The Resonance Corpus: Chinese Caregiver-Child Dialogue for Community-Aligned Language Models

Anonymous submission

Abstract

We introduce the Resonance Corpus, a large-scale collection of natural Chinese caregiver–child conversations designed as infrastructure for aligning language models to underserved communities. The corpus captures everyday, intergenerational talk around child-friendly news prompts and includes rich contextual and cognitive information. We use this resource to argue for a research agenda that treats family dialogue as a key testbed for culturally grounded and developmentally appropriate AI. In particular, we outline how the corpus supports three strands of work: participatory alignment with community-contributed data, lightweight instruction tuning of Chinese LLMs under realistic computational budgets, and evaluation protocols that focus on cognitive fit and cultural robustness, rather than solely on generic benchmark scores. By framing Chinese caregiver–child dialogue as a core low-resource setting, we aim to provide open infrastructure for building language technologies that communicate more effectively with children and caregivers in underrepresented linguistic and cultural contexts.

Introduction

Recent advances in large language models (LLMs) have yielded impressive results on standard NLP benchmarks, yet fundamental alignment challenges remain in multi-turn dialogues (Wang, Morgenstern, and Dickerson 2025; Gao et al. 2025). For example, LLMs often lose coherence or diversity during extended conversations – a phenomenon of dialogic "collapse" (Xiao et al. 2023; Zhou et al. 2024; Wu and Papyan 2024). Even instruction-tuned models that undergo reinforcement learning from human feedback exhibit trade-offs: they may refuse user requests less often but instead produce more plausible-sounding factual errors as task complexity increases (Fanous et al. 2025; Liu et al. 2024). These inconsistencies show that current models still struggle to align with user intent and adapt in interactive contexts.

Addressing alignment for diverse users requires grounding models in natural conversational data. Human conversation is interactive and contextual, rich in pragmatic cues such as turn-taking, backchannels, conversational strategies, and interruptions. Prior works have begun to capture these dynamics. Researchers at the University of Pennsylvania recently released one of the first large-scale corpora of naturalistic human conversations (the CANDOR corpus), comprising over a thousand recorded English dialogues (Reece et al. 2023). Weng et al. applied 3D vision analysis to in-home video recordings of caregiver-child interactions, revealing developmental patterns in children's engagement and turn-taking (Weng et al. 2025). However, the field still lacks an open-source, large-scale Chinese caregiver-child corpus, which hinders cross-cultural comparison.

Yet most alignment datasets focus on high-resource languages, adult speakers, and formal registers, leaving child-directed Chinese and dialect-rich family talk largely underserved. To fill this gap, we introduce the Resonance Corpus, a large-scale collection of natural Chinese caregiver–child conversations. We build the corpus through a participatory data collection process: thousands of families across China record weekly at-home discussions prompted by child-friendly news topics and submit transcripts via a WeChat mini-program. This procedure captures genuine, unconstrained dialogues in everyday settings and reflects a wide range of socio-cultural backgrounds, including some dialectal variation. At the current stage, the corpus contains nearly 60,000 multi-turn dialogues from roughly 3,000 families, making it a large-scale conversational resource and, to our knowledge, one of the largest available corpora of Chinese caregiver–child dialogue. Each dialogue entry includes detailed metadata, and we de-identify all records to protect participant privacy. We also completed extensive human annotation on a subset of dialogues. In this paper, we (i) describe the corpus, and (ii) show how it supports LM alignment in underserved Chinese settings.

Corpus Construction

Data Collection

Each week, we send participating families a set of four brief news story prompts through WeChat, as shown in Figure 1. These prompts (e.g., a science discovery, a cultural event, etc.) serve as conversation starters for caregivers and children. If a topic doesn't work for them, caregivers can request an alternative to engage with content that feels relevant. During the week, families independently pick a convenient time at home to discuss one or more of the prompts in an open-ended conversation. Caregivers and children typically speak Mandarin Chinese (sometimes with regional phrases or accents), and they record their conversations in natural settings using a phone or audio recorder. After they finish talking, the caregiver transcribes the dialogue in our WeChat mini-program and submits both the transcript and contextual information. We then store the transcripts in a central database for further analysis and review. Because we use current news as a stimulus, the conversations stay grounded in shared, timely content rather than scripted questions. This design makes the data particularly suitable for studying how humans and LLMs handle continuously updated knowledge in conversation.

