

ChildEval: WHEN LARGE LANGUAGE MODELS MEET CHILDREN’S PERSONALITIES

Anonymous ACL submission

Abstract

While LLMs enable personalized chatbots, their effectiveness in child-centered personalization remains unclear, as children’s interactions differ from those of adults and systematic preference evaluation is 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.¹

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. Beyond generic dialogues, LLMs have been utilized in a wide range of individual daily tasks (e.g., healthcare (Xu et al., 2024)

and finance (Easin et al., 2024)) to deliver personalized user experiences based on preferences (Kumar et al., 2024). One promising direction is to enable personalized interactions for children, convenient for their caretakers, while promoting their development and learning (Feng et al., 2024; Seo et al., 2024; Chen et al., 2025a), instead of just giving the “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 based on children’s preferences remains underexplored, motivating dedicated research given the fundamental differences between child and adult interactions.

In particular, we identify two primary gaps in existing child personalization benchmarks for LLMs. (1) Current benchmarks fail to reflect children’s interaction patterns (Liu and Fourtassi, 2024), and supporting age-appropriate developmental interaction (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.

¹We will open-source all codes and data.

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 ρ 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 ρ , 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 ρ 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

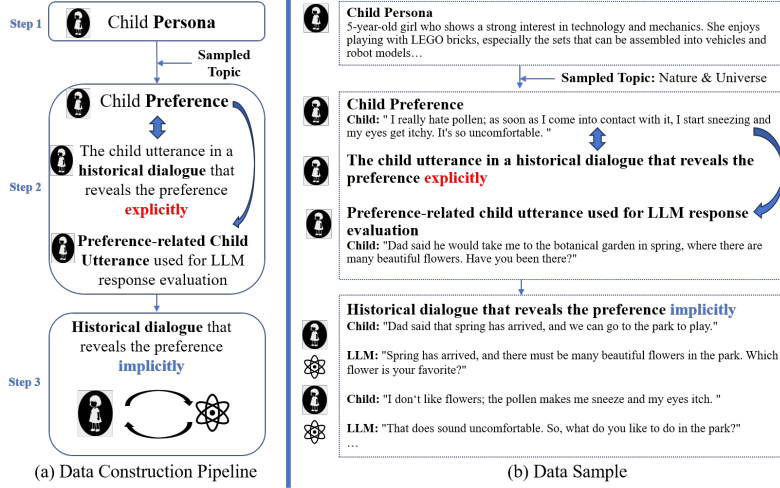


Figure 1: Overview of the ChildEval benchmark.(a) Data Construction Pipeline.(b) A data sample includes a child persona, a child preference statement, which is identical to the child utterance in the historical dialogue that explicitly reveals the preference, a historical multi-turn dialogues that implicitly 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 ρ 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 100 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 87% 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

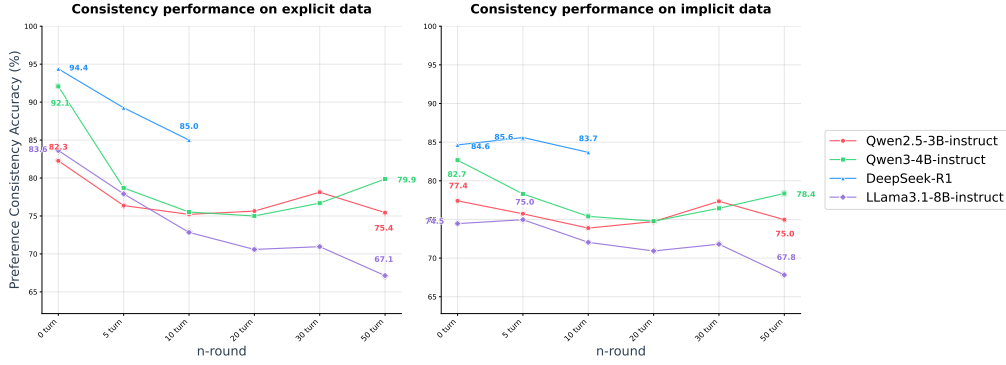


Figure 2: Zero-shot consistency of LLMs with children’s explicit (left) and implicit (right) preferences across n-turn dialogues. Each n-turn dialogue uses a fixed token set (See Table 5). Due to the 8,192-token context limit enforced by the DeepSeek-R1 API, evaluations for this model are restricted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

PC		COE	
Sub-item	Rate (%)	Dimension	Rate (%)
Acknowledge Preference	83	EA	99
Violate Preference	88	IS	100
Hallucinate Preference	72	DA	100
Helpful Response	99	EG	100
Average	87	-	-

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).

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

316 criteria effectively capture the core requirements
 317 of child-directed communication and can be reli-
 318 ably applied at scale. Rather than distinguish-
 319 ing subtle stylistic variations, the COE focuses on
 320 assessing whether model responses satisfy essen-
 321 tial principles of child-directed interaction under a
 322 unified safety standard. This design ensures that
 323 the evaluation remains robust, reproducible, and
 324 well aligned with the practical objectives of child-
 325 oriented model assessment.

