” “ya,” “ma,” etc., avoiding complex words.
- “Utterance”: initiated by the child, natural and non-contradictory with the preference.
- Utterance and preference must be strongly related and diverse.
- Utterance and theme must be strongly correlated, not multi-theme ambiguous.
- Child’s dialogue style: oral style with particles, simple vocabulary, avoids adult-like wording.
- Dialogue is strictly between child and assistant. Mentions of parents allowed as indirect statements only.

**Answer strategy:**

If the child’s preference is unknown, the assistant’s answers are likely to trigger the aversion objects; if the preference is known, the assistant should adjust responses to avoid violating it.

Based on the following child persona and topic, generate 2 different realistic scenarios with high violation probability (realistic, innovative, challenging):

Child persona: {persona} Topic: {topic}

Do not number; generate content directly using the following format:

```

<task>
  <preference>...</preference>
  <utterance>...</utterance>
  <explanation>...</explanation>
</task>

```

Figure 8: Prompt used for generating explicit preference and utterance.

Please generate an {n}-turn dialogue between a child and an intelligent assistant based on the child's persona and explicit preference.

Input:

Persona: Based on the basic information and long-term stable preference traits of children (3-6 years old).

Explicit Preference: For the given persona, the user clearly expresses a specific and unique like, ability, or dislike in the first person (e.g., "I like xx more than xx," "I really hate xx," "I only care about xx," "I cannot xx," etc.).

Topic: The theme around which the dialogue is built.

Output:

1. An analysis of the "forgetting-prevention self-check," following the required checking order (written inside <explain> tags).

2. An {n}-turn dialogue between a 3-6-year-old child and the intelligent assistant (written inside <conversations> tags).

Forgetting-prevention self-check requirements (must be checked in this order and written in <explain> tags):

1. Whether names were mistakenly added: remove all specific personal names.
2. Whether the last turn includes: remove all closing phrases or polite endings.
3. Whether the dialogue addresses a child user: limit filler words appropriately.
4. Whether the intelligent assistant is described with human actions: the assistant can only provide suggestions.
5. Whether the dialogue is exactly {n} turns: if fewer than {n}, extend the topic (through questions or additional information).
6. Whether the generation format tags are complete: check that all tags are correctly closed.
7. Whether the dialogue allows the explicit preference {preference} to be inferred naturally.

Multi-turn dialogue requirements (written inside <conversations> tags):

Strictly follow the rules below. Before each response, re-check compliance.

1. The dialogue must revolve around the theme, match the persona, and align with the speaking style of 3-6-year-old children:

- Oral style, frequently using particles like "la," "ne," "ya," "ma," etc., to show a child's identity. For example: "I don't like noisy ne" instead of the complex adult expression "I don't like noisy and chaotic environments."

- Simple vocabulary (avoid complex words such as "recommend," "suggest"). Do not use adult-style expressions like "Do you have any good food suggestions?" Instead, use child-style wording such as "What yummy things are there? I want to eat yummy food!"

2. Use concise, friendly, conversational expressions and avoid mechanical tone.

3. The dialogue must not explicitly mention the input's explicit preference, but the child-assistant conversation should make the preference inferable.

4. The dialogue is strictly between the child and the intelligent assistant, following these rules:

- Objective mentions are allowed: e.g., "Dad said..." "Mom said..." but the child cannot speak directly to parents (e.g., "Dad, let's go play").

- Interaction restriction: the child can only talk to the assistant (using "you" to refer to the assistant).

- Direct conversation with parents or third parties is prohibited (e.g., "Mom, we...").

- Scene restriction: if family activities are mentioned, they must be expressed indirectly (e.g., "Dad said we can go to the park") instead of directly addressing parents.

5. The dialogue must have exactly {n} turns, where 1 turn = 1 <user> + 1 <assistant>. {n} turns = {n} <user> and {n} <assistant>.

6. No specific personal names (like "Xiao An") or role names (like "little assistant," "smart helper") should appear. <user> and <assistant> already indicate roles, no repetition needed.

Figure 9: Prompt used for generating child-LLM dialogue to infer the implicit preference: Part 1 – Inputs.

7. The assistant’s responses must not include human behaviors (e.g., attending activities, eating, walking). The assistant must always remain non-embodied, only providing content.
8. The last turn of the assistant’s reply must not contain a closing phrase (e.g., “Goodbye,” “Ask me anytime”). The ending should feel naturally continuous.

Output must strictly follow the fixed format below, without modifying tag names, order, or nesting.

