ChildEC: A Developmentally Stratified Chinese Dialogue Corpus for Child Emotion Coaching

Yuelin Ding[1] and Sujuan Liu[2] (✉)

[1] Shangxing Innovation Division, Tianjin University of Science and Technology, Tianjin 300457, China
2 College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China
sjliu@tust.edu.cn

Abstract. Existing dialogue resources for minors mainly focus on general psychological support, adolescent positive mental health promotion, or broad emotional companionship, and therefore do not adequately support the task of everyday emotion coaching for Chinese children. To address this gap, we construct ChildEC, a developmentally stratified multi-turn dialogue corpus for everyday emotion coaching, targeting Chinese children aged 8–12. The corpus contains 1,709 multi-turn dialogues spanning two developmental stages and five core themes. To support corpus construction and evaluation, we further propose the Theory-Grounded Emotion Coaching Framework (TGEC). Within this framework, TGEC-Synth operationalizes the five-step emotion coaching model and Socratic dialogue into a multi-stage dialogue synthesis pipeline, while TGEC-Eval assesses the quality of child emotion coaching dialogues along six key dimensions. Experimental results show that models fine-tuned on ChildEC consistently outperform their corresponding base models under both automatic metrics and LLM-as-a-Judge evaluation, demonstrating the downstream training value of the corpus for child emotion coaching. Overall, ChildEC provides a specialized data resource and an evaluation reference for developmentally sensitive and theory-consistent research on Chinese child emotion coaching. The dataset and prompts are publicly available at https://github.com/ChildEC-Project/ChildEC.

Keywords: Child emotion coaching · Multi-turn dialogue corpus · Large language models

1 Introduction

As large language models are increasingly applied in settings such as educational companionship and psychological support, conversational AI for children has received growing attention [8,12,18]. However, children differ markedly from adults in cognitive development, language comprehension, and expressive style [14]. These differences imply that multi-turn interactions for children require dedicated design in both support strategies and conversational pacing, rather than direct adaptation from adult-oriented settings. Existing studies have begun to show that child-facing scenarios

require targeted interaction design and corresponding resource support [18,27]. However, prior work has focused more on parent–child collaborative education, broad emotional companionship, or more general mental health support, and has yet to provide suitable multi-turn dialogue corpora for everyday child emotion coaching.

More specifically, there is a clear mismatch between the task formulations of existing resources and the requirements of child emotion coaching. On the one hand, existing Chinese psychological support and counseling corpora are primarily designed for general mental health support tasks [19,25]. On the other hand, although some recent studies have extended the target population to adolescents, their main focus remains the promotion of positive mental health and well-being [16]. Existing child-related studies, in contrast, have largely centered on parent–child collaborative education or broader forms of emotional companionship [12,18,27], rather than treating everyday child emotion coaching as the core task setting. This mismatch limits their suitability for the specific task of child emotion coaching.

However, constructing high-quality dialogue resources and developing AI systems for this task still present several challenges. First, developmental alignment is essential. Children at different ages differ substantially in cognitive development, language comprehension, and expressive style [8,14]. Although existing resources provide important insights for child-facing interaction, they typically do not systematically calibrate language style or guidance pacing around developmental differences. As a result, they cannot directly support developmentally sensitive multi-turn emotion coaching. Second, explicit modeling of the child emotion coaching process remains limited. Many mental health corpora are derived from general support dialogues, web text, or reconstructed counseling reports [10,15,19]. These resources tend to emphasize emotion regulation or broader mental health support, rather than operationalizing the relevant theories of child emotion coaching into a controllable process. Third, existing evaluation frameworks remain insufficient. Traditional natural language processing metrics, such as BLEU [13], ROUGE [9], and BERTScore [26], are not well suited to capturing key dimensions of multi-turn dialogue quality, including emotion labeling and empathic response [22]. More general counseling evaluation frameworks are also unable to fully reflect the specific requirements of everyday child emotion coaching [25]. Together, these limitations highlight the need for a task-specific evaluation framework for child emotion coaching.

To address the challenges above, we construct ChildEC, a developmentally stratified multi-turn dialogue corpus for Chinese children aged 8–12. Fig. 1 shows a representative instance from ChildEC. The corpus contains 1,709 high-quality multi-turn dialogues covering everyday emotional difficulties related to academic pressure, peer interaction, family interaction, self-evaluation, and responsibility and guilt. ChildEC focuses on non-clinical everyday child emotion coaching scenarios. To support both corpus construction and evaluation, we further propose a Theory-Grounded Emotion Coaching Framework (TGEC). TGEC consists of two components: TGEC-Synth, a multi-stage dialogue synthesis pipeline, and TGEC-Eval, an evaluation component built around theory-grounded metrics. Specifically, TGEC-Synth integrates John Gottman's five-step model of emotion coaching [6] and Socratic dialogue [3] into a multi-stage dialogue synthesis process, while TGEC-Eval assesses the quality of child

emotion coaching dialogues along six key dimensions. Experimental results demonstrate the clear task utility and downstream training value of ChildEC.