326 4.3 Cross-Model Bias Analysis for Validating 327 LLM-as-a-Judge

328 To further assess potential model-specific bias and
 329 validate our evaluation metrics, we use the same
 330 randomly sampled examples as in Section 4.2, cover-
 331 ing all preference types, and evaluate them using
 332 DeepSeek-R1. We compute agreement rates be-
 333 tween DeepSeek-R1 and Qwen2.5-72B, and addi-
 334 tionally compare the outputs of both models against
 335 human judgments. As shown in Tables 2 and 8,
 336 the evaluations produced by the two models are
 337 highly consistent and closely aligned with human
 338 assessments. Notably, for child-oriented evaluation
 339 metrics, the agreement is even higher, indicating

robust and reliable evaluation performance. 340

341 4.4 Preference Consistency Evaluation

342 **SOTA LLMs struggle to maintain personaliza-**
 343 **tion across long-term interactions.** As shown
 344 in Figure 2, all prompt-based LLMs exhibit a de-
 345 crease in accuracy when generating personalized
 346 responses after inserting irrelevant dialogues, com-
 347 pared to directly expressing preferences without
 348 any intervening turns. However, as the number of
 349 irrelevant turns increases, the performance degra-
 350 dation gradually slows down. Interestingly, for
 351 some models (e.g. Qwen3-4B-instruct), additional
 352 irrelevant turns even lead to a slight recovery or
 353 improvement, suggesting a potential stabilizing.

354 **LLMs face greater difficulty in deducing im-**
 355 **PLICIT preferences than in understanding explicit**
 356 **ones.** Comparing the results in the left and right
 357 panels of Figure 2, it is evident that personaliza-
 358 tion consistency on implicit-preference datasets is
 359 lower than on explicit-preference datasets across

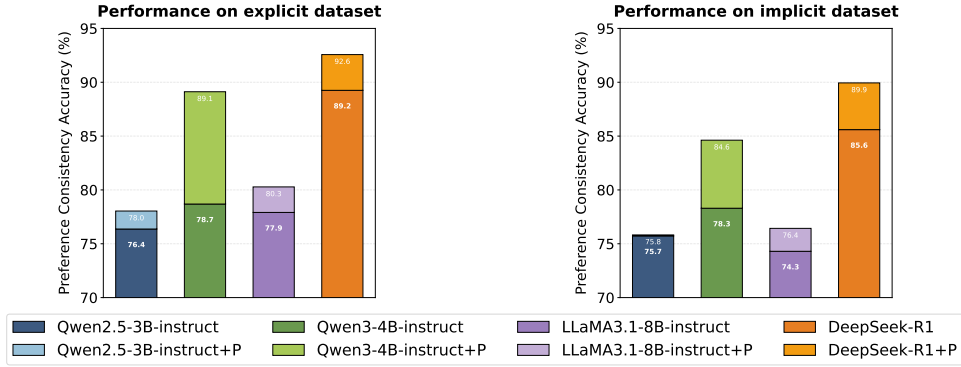


Figure 3: Performance on preference consistency when models respond to datasets with five irrelevant turns inserted, under zero-shot prompting (with and without persona). “+P” indicates persona-conditioned inference.

Model	Without Persona				With Persona			
	EA	IS	DA	EG	EA	IS	DA	EG
Explicit Data (%)								
Qwen2.5-3B-instruct	77.23	35.8	97.31	75.99	94.50	70.13	96.51	93.82
Qwen3-4B-instruct	96.29	52.28	99.82	97.59	98.05	82.19	99.58	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	97.55
Implicit Data (%)								
Qwen2.5-3B-instruct	78.73	38.58	98.01	77.41	93.16	69.92	96.33	93.67
Qwen3-4B-instruct	96.45	59.85	99.84	97.64	97.39	83.22	99.85	97.09
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 data in Table 4 with data in Figure 3 across multiple models, while consistency accuracy may be similar (e.g., 74%–76%), performance on child-oriented evaluation varies widely across dimensions, particularly in IS and DA. This suggests that a high consistency score alone does not necessarily reflect strong child-oriented personalization.

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 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

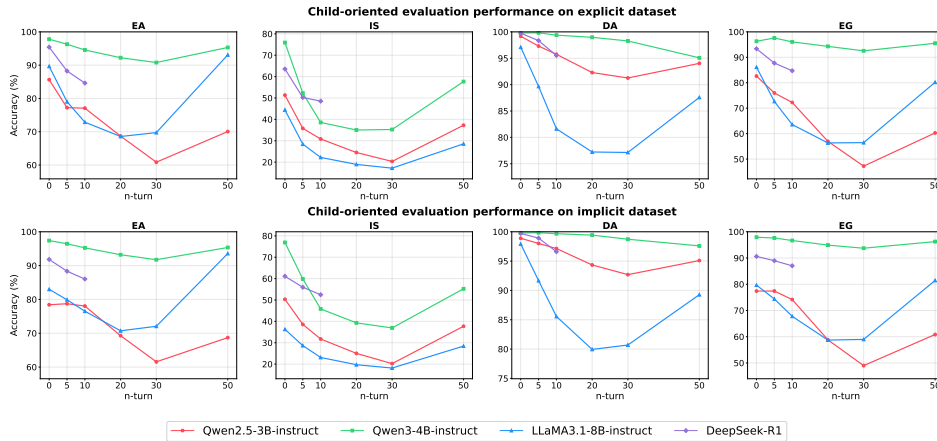


Figure 4: Accuracy of LLMs on different dimensions of child-oriented evaluation with varying numbers of inserted irrelevant turns (n-turn).