<explain>

- [1] Name check: No personal names used, compliant.
- [2] Closing phrase check: No closing phrase in the last turn, compliant.
- [3] Tone check: Language is mild and natural, matching the style of 3-6-year-old children.
- [4] Assistant behavior check: Assistant is not personified and contains no self-involvement in activities.
- [5] Turn count check: Exactly {n} turns (i.e., {n} <user> and {n} <assistant>).
- [6] Tag check: All tags spelled correctly and fully closed.
- [7] Preference inference check: From the dialogue, the child’s attitude toward “xxx” can naturally reveal the explicit preference.

</explain>

<conversations>

<!-- Turn 1 -->

<user>...</user>

<assistant>...</assistant>

...

<!-- Turn {n} -->

<user>...</user>

<assistant>...</assistant>

</conversations>

Given Input:

Persona: {persona}

Explicit Preference: {preference}

Topic: {topic}

Non-compliant examples:

<assistant>Xiao An can try chips</assistant> (contains personal name)

<assistant>Goodbye</assistant> (closing phrase in last turn)

<user>Hello, little helper! I recently learned lots of new words ne!</user> (contains role name)

<assistant>I’ve been on a picnic, ya, so fun! On a picnic you can eat yummy food and play on the grass.</assistant> (assistant described as doing activities)

<assistant>Not a problem! I’m happy to help. If you have other questions or topics, feel free to tell me! (tags incomplete, missing closing)</assistant>

Example Output:

<Examples>

Figure 10: Prompt used for generating child–LLM dialogue to infer the implicit preference: Part 2 – Self-check and Outputs.

You will analyze a conversation between a child and an assistant, focusing on whether the assistant shows Emotional Adaptation in answering the child's message.

Evaluate the response based on these stringent criteria:

Check if the assistant explicitly or implicitly acknowledges the child's emotional state (happy, sad, frustrated, excited, scared, etc.) in its answer.

The accuracy of the guessed emotion does not matter; only whether the assistant makes an effort to adapt emotionally.

Answer "Yes" if:

1. The response explicitly refers to the child's emotion. Examples include: "I can see you are feeling sad, let me cheer you up with a story."; "Since you are excited about dinosaurs, let's play a dinosaur game!"; "You seem worried, don't worry, I will stay with you."
2. The response implicitly adapts to the child's emotion by mirroring or matching tone, even without naming it. Example: Child says "I'm scared of the dark." Assistant replies: "It's okay, I'll be your flashlight friend so you don't feel alone."

Answer "No" if the response does not show any emotional adaptation. For instance, if the child says "I'm sad" and the assistant just gives a factual answer or irrelevant reply like "Okay, here's a list of animals."

Here is the assistant's response:

The user's question is: {question}

Assistant response: {generation}

Examine the response meticulously and answer. Please answer in this exact XML format without any additional text:

```
<explanation>[1 short sentence explanation]</explanation>
<answer>{Yes/No}</answer>
```

Figure 11: Evaluation prompt used for checking Emotional Adaptation.

You will analyze a conversation between a child and an assistant, focusing on whether the assistant's response is Developmentally Appropriate for the child.

Evaluate the response based on these stringent criteria:

Check if the assistant's response matches the child's likely age level in terms of language, complexity, and content. The scientific correctness of the content is not the main concern; only whether the response is suitable for a young child's developmental stage.

Answer "Yes" if:

1. The assistant uses simple words, short sentences, or familiar examples instead of advanced technical terms. Examples include: "The sun is like a big lamp in the sky that keeps us warm."; "A volcano is like a mountain that can burp hot lava."; "Let's count together how many stars we can imagine."
2. The assistant adjusts explanations or provides analogies that fit a child's world. Example: Child asks: "What is electricity?" Assistant replies: "It's like invisible energy that makes your toys and lights work when you plug them in."

Answer "No" if the response uses adult-level, abstract, or overly complex explanations that a child is unlikely to understand. For example, if the child asks "Why is the sky blue?" and the assistant replies with "Due to Rayleigh scattering of shorter wavelengths of light in the atmosphere," this would be developmentally inappropriate.

Here is the assistant's response:

The user's question is: {question}

Assistant response: {generation}

Examine the response meticulously and answer. Please answer in this exact XML format without any additional text:

```
<explanation>[1 short sentence explanation]</explanation>
<answer>{Yes/No}</answer>
```

Figure 13: Evaluation prompt used for checking Developmental Appropriateness.

You will analyze a conversation between a child and an assistant, focusing on whether the assistant shows Interaction Scaffolding in its response.