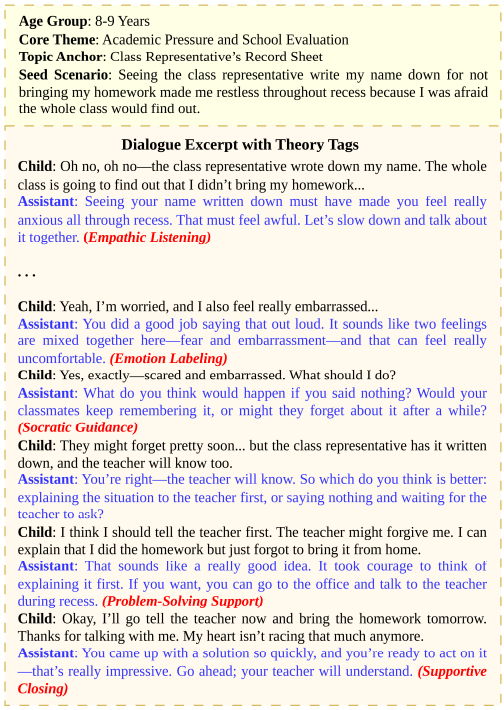

Fig. 1. A representative instance from ChildEC, including its developmental stage, core theme, topic anchor, seed scenario, and a dialogue excerpt annotated with theory tags.

Our main contributions are as follows:

- We construct ChildEC, a developmentally stratified Chinese dialogue corpus for child emotion coaching. The corpus provides a specialized data resource for research on child emotion coaching in Chinese.
- We propose a theory-grounded multi-stage dialogue synthesis pipeline. This pipeline balances theoretical consistency, process controllability, and scalable data generation in the construction of child emotion coaching dialogues.
- We design a metric-based evaluation framework for child emotion coaching and use it to validate the downstream value of the corpus.

2 Related Work

Recent research on affective interaction in child-facing settings has begun to show that children require interaction goals and modes of use specifically designed for them. For

example, ChaCha combines a state machine with a large language model to encourage children aged 8–12 to share personal experiences and the emotions associated with them [18]. PACEE targets emotion education for children aged 3–6 and adopts a parent-centered AI scaffolding framework [12]. QiaoBan investigates parental emotion coaching in parent–child interaction [27], whereas the Sahar Dataset integrates child-centric empathetic support with knowledge-oriented interaction [1]. However, these studies do not sufficiently account for developmental differences across childhood, nor do they build more targeted multi-turn interactions around everyday emotional issues.

Beyond child-facing settings, the broader literature on adolescent mental health promotion, general psychological support, and counseling has accumulated a substantial body of data resources and task formulations. DeepWell-Adol focuses on promoting positive mental health and well-being among adolescents [16]. ESConv constructs a multi-turn emotional support dialogue dataset with explicit support strategy annotations [10], while SMILE expands single-turn data into multi-turn mental health support conversations [15]. Meanwhile, research in counseling-oriented settings places greater emphasis on modeling the counseling process as a multi-turn interaction. For example, CPsyCoun reconstructs multi-turn Chinese counseling dialogues from counseling reports [25], and CPsDD builds a large-scale Chinese psychological support dialogue dataset through path-guided constraints and multi-agent collaboration [19]. However, these studies still primarily target adolescents or general populations, and their task formulations remain centered on broad mental health support scenarios rather than the construction of resources for everyday emotion coaching dialogues for Chinese children.

Beyond differences in task formulation and resource scope, related work has also shown a clear methodological shift in data construction, moving from direct generation toward explicit planning and stage-aware control. DeepWell-Adol adopts a construction framework that combines expert-written data with two-stage scenario-based augmentation [16]. CPsDD employs a framework that integrates path-guided constraints with generation and revision [19]. MindChat constructs psychological support corpora through a multi-agent synthesis pipeline based on role-playing [23], while DeepPsyAgent models psychological support as a dialogue system capable of perceiving stage transitions [4]. Taken together, these studies suggest that a construction paradigm based on planning, generation, and review is becoming an increasingly compelling methodological route. This trend provides an important methodological reference for our data construction framework.

3 A Theory-Grounded Emotion Coaching Framework

We propose a Theory-Grounded Emotion Coaching Framework (TGEC) to guide the construction and evaluation of ChildEC. The framework targets everyday emotional difficulties encountered by Chinese children aged 8–12, with a clear focus on non-clinical and non-therapeutic scenarios. TGEC is grounded in John Gottman's five-step model of emotion coaching [6], which proceeds through emotion awareness, viewing emotional moments as opportunities, empathic listening, emotion labeling, and setting

limits while supporting problem solving. It further incorporates Socratic dialogue [3], using open-ended, concrete, and non-leading questions to help children understand their own thoughts and develop constructive coping strategies. TGEC consists of two components: TGEC-Synth, a multi-stage dialogue synthesis pipeline, and TGEC-Eval, an evaluation component consisting of six theory-grounded metrics. Together, they provide an efficient and reproducible methodology for constructing high-quality dialogue corpora for child emotion coaching tasks.