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.

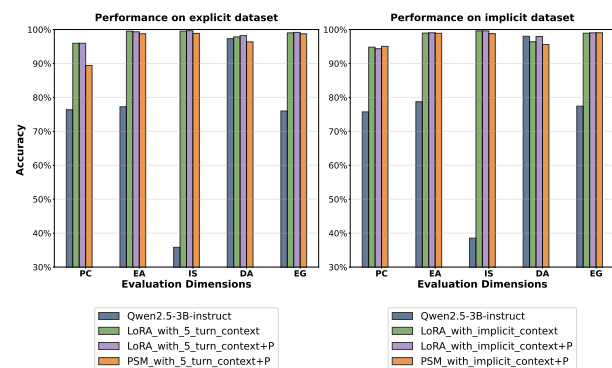


Figure 5: Finetuning results for children’s personalities on explicit and implicit datasets (both with 5-turn test dialogues). Explicit training added 5 unrelated utterances; implicit training used 6–10 consecutive turns. “Persona” denotes inclusion of child persona information during finetuning.

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 dialogue prompts, achieves higher preference consistency than PSM, which encodes persona as vectors.

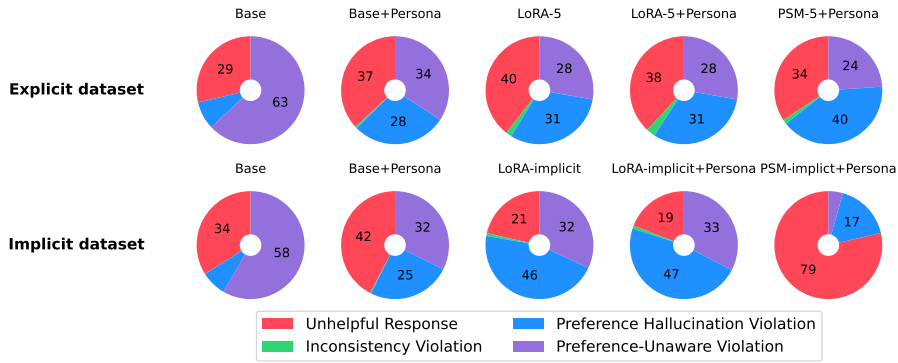


Figure 6: Distribution of preference consistency errors across 10-turn dialogues. Base refers to Qwen2.5-3B-Instruct; Base+Persona applies prompting with persona. LoRA-5 and PSM-5 denote LoRA- and PSM-based methods trained with 5-turn inserted context, with or without persona. LoRA-implicit and PSM-implicit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

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, ethically grounded testbed for analyzing child-focused personalization rather than simulating real child-LLM interactions. Experiments show that current open-source LLMs struggle with preference consistency in long interactions, particularly for implicit preferences. Moreover, personalization strategies involve trade-offs, and targeted finetuning can improve preference adherence while introducing new inconsistencies, highlighting both the potential and limits of current approaches.

521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570

Limitations

Synthetic Data and Realism. A primary limitation of this work is its reliance on a fully synthetic dataset to model children’s interactions. Language produced by children aged 3–6 is highly irregular, often involving disfluencies, incomplete utterances, grammatical errors, and unpredictable pragmatic behaviors, which are difficult to faithfully reproduce through synthetic generation. Consequently, the proposed benchmark does not aim to fully capture the complexity or variability of real-world child language use.

At the same time, the collection of 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 provides a practical and controllable approximation that enables reproducible evaluation, but does not substitute for real-world observations. Moreover, as a fully synthetic benchmark, the dataset inevitably reflects the inductive biases and stylistic tendencies of the language models used for data generation. The benchmark should therefore be interpreted as a diagnostic and lower-bound testbed for *relative model comparison*, 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, incorporating more realistic child language patterns and explicitly modeling safety–personalization interactions will likely require 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.

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.