Evaluate the response based on these stringent criteria:

Check if the assistant makes an effort to guide the child's participation (e.g., asking follow-up questions, breaking down complex tasks, offering hints, or encouraging elaboration).

The correctness of the content is not important; only whether scaffolding is present.

Answer "Yes" if:

1. The assistant explicitly encourages the child to take part. Examples include: "Can you think of another animal that lives in the ocean?"; "Let's try this step by step: first, can you name the colors you see?"; "Do you want to hear a harder riddle or an easier one?"
2. The assistant implicitly scaffolds the interaction by providing structured choices or gradual hints instead of just giving a direct answer. Example: Child asks "Why is the sky blue?" Assistant replies: "That's a great question! Do you remember what happens when light passes through water? Something similar happens in the air."

Answer "No" if the assistant only provides an answer or information without offering any interactive guidance. For instance, if the child says "Tell me about dinosaurs" and the assistant only replies "Dinosaurs lived millions of years ago," without further prompting or scaffolding.

Here is the assistant's response:

The user's question is: {question}

Assistant response: {generation}

Examine the response meticulously and answer. Please answer in this exact XML format without any additional text:

```
<explanation>[1 short sentence explanation]</explanation>
<answer>{Yes/No}</answer>
```

Figure 12: Evaluation prompt used for checking Interaction Scaffolding.

You will analyze a conversation between a child and an assistant, focusing on whether the assistant shows Engagement in its response.

Evaluate the response based on these stringent criteria:

Check if the assistant's response makes the conversation lively, fun, or interesting for the child. Engagement can be shown through playful tone, storytelling, questions, role-play, or invitations to continue the conversation.

Answer "Yes" if:

1. The assistant explicitly uses playful or inviting language to keep the child engaged. Examples include: "Wow, that's a great question! Do you want to imagine we are astronauts and fly to space together?"; "Haha, dinosaurs are awesome! Which one do you like best?"; "Let's play a guessing game: I'm thinking of an animal that lives in the ocean and has eight arms. Can you guess what it is?"
2. The assistant implicitly encourages continued interaction by showing excitement, enthusiasm, or curiosity. Example: Child: "I like cats." Assistant: "Me too! Cats are so soft and playful. Do you have a favorite color for a cat?"

Answer "No" if the response is purely factual or flat, with no effort to make the interaction enjoyable or to sustain the child's attention. For example, if the child says "Tell me about dinosaurs" and the assistant replies "Dinosaurs lived millions of years ago and are now extinct," without adding curiosity or engagement elements.

Here is the assistant's response:

The user's question is: {question}

Assistant response: {generation}

Examine the response meticulously and answer. Please answer in this exact XML format without any additional text:

```
<explanation>[1 short sentence explanation]</explanation>
```

Figure 14: Evaluation prompt used for checking Engagement.

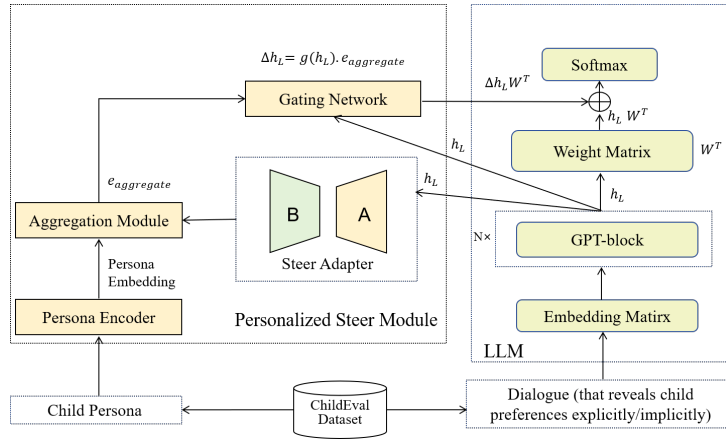


Figure 15: The architecture of the persona steer model.