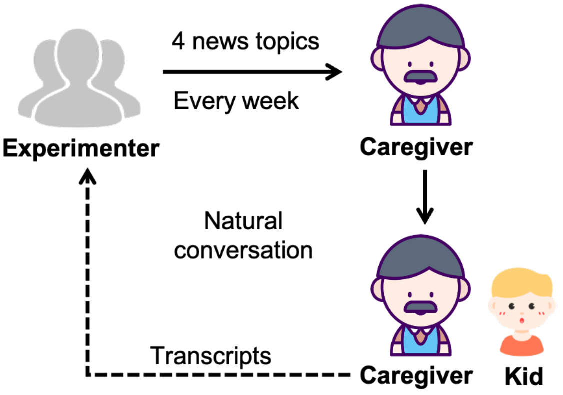

Figure 1: Corpus collection process.

The data collection began in 2023 and is ongoing. As of the current release, the Resonance Corpus contains approximately 60,000 caregiver-child conversations contributed by about 3,000 distinct families. The dialogues range from a few turns to more than 50 utterances, with an average duration of roughly 5–12 minutes. The participating children span a wide age range (most are in early childhood to early adolescence), and families come from diverse regions and socio-economic backgrounds across China. This diversity introduces a rich mix of linguistic variety – while all transcripts use standard Chinese characters, many include colloquial expressions, and some reflect regional vocabulary or the influence of local dialects. Such breadth is particularly valuable for training and testing models on dialect robustness and informal speech. Table 1 summarizes the key data fields recorded for each conversation.

Annotation Scheme

In addition to raw transcripts, the Resonance Corpus provides a hand-annotated subset with utterance-level labels that capture cognitive and pragmatic dimensions of the dialogue. We recruited and trained a team of coders (primarily psychology and education students) to label two months of conversation data, totaling 6,840 dialogues (nearly 90,000 utterances). We coded each utterance in these dialogues along three axes:

(1) Discourse Function: The speech act or conversational move, such as a question, an answer, a command, feedback, or statement. This helps identify how each turn functions in the dialogue, for instance, whether a caregiver's utterance is prompting or affirming the child's response.

Table 1: Sample from the corpus.

Field	**Sample**
New topic	According to a report by the BBC(2023-10-10), the University of Exeter in the United Kingdom has announced that it will introduce a new Master's degree programme in Magic Studies in 2024, established in response to the "recent surge of public interest in magic." It is noteworthy that J. K. Rowling, author of the Harry Potter series, is a graduate of the University of Exeter.
caregiver's name	Wang (surname only, for privacy)
caregiver's age	39 years old
caregiver's gender	Female (mother)
caregiver's emotion	calm
child's name	Du (surname only, for privacy)
child's age	10 years old
child's gender	Female
child's emotion	Really looking forward to it
Date	2023-10-12
Duration	8 min
Scenario	After dinner
Rating	2 (1–5, assigned retrospectively by the caregiver)
Post-conversation reflection	The caregiver said, "During the conversation, the child mentioned asking the director to teach the students magic, but I didn't understand what they meant at the time. When I was organizing the transcript, I realized I should have given the child more time to express themself."
Commentary on the news topic	Very eye-catching news material; it's suitable for exploring the human mind, and I hope to see it incorporate Eastern philosophy.
Conversation content	712 characters (punctuation and speaker labels removed); 31 utterances (caregiver: 17, child: 14); Child output ratio: 30.01%.

(2) Abstraction Level: The level of abstraction or cognitive demand of the utterance, categorized using Marion Blank's four levels (L1–L4) from developmental psychology (Blank 1974; Blank, Rose, and Berlin 1978). An L1 utterance involves concrete labeling or description of objects and events (e.g. "This is a cat."); L2 involves short-term retelling or recall (e.g. "What did we see at the zoo today?"); L3 performs an internal reorganization or reconstruction of perceptual information (e.g. "From these few things, we can see that he is a brave person."); and L4 involves reasoning or prediction (e.g. "How would you feel if you had a cat that could talk?"). By coding the level of each conversational turn, we can examine how caregivers calibrate the complexity of their language to the developmental stage of children, as well as how children's own contributions become increasingly abstract with age.