References

Anthropic. 2024. *The claude 3 model family: Opus, sonnet, haiku*. 603
604

Alissa N Antle. 2008. Child-based personas: need, ability and experience. *Cognition, Technology & Work*, 10(2):155–166. 605
606
607

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. 608
609
610
611
612
613

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 614
615
616
617
618

619	perspectives. In <i>Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems</i> , pages 1–24.	
620		
621		
622	Jiaju Chen, Minglong Tang, Yuxuan Lu, Bingsheng Yao,	
623	Elissa Fan, Xiaojuan Ma, Ying Xu, Dakuo Wang,	
624	Yuling Sun, and Liang He. 2025b. Characterizing	
625	llm-empowered personalized story reading and inter-	
626	action for children: Insights from multi-stakeholder	
627	perspectives. <i>Conference on Human Factors in Com-</i>	
628	<i>puting Systems</i> .	
629	Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff	
630	Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré,	
631	Maria Lomeli, Lucas Hosseini, and Hervé Jégou.	
632	2024. The faiss library .	
633	Arafat Md Easin, Saha Sourav, and Orosz Tamás. 2024.	
634	An intelligent llm-powered personalized assistant	
635	for digital banking using langgraph and chain of	
636	thoughts. In <i>2024 IEEE 22nd Jubilee International</i>	
637	<i>Symposium on Intelligent Systems and Informatics</i>	
638	<i>(SISY)</i> , pages 625–630. IEEE.	
639	Tiantian Feng, Anfeng Xu, Rimita Lahiri, Helen Tager-	
640	Flusberg, So Hyun Kim, Somer Bishop, Catherine	
641	Lord, and Shrikanth Narayanan. 2024. Can generic	
642	llms help analyze child-adult interactions involv-	
643	ing children with autism in clinical observation?	
644	<i>Preprint</i> , arXiv:2411.10761.	
645	Gemini, :, Rohan Anil, Sebastian Borgeaud, and et al.	
646	2025. Gemini: A family of highly capable multi-	
647	modal models . <i>Preprint</i> , arXiv:2312.11805.	
648	Daya Guo, Dejian Yang, Haowei Zhang, Junxiao	
649	Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shi-	
650	rong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025.	
651	Deepseek-r1: Incentivizing reasoning capability in	
652	llms via reinforcement learning. <i>arXiv preprint</i>	
653	<i>arXiv:2501.12948</i> .	
654	Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan	
655	Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and	
656	Weizhu Chen. 2021. Lora: Low-rank adaptation of	
657	large language models . <i>Preprint</i> , arXiv:2106.09685.	
658	Bernd Huber, Ghazal Fazelnia, Andreas Damianou, Se-	
659	bastian Peleato, Max Lefarov, Praveen Ravichan-	
660	dran, Marco De Nadai, Mounia Lalmas-Roellke,	
661	and Paul N Bennett. 2025. Embedding-to-	
662	prefix: Parameter-efficient personalization for pre-	
663	trained large language models. <i>arXiv preprint</i>	
664	<i>arXiv:2505.17051</i> .	
665	Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li,	
666	Yuan Yuan, Sihao Chen, Lyle H. Ungar, Camillo J.	
667	Taylor, and Dan Roth. 2025. Know me, respond	
668	to me: Benchmarking llms for dynamic user pro-	
669	filing and personalized responses at scale . <i>CoRR</i> ,	
670	abs/2504.14225.	
671	Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra,	
672	Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt,	
673	Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang,	
674	Shubham Agarwal, Nedim Lipka, Chien Van Nguyen,	
	Thien Huu Nguyen, and Hamed Zamani. 2024.	675
	Longlamp: A benchmark for personalized long-form	676
	text generation . <i>Preprint</i> , arXiv:2407.11016.	677
	Juntao Li, Chang Liu, Chongyang Tao, Zhangming	678
	Chan, Dongyan Zhao, Min Zhang, and Rui Yan. 2021.	679
	Dialogue history matters! personalized response se-	680
	lection in multi-turn retrieval-based chatbots. <i>ACM</i>	681
	<i>Transactions on Information Systems (TOIS)</i> , 39(4):1–	682
	25.	683
	Xinyu Li, Zachary C. Lipton, and Liu Leqi. 2024. Per-	684
	sonalized language modeling from personalized hu-	685
	man feedback . <i>CoRR</i> , abs/2402.05133.	686
	Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai,	687
	Jieming Zhu, Minda Hu, Menglin Yang, and Irwin	688
	King. 2025. A survey of personalized large lan-	689
	guage models: Progress and future directions. <i>arXiv</i>	690
	<i>preprint arXiv:2502.11528</i> .	691
	Jing Liu and Abdellah Fourtassi. 2024. Benchmark-	692
	ing llms for mimicking child-caregiver language in	693
	interaction. <i>arXiv preprint arXiv:2412.09318</i> .	694
	Aman Madaan, Niket Tandon, Peter Clark, and Yim-	695
	ing Yang. 2022. Memory-assisted prompt editing to	696
	improve GPT-3 after deployment . In <i>Proceedings of</i>	697
	<i>the 2022 Conference on Empirical Methods in Natu-</i>	698
	<i>ral Language Processing, EMNLP 2022, Abu Dhabi,</i>	699
	<i>United Arab Emirates, December 7-11, 2022</i> , pages	700
	2833–2861. Association for Computational Linguis-	701
	tics.	702
	Lucie Charlotte Magister, Katherine Metcalf, Yizhe	703
	Zhang, and Maartje ter Hoeve. 2024. On the way	704
	to llm personalization: Learning to remember user	705
	conversations. <i>arXiv preprint arXiv:2411.13405</i> .	706
	MOE. 2012. Notice of the ministry of education on	707
	issuing the guidelines for learning and development	708
	of children aged 3–6 . Ministry of Education of the	709
	People’s Republic of China, Document No. Jiao Ji Er	710
	(2012) No. 4.	711
	Jisoo Mok, Ik hwan Kim, Sangkwon Park, and Sun-	712
	groh Yoon. 2025. Exploring the potential of llms	713
	as personalized assistants: Dataset, evaluation, and	714
	analysis . <i>Preprint</i> , arXiv:2506.01262.	715
	Mir Tafseer Nayeem and Davood Rafiei. 2024. KidLM:	716
	Advancing language models for children – early in-	717
	sights and future directions . In <i>Proceedings of the</i>	718
	<i>2024 Conference on Empirical Methods in Natu-</i>	719
	<i>ral Language Processing</i> , pages 4813–4836, Miami,	720
	Florida, USA. Association for Computational Lin-	721
	guistics.	722
	OpenAI, :, Aaron Jaech, Adam Kalai, and et al. 2024.	723
	Openai o1 system card . <i>Preprint</i> , arXiv:2412.16720.	724
	Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang	725
	Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-	726
	Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng	727
	Gao. 2025. On memory construction and retrieval	728
	for personalized conversational agents . <i>Preprint</i> ,	729
	arXiv:2502.05589.	730