3.1 TGEC-Synth: Multi-Stage Dialogue Synthesis Pipeline

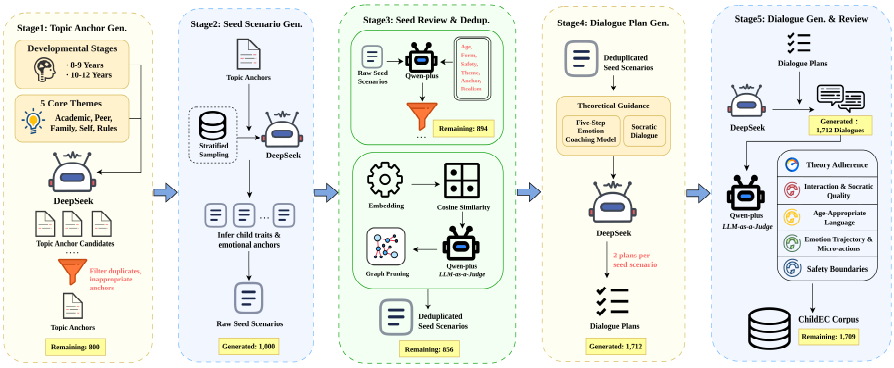

Fig. 2. Overview of TGEC-Synth, the theory-grounded multi-stage pipeline for constructing ChildEC. The pipeline consists of five stages: topic anchor generation, seed scenario generation, seed scenario review and deduplication, dialogue plan generation, and dialogue generation with dialogue review.

TGEC-Synth serves as the dialogue construction module of TGEC, operationalizing the five-step emotion coaching model and Socratic dialogue into a multi-stage dialogue synthesis pipeline. As shown in Fig. 2, the pipeline consists of five stages: topic anchor generation, seed scenario generation, seed scenario review and deduplication, dialogue plan generation, and dialogue generation with dialogue review. TGEC-Synth enables AI systems to generate multi-turn dialogue content that is developmentally appropriate for children and aligned with the logic of emotion coaching. Through this process, the resulting dialogues are designed to guide children to gradually recognize their own emotions while expressing and understanding everyday emotional difficulties, thereby supporting more constructive coping and self-growth.

3.2 TGEC-Eval: Evaluation Metrics

To ensure that the generated dialogues effectively support child emotion coaching, TGEC introduces TGEC-Eval as the evaluation component of the framework. Rather than merely assessing whether a dialogue alleviates children's emotional distress, these metrics examine whether the dialogue can appropriately advance the emotion coaching process. TGEC-Eval evaluates six dimensions: (1) emotion coaching pacing, (2)

listening and empathy, (3) emotion understanding and labeling, (4) practical usefulness, (5) quality of Socratic guidance, and (6) age-appropriate expression. These dimensions are grounded in emotion coaching, Socratic guidance, and child developmental considerations [6,14]. Each dimension is operationalized as a quantifiable scoring criterion in the unified TGEC-Eval scoring prompt, which is summarized in Table 5 in Appendix A.

TGEC-Eval is designed to complement conventional automatic text-based metrics. It provides a task-relevant and theory-consistent quality measure for child emotion coaching dialogues. In ChildEC, it can be used to compare the child emotion coaching capabilities of different models.

4 ChildEC Dataset Construction

Given the profound domain expertise required for child emotion coaching, manually constructing a large-scale, high-quality dialogue dataset is prohibitively expensive. To address this challenge, we propose TGEC-Synth, a multi-stage pipeline for emotion coaching dialogue synthesis guided by the TGEC framework. By simulating an expert-informed corpus construction process, this pipeline decomposes the complex task of generating child emotion coaching dialogues into multiple controllable sub-tasks, thereby facilitating the efficient and reproducible generation of multi-turn dialogue corpora. Based on this pipeline, we construct ChildEC, a developmentally stratified, multi-turn emotion coaching dialogue dataset specifically tailored for Chinese children aged 8–12.

4.1 Scenario Space and Topic Anchor Generation

The design of the scenario space determines the range of everyday emotional difficulties covered by ChildEC. As shown in Fig. 2, we first divide children aged 8–12 into two developmental stages, namely 8–9 and 10–12 years old, based on child developmental psychology and cognitive development theories [11,14]. Building on this developmental stratification, we define five core themes: (1) academic pressure and school evaluation, (2) peer interaction and interpersonal difficulties, (3) family interaction and parent–child conflict, (4) self-evaluation and developmental setbacks, and (5) rule violation, responsibility, and guilt. These themes cover common everyday emotional difficulty scenarios for children aged 8–12 and provide a stable scenario space for subsequent construction.

We further construct topic anchors to constrain the generation boundary of subsequent seed scenarios. For each developmental stage, we design prompts conditioned on the corresponding age group and core theme, and use DeepSeek-V3.2-chat [5] to generate developmentally grounded topic anchors. The full prompts used for all TGEC-Synth stages and TGEC-Eval scoring are summarized in Table 5 in Appendix A and released in full in the project repository. A topic anchor is a concrete everyday element with emotional potential that may give rise to the kinds of emotional difficulties children encounter in daily life. To ensure the quality and diversity of the generated topic

anchors, we remove anchors that are duplicated, developmentally inappropriate, or contain explicit emotional judgments. Finally, we retain 800 high-quality topic anchors for subsequent seed scenario generation.

4.2 Seed Scenario Generation

After obtaining high-quality topic anchors, we further generate seed scenarios as the scenario starting points for subsequent dialogue planning and dialogue generation. Specifically, we prompt DeepSeek-V3.2-chat to serve as a child life-scenario creator. The LLM receives topic anchors conditioned on a specific developmental stage and core theme as input. It first infers the cognitive and psychological characteristics of children in that group, then derives a specific emotional anchor for each topic anchor and generates the corresponding seed scenario in a one-to-one manner. Each seed scenario is a concise scenario description that depicts an everyday emotional difficulty encountered by a child.