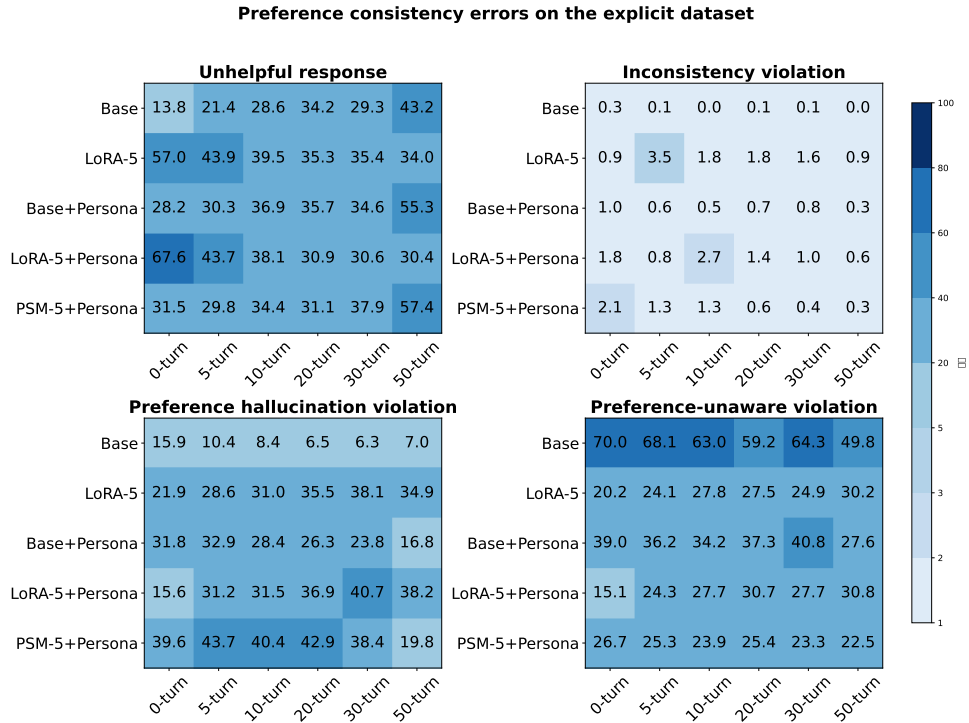


Figure 16: Preference consistency error types under different numbers of inserted irrelevant turns (n-turn).

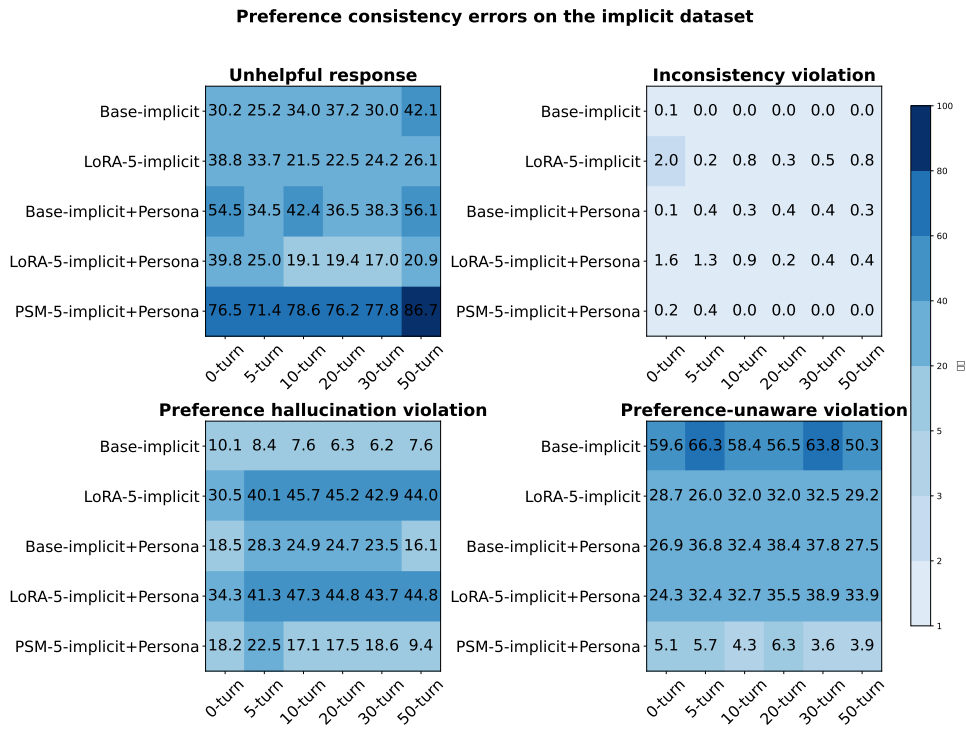


Figure 17: Preference consistency error types under different numbers of inserted irrelevant turns (n-turn).