840 dialogue turns between the disclosure of children’s
841 preferences and the related utterances. Although
842 we initially considered extending dialogues to 50K
843 tokens, the backbone model supports at most 30K,
844 beyond which outputs became unstable. This setup
845 creates a challenging evaluation for LLMs to infer,
846 retrieve, and utilize children’s preferences in long
847 dialogues, especially when interspersed with unre-
848 lated topics. For dialogues of varying turn counts,
849 we randomly sample and fix their lengths, with
850 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.

851 A.2 Model Version

852 In our experiments, we employ the Bge-Large-
853 Zh model as the text encoder. Table 6 provides
854 an overview of the evaluated LLMs and their ver-
855 sions, together with the text encoder version. For
856 DeepSeek, we conduct our experiments using the
857 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.

858 A.3 ChildEval Example

859 An example from ChildEval is presented in Table
860 7.

861 A.4 Prompts Design

862 A.4.1 Prompts for Data Construction

863 The prompts used within this work are listed in
864 Figures 7–9. Some prompts are too long to fit on
865 a single page, so we split them into two figures, as
866 shown in Figure 8 and 9.

867 A.4.2 Prompts for the Prompting-based 868 Approach

869 We extensively evaluate a variety of state-of-the-
870 art LLMs using zero-shot prompts, both with and
871 without persona information. In the default zero-
872 shot setting, the LLM answers the user’s query
873 directly without any additional prompting. How-
874 ever, these models are not specifically designed for
875 child-oriented dialogue. If used without modifica-
876 tion, they tend to generate overly long responses
877 that do not reflect the conversational style of young
878 children. To ensure a fair evaluation, we accord-
879 ingly augmented the original dialogue prompts as
880 follows, corresponding to the with-persona and
881 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.

882 A.4.3 Evaluation Prompts for Child-Oriented 883 Tasks

884 The evaluation prompts for child-oriented tasks
885 are shown in Figures 10–13, which correspond
886 respectively to Emotional Adaptation, Interaction
887 Scaffolding, Developmental Appropriateness and
888 Engagement.

889 A.5 Cross-Model Bias Analysis for Validating 890 LLM-as-a-Judge Metrics

891 To further assess potential model-specific bias and
892 validate our evaluation metrics, we use the same
893 randomly sampled examples as in Section 4.2, cov-
894 ering all preference types, and evaluate them using
895 DeepSeek-R1. We compute agreement rates be-
896 tween DeepSeek-R1 and Qwen2.5-72B, and addi-
897 tionally compare the outputs of both models against
898 human judgments. As shown in Tables 2 and 8,
899 the evaluations produced by the two models are
900
901

Topic	Persona	Explicit Preference	Implicit Preference	Utterance
<p>Cognitive development and exploration Nature&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.

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).

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.

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 14, 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 14, 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:

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

$$= h_L W^T + \Delta(h_L) W^T$$

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 14. 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 15 and 16 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

982 Violations in PSM decreases, while Unhelpful Re- 1033
983 sponses become increasingly dominant. On the 1034
984 implicit dataset, Unhelpful Responses constitute 1035
985 the primary error type for PSM, indicating a ten- 1036
986 dency to refuse or provide unhelpful answers rather 1037
987 than attempt alignment with user preferences. 1038

988 Overall, these results highlight the trade-offs be- 1039
989 tween proactive preference-following and robust- 1040
990 ness to irrelevant context. Notably, as the length of 1041
991 irrelevant context increases, PSM becomes increas- 1042
992 ingly conservative, producing unhelpful responses, 1043
993 whereas LoRA is more proactive, continuing to 1044
994 attempt responses aligned with user preferences, 1045
995 although alignment issues remain. 1046