To ensure the coverage and distributional stability of seed scenarios, we introduce a stratified sampling strategy at the input level. Specifically, we partition the topic-anchor space into 10 age-group–theme buckets, each defined by one combination of the two developmental stages and five core themes. Each bucket contains 80 topic anchors. On this basis, we first ensure that all anchors within each bucket are used once to generate 80 base seed scenarios. We then sample an additional 20 anchors from the same bucket to generate 20 supplementary seed scenarios. The entire sampling process is conducted only within the same bucket, avoiding mixture across developmental stages or core themes. Finally, we obtain 1,000 raw seed scenarios.

4.3 Seed Scenario Review and Deduplication

Seed Scenario Review. To ensure that the seed scenarios can be used in the subsequent dialogue planning stage, we review all 1,000 seed scenarios. Specifically, we use Qwen-plus to conduct a multi-dimensional quality assessment of the seed scenarios. In this paper, Qwen-plus refers to the Qwen-plus-2025-09-11 model version [24]. The review prompt asks the LLM to judge each seed scenario along six dimensions and provide a final decision: (1) age appropriateness, (2) formal validity and non-intervention, (3) safety boundary, (4) thematic relevance and situational tension, (5) anchor binding and emotional consistency, and (6) local realism. After the initial review, we manually revise the samples judged as revisable and then use Qwen-plus to review them again. This process is repeated until the revised samples meet the review criteria. Finally, we obtain 894 seed scenarios that pass the review.

Seed Scenario Deduplication. After obtaining the 894 reviewed seed scenarios, we further perform semantic deduplication to reduce scenario redundancy in the corpus. This stage targets training-value redundancy, where seed scenarios with different surface forms may still induce nearly identical dialogue trajectories. Specifically, we first

divide the seed scenarios into 10 non-overlapping buckets according to the two age groups and five themes. Deduplication is performed only within the same bucket to avoid mistaken deletion across developmental stages or themes. We then use bge-large-zh-v1.5 [21] to vectorize the content of each seed scenario and compute a cosine similarity matrix within each bucket. Based on the number of candidate pairs retained under different thresholds and the overall similarity distribution, we set the final recall threshold to 0.8. Seed scenario pairs above this threshold are retained as candidate duplicates for further LLM-based judgment.

For these candidate pairs, we further use Qwen-plus as an LLM judge to determine whether two seed scenarios are redundant in scenario value and may therefore lead to similar dialogue trajectories. For seed scenario pairs judged as redundant, we treat them as edges in an undirected graph and conduct graph pruning based on connected components. In each duplicate cluster, only the node with the smallest seed scenario identifier is retained, while the other nodes are removed. This process finally yields 856 deduplicated seed scenarios.

4.4 Dialogue Plan Generation

To avoid insufficient theory adherence in direct dialogue generation, we introduce a dialogue planning stage before generating full conversations. In this stage, we prompt DeepSeek-V3.2-chat to act as a dialogue planning expert for child emotion coaching. Specifically, given a seed scenario, a target number of dialogue turns, and other contextual information, the LLM first analyzes the child's state in the scenario. It then maps the five-step emotion coaching model and Socratic dialogue into the rhythm of a multi-turn conversation, and generates a dialogue plan. In this stage, we generate two dialogue plans with different turn counts for each seed scenario, resulting in 1,712 dialogue plans in total.

4.5 Dialogue Generation and Review

Dialogue Generation. Guided by the fine-grained dialogue plans obtained from the previous stage, we proceed to generate concrete multi-turn emotion coaching dialogues. At this stage, we prompt DeepSeek-V3.2-chat to act as a child emotion coaching dialogue designer. Specifically, the LLM is required to follow the dialogue plan and translate the five-step emotion coaching model and Socratic dialogue into a coherent interactive conversational flow. This stage yields 1,712 initial multi-turn dialogues.

Dialogue Review. To ensure corpus quality, we introduce a review stage after dialogue generation. Specifically, we use Qwen-plus as an LLM judge to assess each dialogue along multiple dimensions. The review criteria include: (1) adherence to the theoretical framework, (2) interaction quality and Socratic questioning, (3) age-appropriate language and cognitive alignment, (4) emotional trajectory and the executability of scenario-specific coping actions, and (5) safety boundaries. After this review stage, we

filter out 3 unqualified dialogues and retain 1,709 high-quality multi-turn dialogues as the core corpus of ChildEC.

5 Dataset Analysis

5.1 Statistics of ChildEC

Distribution Across Developmental Stages and Core Themes. As shown in Table 1, ChildEC exhibits a generally balanced dialogue distribution across the two developmental stages and five core themes. No significant bias toward either age group is observed. The number of dialogues is comparable across themes, and the sample sizes of individual age-group × theme buckets also remain at similar levels. This balanced distribution preserves the developmentally stratified design of the dataset. It also ensures adequate coverage of the diverse everyday emotional difficulty scenarios that children encounter.

Table 1. Distribution of ChildEC dialogues across developmental stages and core themes. Academic, Peer, Family, Self, and Rules denote the five core themes defined in Section 4.1.

Age Group	Academic	Peer	Family	Self	Rules	Total
8–9	174	166	172	154	170	836
10–12	187	167	172	181	166	873
Total	361	333	344	335	336	1,709

Distribution of Emotion Categories. The distribution of emotion categories is presented in Fig. 3. Overall, ChildEC covers a broad range of children's emotional difficulties. Among them, anxiety and dejection are the most frequent emotion categories in the dataset. Shame, embarrassment, and frustration also account for substantial proportions. This pattern is consistent with children's real emotional experiences in everyday situations. It also ensures broad emotional coverage and diversity in ChildEC.