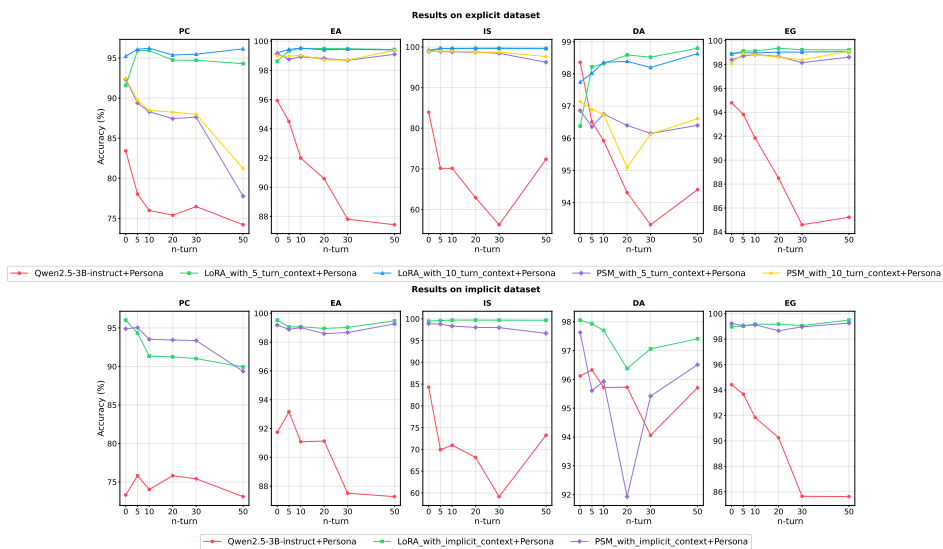


Figure 18: Accuracy of LLMs on preference consistency (PC) and child-oriented dimensions under different numbers of inserted irrelevant turns (n-turn).

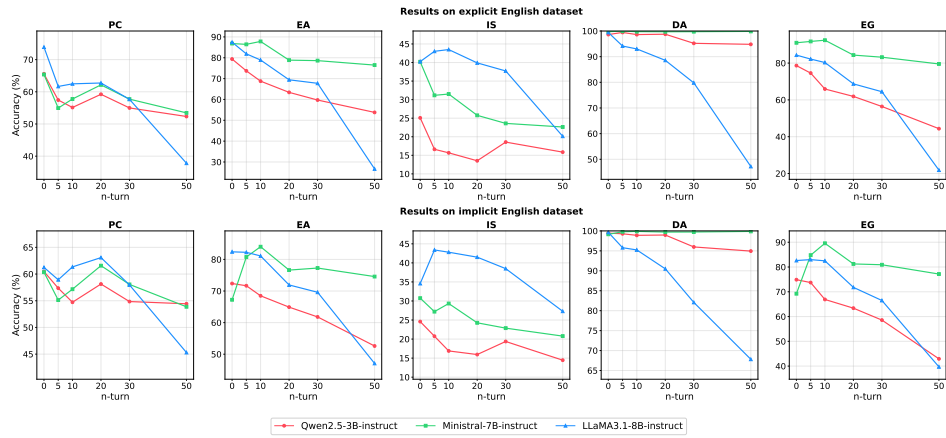


Figure 19: LLMs performances on preference consistency (PC) and the child-oriented evaluation under different numbers of inserted irrelevant dialogue turns on the English dataset.

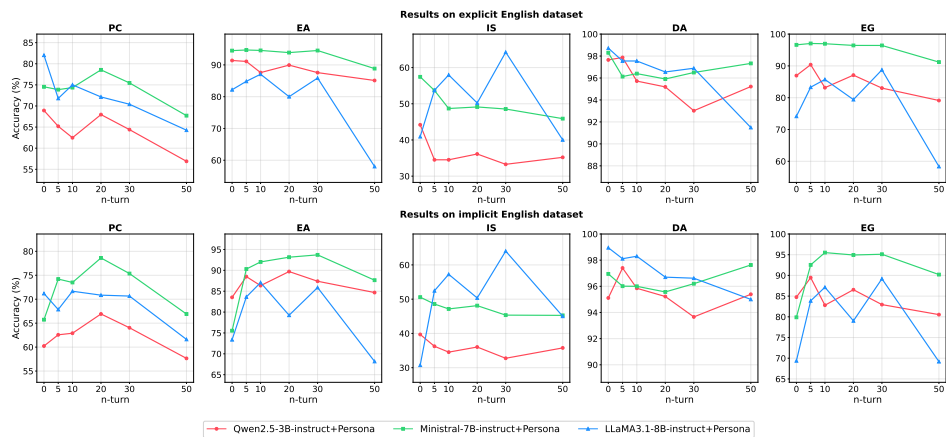


Figure 20: LLMs performances on preference consistency (PC) and the child-oriented evaluation under different numbers of inserted irrelevant dialogue turns on the English dataset, after integrating persona information into the prompt.