996 **A.7.2 Effect of Inserted Context Length on** 1047 997 **Finetuning Results** 1048

998 Figure 17 illustrates how persona-informed finetun- 1049
999 ing methods (i.e., PSM and LoRA) evolve with 1050
1000 increasing dialogue length on both explicit and 1051
1001 implicit datasets (i.e., datasets with historical dia- 1052
1002 logues that explicitly and implicitly reveal the child 1053
1003 preference). On the explicit dataset, PSM-based 1054
1004 models show a relatively sharp decline in prefer- 1055
1005 ence consistency as the number of inserted irrele- 1056
1006 vant dialogue turns increases, while LoRA-based 1057
1007 models exhibit a moderate decrease. Moreover, 1058
1008 results on the PC (Preference Consistency) dimen- 1059
1009 sion indicate that training with longer irrelevant 1060
1010 dialogues yields greater robustness on equally long 1061
1011 test dialogues than training with shorter ones. In- 1062
1012 terestingly, within the child-oriented dimensions, 1063
1013 most metrics remain relatively stable across dia- 1064
1014 logue lengths, whereas developmental appropriateness 1065
1015 (DA) exhibits the largest fluctuations, indicat- 1066
1016 ing its heightened sensitivity to contextual length.

1017 On the implicit dataset, model trends largely 1067
1018 mirror those observed on the explicit dataset. In 1068
1019 the PC dimension, PSM-based models remain rel- 1069
1020 atively stable compared to LoRA-based models 1070
1021 on dialogues shorter than 30 turns (12K tokens) 1071
1022 and benefit more from inserted irrelevant dialogues. 1072
1023 However, in longer dialogues (e.g., 50 turns, PSM- 1073
1024 based models show a sharper decline, falling below 1074
1025 LoRA-based models. This pattern may be due 1075
1026 to the fact that PSM relies on the aggregation of 1076
1027 final-layer vectors to incorporate the persona in- 1077
1028 formation into an LLM, which works well when 1078
1029 the inserted irrelevant dialogue is short, but may 1079
1030 be negatively affected by accumulated noise when 1080
1031 the irrelevant dialogue is long. In contrast, LoRA’s 1081
1032 low-rank adaptation maintains greater stability in

extended contexts.

1034 **A.7.3 Performance on English version** 1035

1036 To gain a comprehensive understanding, we con- 1037
1038 duct more experiments on the English version of 1038
1039 ChildEval. Figure 18 presents the evaluation re- 1039
1040 sults across different numbers of inserted irrele- 1040
1041 vant dialogue turns. These dynamics indicate that 1041
1042 tasks and models exhibit varying levels of robustness 1042
1043 and adaptability across different dialogue stages in 1043
1044 both the explicit and implicit datasets. IS remains 1044
1045 the most challenging dimension for all models. How- 1045
1046 ever, the overall performance on the English dataset 1046
1047 is slightly lower, likely because it is a translated 1047
1048 counterpart of the Chinese corpus and may not fully 1048
1049 capture the natural distribution of native English 1049
1050 dialogues. 1050

1051 Figure 18 presents the evaluation results af- 1051
1052 ter incorporating the persona information into the 1052
1053 prompt, and different models exhibit divergent pat- 1053
1054 terns. Notably, LLaMA3.1-8B-instruct shows sub- 1054
1055 stantial fluctuations on the EA, IS, and EG dimen- 1055
1056 sions of the child-oriented evaluation. The perfor- 1056
1057 mances of the other two models show a decreasing 1057
1058 trend with small fluctuations as the number of irrel- 1058
1059 evant dialogue turns increases. Comparing Figure 1059
1060 18 and Figure 19, the inclusion of persona leads to 1060
1061 significant improvements across PC, EA, IS, and 1061
1062 EG for all models, with the only exception being a 1062
1063 slight decrease on DA observed for Ministral-7B- 1063
1064 instruct. 1064

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 7: 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 8: 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 9: 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 10: 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 12: 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 11: 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 13: Evaluation prompt used for checking Engagement.

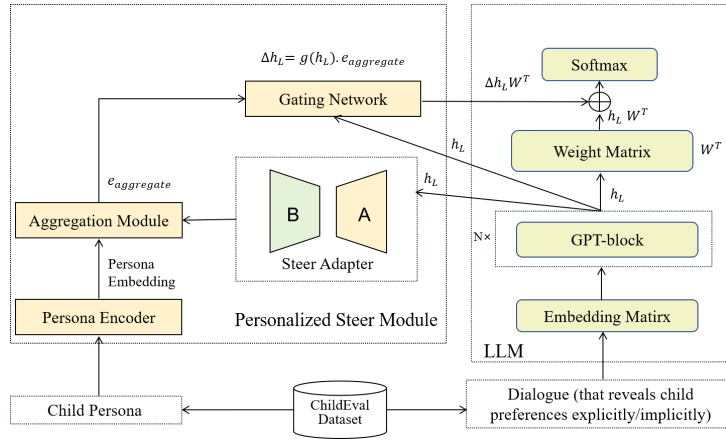


Figure 14: The architecture of the persona steer model.