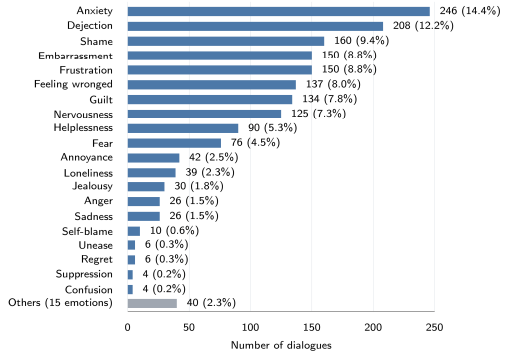

Fig. 3. Distribution of the top 20 most frequent emotion categories in ChildEC dialogues. The remaining categories are merged into “Others”.

5.2 Dataset Comparison

Table 2 compares ChildEC with two related Chinese dialogue datasets in the broader child and adolescent emotional support domain. ChildEC achieves a multi-turn interaction depth comparable to DeepWell-Adol [16], yet with more concise and developmentally appropriate language. Compared with QiaoBan [27], ChildEC distributes the emotion coaching process across more dialogue turns while maintaining concise per-turn utterances. This design aligns closely with our goal of supporting everyday emotion coaching for children.

Table 2. Corpus-level statistical comparison of QiaoBan, DeepWell-Adol, and ChildEC.

		QiaoBan	DeepWell-Adol	ChildEC (Ours)
Num. Dialogues		5100	1337	1709
Avg. Utts. per Dialogue		11.84	19.63	19.84
Avg. Len. per Dialogue		403.80	881.32	666.65
Num. Utterances		60.4K	26.3K	33.9K
Avg. Len. per Utterance		34.10	44.89	33.60
User	Avg. Len. per Dialogue	131.33	287.02	218.70
	Avg. Len. per Utterance	21.67	29.24	22.04
Assistant	Avg. Len. per Dialogue	272.47	594.30	447.95
	Avg. Len. per Utterance	47.14	60.54	45.15

6 Experiments

6.1 Dataset Splits

We divide the 1,709 dialogues in ChildEC into training, validation, and evaluation sets. To prevent data leakage while reducing distributional discrepancies across splits, we use the seed scenario as the basic unit of splitting. Specifically, the two dialogues derived from the same seed scenario are always assigned to the same subset. In addition, we perform stratified splitting over the 10 buckets defined by developmental stage × core theme, in order to preserve distributional stability across developmental stages and core themes to the greatest extent possible.

We first split out an evaluation set from the full dataset. This evaluation set is used for subsequent automatic evaluation and LLM-as-a-Judge evaluation, and it is excluded from training. Specifically, we sample 5 seed scenarios from each bucket and include their corresponding 10 dialogues in the evaluation set. Within each bucket, we also consider the diversity of emotion categories and topic anchors to enhance the quality of the evaluation set. In the end, we obtain 100 evaluation dialogues.

On this basis, the remaining 1,609 dialogues are used for supervised fine-tuning. We further divide them into a training set of 1,409 dialogues (87.57%) and a validation set of 200 dialogues (12.43%). Specifically, following the same splitting logic, we sample 10 seed scenarios from each bucket and include their corresponding 20 dialogues in the validation set. This yields 200 validation dialogues in total. The remaining 1,409 dialogues are used as the training set. In the end, we obtain mutually exclusive training,

validation, and evaluation sets. This design avoids data leakage while preserving distributional stability to the greatest extent possible.

To support the subsequent experiments, we construct three strictly nested training subsets within the training set, namely Train-25%, Train-50%, and Train-100%. Specifically, we continue to follow the same splitting logic and perform stratified sampling over the 10 buckets. The resulting subsets contain 352, 704, and 1,409 dialogues, respectively, and satisfy a strict nested relationship.

6.2 Training

We fine-tune three base models on the ChildEC corpus: Qwen2.5-7B-Instruct [17], ChatGLM3-6B [20], and InternLM2.5-7B-Chat [2]. For each model, we use LlamaFactory [28] for fine-tuning under three data scales, namely Train-25%, Train-50%, and Train-100%, while using the same validation set of 200 dialogues.

All models are trained for 3 epochs with the AdamW optimizer. The initial learning rate is set to 1.0×10^{-4}, with linear learning rate decay and a 5% warm-up ratio. The weight decay is set to 0.01. We apply Low-Rank Adaptation (LoRA) to all target linear layers, with rank = 8, α = 32, and dropout = 0.1 [7]. The maximum sequence length is set to 2048. Training is conducted with bfloat16 precision and gradient checkpointing. The per-device batch size is set to 1, and gradients are accumulated over 8 steps, resulting in an effective batch size of 8. Validation and checkpoint saving are performed every 100 steps. All fine-tuning experiments are conducted on a single NVIDIA RTX 4090 GPU.

6.3 Automatic Evaluation

To evaluate the response generation ability of LLMs in multi-turn child emotion coaching dialogues, we adopt the following automatic evaluation algorithm. Given an evaluation dialogue with m turns, we represent it as an ordered sequence of paired utterances $\{(q_i, r_i)\}_{i=1}^{m}$, where q_i denotes the child's utterance at turn i, and r_i denotes the corresponding assistant response.