(3) Analogical Reasoning: A binary label marking whether the utterance contains an analogy or metaphor (broadly defined as mapping one situation or concept onto another), which follows the structure-mapping theory (Gentner 1983, 2016). For example, a caregiver saying "Our brain is like a computer" would be tagged as an analogy. Such instances are relatively rare but indicate higher-order reasoning and teaching moments in conversations.

All annotators underwent extensive training to ensure reliable and objective coding. We recruited a pool of coders and conducted a two-round training process. In Round 1, candidates learned the coding scheme (through examples and a short practice test). Those who advanced received a comprehensive Language Coding Manual with guidelines and examples for each label dimension – including utterance-level coding standards (informed by the University of Chicago's Language Development Project) as well as detailed definitions for each cognitive level (Blank's L1–L4) and analogical reasoning cues. In Round 2, candidates coded sample dialogues both individually and in teams, refining coding consistency through discussion and achieving a high level of within-team agreement (targeting more than 80% agreement on labels).

For the actual corpus annotation, we organized coders into small teams and assigned each team batches of dialogues. We implemented a rigorous quality-control pipeline: teams first coded a common subset independently, then reconciled disagreements through discussion until they exceeded 80% consensus. Independent auditors—researchers who worked outside the coding teams and remained blind to the study's hypotheses—conducted stratified spot checks on roughly 20% of the utterances. Whenever an audit showed that a team's coding consistency in any dialogue fell below the 80% threshold, a cross-team panel re-audited and adjudicated the entire dialogue. This multi-layered protocol yielded highly reliable labels. The resulting annotated subset forms a rich, structured dataset of caregiver-child interactions, with over 90,000 utterances coded for conversational role, cognitive complexity, and reasoning. We can use these annotations to derive quantitative metrics (e.g., distributions of cognitive-level usage, frequency of analogies, turn-taking patterns).

Alignment Applications

The Resonance Corpus creates several avenues to advance and assess language model alignment in low-resource and culturally diverse settings. We focus on three central application areas.

Participatory Alignment with Community Data

The corpus arises from a participatory process, with thousands of families actively contributing their conversations. Researchers can extend this participatory ethos into model development by fine-tuning or conditioning LLMs on data drawn directly from the target community, thereby aligning models with the community's linguistic styles, values, and needs. For example, our dataset documents how Chinese caregivers explain news and abstract concepts to children, using culturally grounded examples and modulating their tone. An LLM trained on these patterns can learn to adjust its register for users of different ages (e.g., simplifying explanations for children) and to respect local communication norms, such as politeness conventions and storytelling practices. Because the data originates from real users, the resulting alignment operates in a bottom-up and inclusive manner, reflecting community preferences rather than only designer assumptions.

Future work can place caregivers and educators explicitly in the loop, turning corpus contributors into stakeholders who iteratively shape model behavior—a form of human-in-the-loop participatory alignment. This approach becomes important for underserved languages and dialects, where input from local speakers plays a crucial role in steering models toward contextually accurate and culturally appropriate outputs.

Lightweight Instruction Fine-Tuning

Large-scale LLMs with tens of billions of parameters can adapt to new tasks or domains through efficient fine-tuning methods that use relatively modest datasets and compute. With tens of thousands of conversations, the Resonance Corpus provides an ideal, high-quality resource for instruction tuning in multi-turn, intergenerational dialogue. We propose applying QLoRA (Quantized Low-Rank Adaptation)—a recent method for memory-efficient fine-tuning of LLMs—to adapt existing Chinese language models on this corpus (Dettmers et al. 2023).