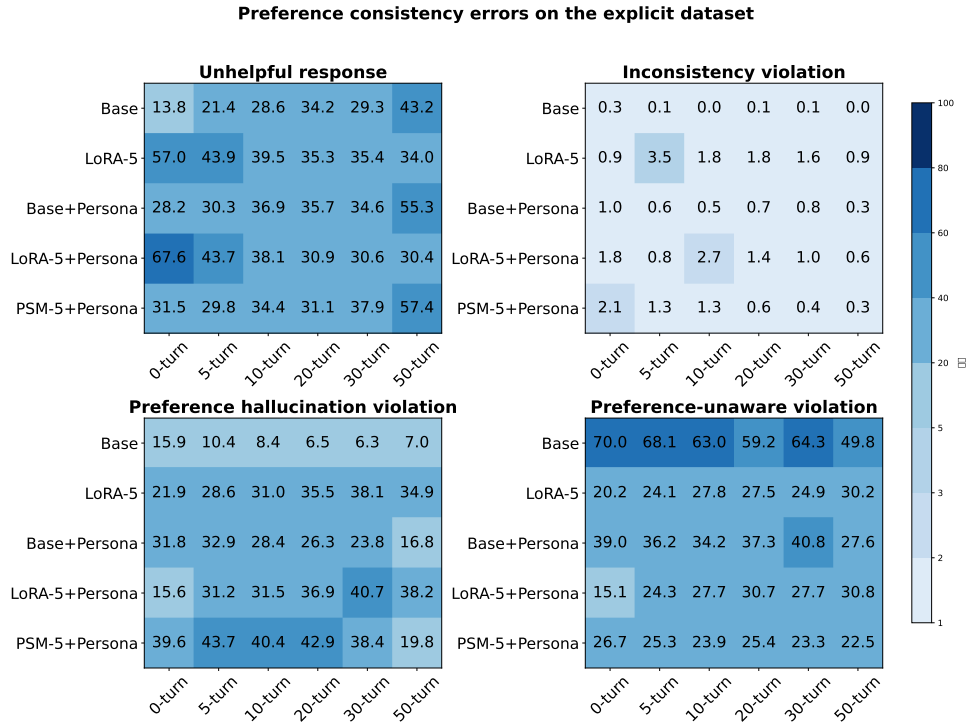


Figure 15: Preference consistency error types under different numbers of inserted irrelevant turns (n-turn).

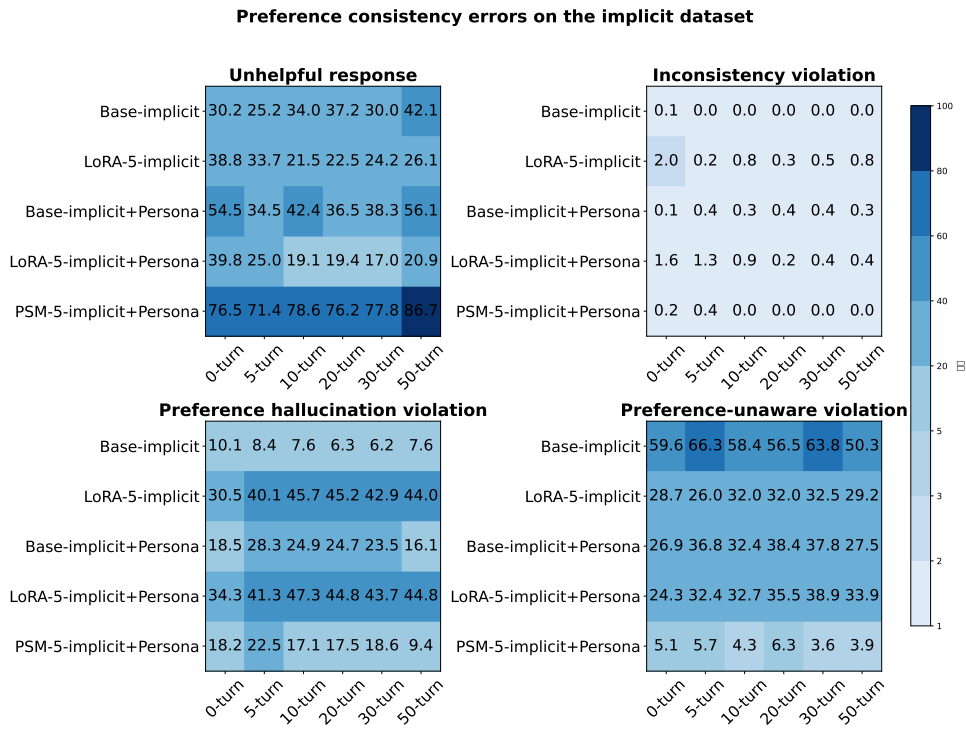


Figure 16: Preference consistency error types under different numbers of inserted irrelevant turns (n-turn).