For each turn, the response generated by the LLM is denoted by $\hat{r}_i$:

$$\hat{r}_i = \begin{cases} f_{\text{LLM}}(q_1), & i = 1 \\ f_{\text{LLM}}(h_i, q_i), & 1 < i \leq m \end{cases} \tag{1}$$

where $h_i = \{(q_j, r_j) \mid j = 1, \ldots, i-1\}$ denotes the dialogue history in the evaluation dialogue before turn i [25].

Each generated response $\hat{r}_i$ is then compared with the reference response r_i, and an automatic metric score is computed as

$$\hat{s}_i = \text{Metric}(\hat{r}_i, r_i). \tag{2}$$

The dialogue-level score is obtained by averaging the scores over all m turns:

$$s = \frac{1}{m}\sum_{i=1}^{m} \hat{s}_i . \tag{3}$$

Finally, we average the dialogue-level scores over all evaluation dialogues to obtain the overall automatic evaluation result on the evaluation set. To ensure a fair comparison, all models use the same system prompt during inference and the same deterministic decoding settings, with temperature set to 0.0 and repetition penalty set to 1.05.

6.4 Results and Data Scale Ablation

Using the same held-out evaluation set of 100 complete dialogues (993 turns in total), we compare the models fine-tuned under three data scales (Train-25%, Train-50%, and Train-100%) with their corresponding base models, namely Qwen2.5-7B-Instruct, ChatGLM3-6B, and InternLM2.5-7B-Chat. As shown in Table 3, all fine-tuned models outperform their base models on ROUGE [9], BLEU-4 [13], and BERTScore [26]. This result indicates improved lexical overlap and semantic similarity between the generated responses and the reference responses.

Table 3. Automatic evaluation results on the held-out evaluation set for base models and models fine-tuned under three data scales. Values in parentheses indicate relative improvement over the corresponding base model, calculated from the original unrounded scores.

Model	Condition	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4	BERTScore
Qwen2.5-7B-Instruct	Base	0.229 *(–)*	0.042 *(–)*	0.161 *(–)*	0.017 *(–)*	0.642 *(–)*
	Train-25%	0.364 *(↑58.6%)*	0.110 *(↑159.8%)*	0.295 *(↑82.7%)*	0.062 *(↑271.4%)*	0.682 *(↑6.2%)*
	Train-50%	0.382 *(↑66.7%)*	0.124 *(↑193.1%)*	0.312 *(↑93.4%)*	0.074 *(↑340.5%)*	0.692 *(↑7.7%)*
	Train-100%	**0.386** ***(↑68.1%)***	**0.128** ***(↑203.3%)***	**0.315** ***(↑95.4%)***	**0.077** ***(↑358.3%)***	**0.694** ***(↑8.1%)***
ChatGLM3-6B	Base	0.231 *(–)*	0.035 *(–)*	0.173 *(–)*	0.015 *(–)*	0.627 *(–)*
	Train-25%	0.347 *(↑50.3%)*	0.096 *(↑173.4%)*	0.276 *(↑59.8%)*	0.052 *(↑244.1%)*	0.671 *(↑7.0%)*
	Train-50%	0.358 *(↑54.9%)*	0.106 *(↑202.9%)*	0.285 *(↑65.3%)*	0.060 *(↑291.4%)*	0.678 *(↑8.1%)*
	Train-100%	**0.367** ***(↑59.1%)***	**0.115** ***(↑228.9%)***	**0.298** ***(↑72.6%)***	**0.065** ***(↑327.6%)***	**0.684** ***(↑9.0%)***
InternLM2.5-7B-Chat	Base	0.192 *(–)*	0.035 *(–)*	0.123 *(–)*	0.012 *(–)*	0.635 *(–)*
	Train-25%	0.331 *(↑72.1%)*	0.087 *(↑148.1%)*	0.263 *(↑114.5%)*	0.044 *(↑273.5%)*	0.660 *(↑4.0%)*
	Train-50%	0.346 *(↑79.9%)*	0.099 *(↑182.6%)*	0.279 *(↑127.8%)*	0.052 *(↑347.9%)*	0.672 *(↑5.8%)*
	Train-100%	**0.368** ***(↑91.2%)***	**0.112** ***(↑219.1%)***	**0.296** ***(↑141.4%)***	**0.064** ***(↑447.0%)***	**0.681** ***(↑7.3%)***

To examine the effect of training data scale on model performance, we conduct a data scale ablation study by varying the amount of training data while keeping all other experimental settings unchanged. The results show a consistent improvement trend

across the three models under the three training scales. Even the smaller training subsets of ChildEC bring clear gains. As the training data scale increases, model performance continues to improve. This suggests that the additional data provide effective training signals. Overall, these findings verify the downstream utility of ChildEC as training data for child emotion coaching.

6.5 LLM-as-a-Judge Evaluation

To comprehensively assess dialogue quality beyond automatic metrics, we conduct an LLM-as-a-Judge evaluation over the assistant responses in the same 100 held-out multi-turn dialogue instances for each of seven sources. These sources comprise the reference dialogues in the evaluation set and the Base and Train-100% outputs of three models: Qwen2.5-7B-Instruct, ChatGLM3-6B, and InternLM2.5-7B-Chat. Specifically, we use TGEC-Eval as the evaluation framework and score response quality in child emotion coaching scenarios along six dimensions. Each dimension is rated on a 1–5 scale. All evaluations are conducted with Qwen-plus using a unified standardized scoring prompt to ensure consistency and reproducibility. The prompt is summarized in Table 5.