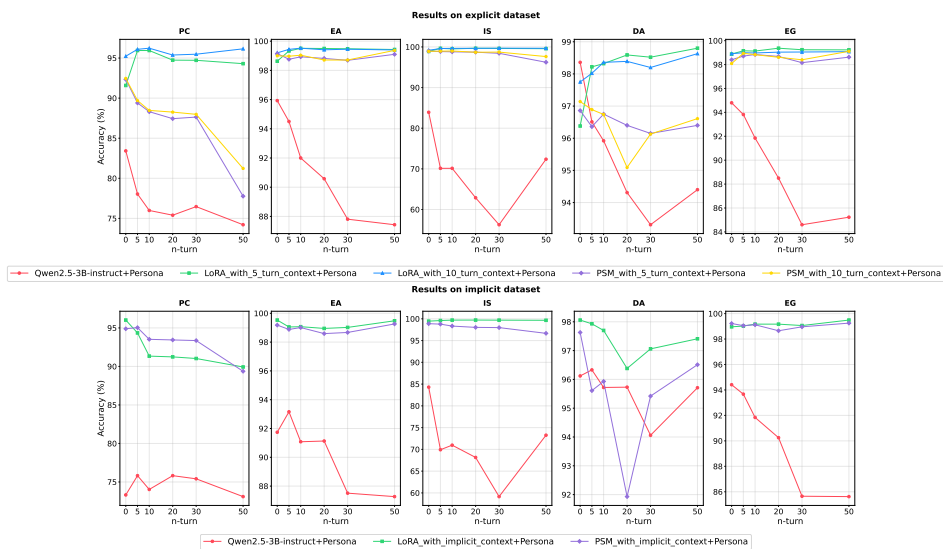


Figure 17: Accuracy of LLMs on preference consistency (PC) and child-oriented dimensions under different numbers of inserted irrelevant turns (n-turn).

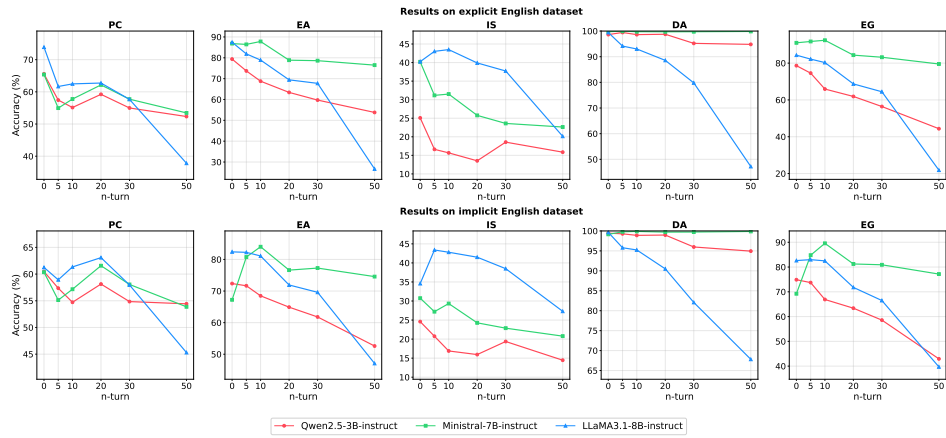


Figure 18: LLMs performances on preference consistency (PC) and the child-oriented evaluation under different numbers of inserted irrelevant dialogue turns on the English dataset.

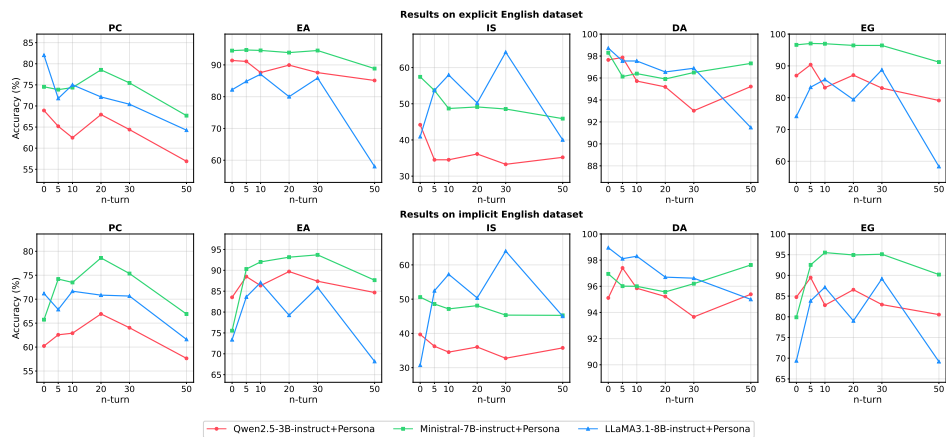


Figure 19: LLMs performances on preference consistency (PC) and the child-oriented evaluation under different numbers of inserted irrelevant dialogue turns on the English dataset, after integrating persona information into the prompt.