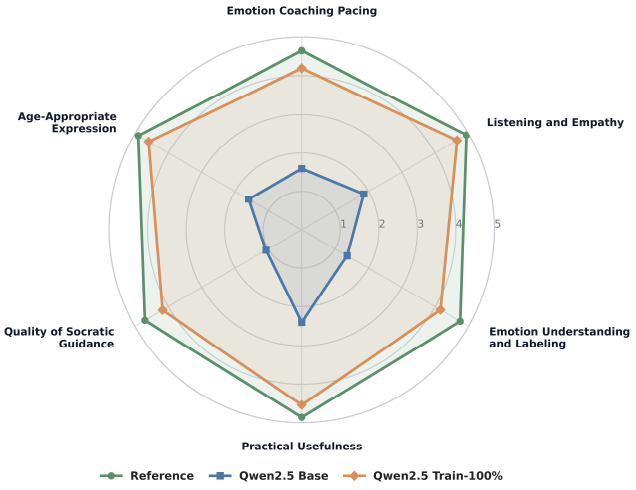

Fig. 4. LLM-as-a-Judge scores for the reference dialogues, Qwen2.5-7B-Instruct Base, and Qwen2.5-7B-Instruct Train-100% on the six TGEC-Eval dimensions.

The results in Fig. 4 and Table 4 show that, although all three fine-tuned models remain below the reference dialogues across all six dimensions, they consistently outperform their respective base models. This result shows that ChildEC can effectively improve model performance in child emotion coaching.

Table 4. LLM-as-a-Judge scores for seven sources across the six TGEC-Eval dimensions: Pacing (Emotion Coaching Pacing), Emp. (Listening and Empathy), Labeling (Emotion Understanding and Labeling), Pract. (Practical Usefulness), Socratic (Quality of Socratic Guidance), and Age App. (Age-Appropriate Expression).

Source	Condition	Pacing	Emp.	Labeling	Pract.	Socratic	Age App.	Avg.
Reference Dialogues	Reference	4.64	4.94	4.75	4.86	4.70	4.90	4.80
Qwen2.5-7B-Instruct	Base	1.59	1.86	1.36	2.40	1.07	1.59	1.65
	Train-100%	4.19	4.66	4.16	4.54	4.17	4.59	4.39
ChatGLM3-6B	Base	1.21	1.41	1.22	2.14	1.05	1.28	1.39
	Train-100%	3.87	4.44	3.97	4.37	3.63	4.32	4.10
InternLM2.5-7B-Chat	Base	1.52	1.75	1.30	2.21	1.01	1.32	1.52
	Train-100%	3.75	4.25	3.78	4.09	3.53	4.06	3.91

7 Conclusion

In this paper, we construct ChildEC, a developmentally stratified multi-turn dialogue corpus for everyday child emotion coaching targeting Chinese children aged 8–12, and propose a Theory-Grounded Emotion Coaching Framework (TGEC) to guide its construction and evaluation. Experimental results show that ChildEC not only provides a specialized data resource for research on child emotion coaching in Chinese, but also effectively supports the training and evaluation of dialogue models and improves their generation quality on child emotion coaching tasks. We hope that ChildEC will facilitate future research on child-facing emotion coaching dialogue systems that are developmentally sensitive and theoretically grounded.

8 Limitations

Although this paper validates the effectiveness of the proposed framework for dialogue generation and evaluation in child emotion coaching, several limitations remain. First, ChildEC is primarily constructed through a theory-guided multi-stage dialogue synthesis pipeline, and thus its dialogue distribution may still deviate from that of real-world child emotion coaching interactions. Second, the corpus does not directly cover a broader range of child support tasks and is not intended for clinical psychological support settings. Furthermore, the current fine-tuning experiments only use small open-source models with around 7B parameters, and the applicability of ChildEC to larger-scale models has not yet been examined. Future work may incorporate systematic human expert evaluation and expand the corpus and model evaluation scope to cover more diverse populations, scenarios, and model scales, thereby further improving its reliability and practical reference value in real-world applications.

9 Ethics Statement

We are keenly aware of the ethical responsibilities inherent in the sensitive research setting of child emotion coaching. ChildEC is constructed through a multi-stage

dialogue synthesis pipeline, with review and filtering procedures applied to control content quality and safety boundaries. Nevertheless, dialogue models trained on this corpus may still produce inappropriate responses and should therefore not be regarded as substitutes for professional psychological services or medical advice. In particular, such systems should be deployed with caution and under necessary human supervision in settings involving minors.

Acknowledgments. This study was supported by the National Undergraduate Innovation and Entrepreneurship Training Program, Project No. 202510057049.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

A Prompt Resources

The full prompts used in TGEC-Synth and TGEC-Eval have been released alongside the ChildEC dataset in the project repository. For brevity, this appendix summarizes each prompt in terms of its corresponding stage, model, main inputs, and output.

Table 5. Summary of the released prompt resources used in TGEC-Synth and TGEC-Eval.

Stage	Model	Main inputs	Output
Topic anchor generation	DeepSeek-V3.2-chat	age group, core theme	topic anchors
Seed scenario generation	DeepSeek-V3.2-chat	age group, core theme, topic anchors	seed scenarios
Seed review	Qwen-plus	seed scenarios	seed quality judgments
Dialogue plan generation	DeepSeek-V3.2-chat	seed scenario, age group, core theme, topic anchor	dialogue plan
Dialogue generation	DeepSeek-V3.2-chat	dialogue plan, seed scenario, age group	multi-turn dialogue
Dialogue review	Qwen-plus	generated dialogue, seed scenario, age group	dialogue quality judgments
TGEC-Eval scoring	Qwen-plus	complete dialogue, seed scenario, age group	six-dimensional scores

References

1. Al Khansa, H., Mustapha, A., Awad, M.: Sahar dataset: A validated dialogue based dataset for a child-centric, empathetic and knowledge-driven chatbot. In: Proceedings of the AAAI Symposium Series. vol. 6, pp. 159–166 (2025)
2. Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., et al.: InternLM2 technical report. arXiv preprint arXiv:2403.17297 (2024)
3. Carona, C., Handford, C., Fonseca, A.: Socratic questioning put into clinical practice. BJPsych Advances **27**(6), 424–426 (2021)
4. Chen, K., Sun, Z.: DeepPsy-Agent: A stage-aware and deep-thinking emotional support agent system. arXiv preprint arXiv:2503.15876 (2025)

5. DeepSeek-AI: DeepSeek-V3.2: Pushing the frontier of open large language models. arXiv preprint arXiv:2512.02556 (2025)
6. Gottman, J.M., Katz, L.F., Hooven, C.: Parental meta-emotion philosophy and the emotional life of families: Theoretical models and preliminary data. Journal of Family Psychology **10**(3), 243–268 (1996)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
8. Jiao, J., Afroogh, S., Chen, K., Murali, A., Atkinson, D., Dhurandhar, A.: Safe-Child-LLM: A developmental benchmark for evaluating LLM safety in child-LLM interactions. arXiv preprint arXiv:2506.13510 (2025)
9. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004)
10. Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., Huang, M.: Towards emotional support dialog systems. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 3469–3483. Association for Computational Linguistics, Online (2021)
11. Markus, H.J., Nurius, P.S.: Self-understanding and self-regulation in middle childhood. In: Collins, W.A. (ed.) Development During Middle Childhood: The Years From Six to Twelve. National Academies Press, Washington, DC (1984)
12. Mei, Y., Wang, X., Zhang, Z., Fu, Y., Wang, S., Wan, Q., Lan, Q., Liu, C., Cai, J., Yu, C., Shi, Y.: PACEE: Parent-centered AI scaffolding for emotion education in early childhood conversations. arXiv preprint arXiv:2511.14414 (2025)
13. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, USA (2002)
14. Piaget, J., Inhelder, B.: The Psychology of the Child. Basic Books, New York (1972)
15. Qiu, H., He, H., Zhang, S., Li, A., Lan, Z.: SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 615–636. Association for Computational Linguistics, Miami, Florida, USA (2024)
16. Qiu, W., Wang, Y., Tan, J., Hou, H., Liu, Q., Yao, W., Ni, S.: DeepWell-Adol: A scalable expert-based dialogue corpus for adolescent positive mental health and wellbeing promotion. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 12786–12810. Association for Computational Linguistics, Suzhou, China (2025)
17. Qwen Team: Qwen2.5: a party of foundation models. Blog (2024). https://qwen.ai/blog?id=qwen2.5
18. Seo, W., Yang, C., Kim, Y.H.: ChaCha: Leveraging large language models to prompt children to share their emotions about personal events. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery (2024)
19. Shi, Y., Zhang, L., Li, M., Zheng, Y., Wang, X., Kong, F.: Simulating human-like counseling: A path- and scenario-guided framework for psychological support dialogue. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 17688–17696 (2026)

20. Team GLM, Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., et al.: ChatGLM: A family of large language models from glm-130b to GLM-4 all tools. arXiv preprint arXiv:2406.12793 (2024)
21. Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., Nie, J.Y.: C-Pack: Packed resources for general chinese embeddings. arXiv preprint arXiv:2309.07597 (2023)
22. Xu, Z., Jiang, J.: Multi-dimensional evaluation of empathetic dialogue responses. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 2066–2087. Association for Computational Linguistics, Miami, Florida, USA (2024)
23. Xue, D., Tu, J., Wang, M., Yan, X., Liu, F., Hu, J.: Towards privacy-preserving mental health support with large language models. arXiv preprint arXiv:2601.01993 (2026)
24. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
25. Zhang, C., Li, R., Tan, M., Yang, M., Zhu, J., Yang, D., Zhao, J., Ye, G., Li, C., Hu, X.: CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 13947–13966. Association for Computational Linguistics, Bangkok, Thailand (2024)
26. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations (2020)
27. Zhao, W., Wang, S., Tong, Y., Lu, X., Li, Z., Zhao, Y., Wang, C., Zheng, T., Qin, B.: QiaoBan: A parental emotion coaching dialogue assistant for better parent-child interaction. GitHub repository (2023). https://github.com/HIT-SCIR-SC/QiaoBan
28. Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z.: LlamaFactory: Unified efficient fine-tuning of 100+ language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). pp. 400–410. Association for Computational Linguistics, Bangkok, Thailand (2024)