By training a base model on Resonance dialogues, using conversation prompts and responses as supervised signals, we enable the model to approximate the interactive behavior of caregivers and children. Concretely, we can fine-tune models such as LLaMA-2 or Chinese-BERT via QLoRA to take a news prompt and generate a multi-turn discussion between a caregiver and a child, or to continue an ongoing conversation in the style of the corpus. This low-cost fine-tuning produces specialized models that better align with family-style conversational dynamics: they can learn to ask follow-up questions, encourage the user (as a caregiver does), and avoid the common LLM failure mode of long, overly formal monologues. Because QLoRA maintains model quality with minimal computational overhead, this strategy accords with principles of sustainable compute. It enables researchers in low-resource settings, who often have limited hardware, to build community-specific chatbot models without training from scratch. We envision such fine-tuned models serving as educational assistants or tools for caregiver training, where model responses display the warmth and adaptivity characteristic of real caregiver-child interaction.

Culturally Grounded Evaluation of LLMs

Standard benchmarks for dialogue agents often rely on generic metrics, such as coherence or factual accuracy on Wikipedia-style questions, but these metrics say little about cultural or cognitive alignment. The Resonance Corpus supports new evaluation frameworks that draw directly on real human–human interaction. One immediate use is to treat the annotated subset as a benchmark for alignment. We can set up tasks in which an LLM plays the role of the caregiver or the child in an actual dialogue from the corpus and continues the conversation; human evaluators (or automatic metrics) then judge how well the model's turns match the expected behavior.

Because the data includes cognitive-level labels, we can define a Cognitive Alignment Score. This score asks whether the model's response stays within an appropriate complexity range, given the child's last utterance and age. A well-aligned model, for instance, may rephrase a complex question into an L1 or L2 form when it "notices" that the child is younger or seems confused. We can quantify this adaptivity by comparing model utterances with human transcripts and labels. The analogical reasoning tags provide another evaluation channel: they let us test whether models use analogy in explanations as effectively as human caregivers do, which is crucial for teaching new concepts.

The corpus also enables tests of dialectal and cultural robustness. We can sample region-specific expressions or idioms from the transcripts and construct prompts that check whether a given LLM understands and answers appropriately in the presence of dialectal variation. For example, if a child uses a local phrase for a common food or game, we can ask whether the model interprets it correctly or fails. Zero-shot and few-shot prompting both fit this setting: we either expose the model directly to dialectal terms or first provide a few contextual examples and then examine whether it can sustain a natural conversation. Success in these tasks signals a model that respects linguistic diversity. At the same time, systematic errors—such as repeated misinterpretation of a particular dialect phrase—reveal concrete gaps that further fine-tuning or data augmentation can target.

More broadly, we can design interactive tasks based on Resonance, such as collaborative storytelling with a child or answering a child's "why" questions. These tasks embed cultural and developmental context into alignment evaluation. They move us toward holistic benchmarks that value not only task completion but also how the AI communicates: whether it speaks in a respectful, engaging, and understandable way for users in their cultural context. The Resonance Corpus offers rich, realistic scenarios that make such assessments possible.

Conclusion

We have introduced the Resonance Corpus and outlined its construction and alignment with key goals for LLMs. This corpus includes roughly 60,000 dialogues and fills a critical gap in available data: it offers a lens into intergenerational communication that traditional NLP corpora rarely capture. Our case study focuses on Chinese families, but the principles and methodologies extend to any community seeking to align technology with their linguistic and cultural context.

Moving forward, we are committed to releasing the Resonance Corpus (with robust privacy safeguards) and accompanying it with documentation and tools to lower the barrier for its use. We encourage researchers in natural language processing, education, and cognitive science to leverage this corpus in developing next-generation dialogue models. Such applications embody the vision of language technology that empowers underserved communities – by being accessible, adaptive, and aligned with the people it serves. We believe that bridging real-world conversational data with modern AI techniques is not only scientifically fruitful, uncovering new challenges and evaluation metrics for LLMs, but also socially beneficial. It paves the way for AI that respects linguistic diversity, engages users on their own terms, and supports human communication rather than undermining it.

References

Blank, M. 1974. Cognitive functions of language in the preschool years. *Developmental Psychology*, 10(2): 229.

Blank, M.; Rose, S. A.; and Berlin, L. J. 1978. The language of learning: The preschool years. *(No Title)*.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36: 10088–10115.

Fanous, A.; Goldberg, J.; Agarwal, A.; Lin, J.; Zhou, A.; Xu, S.; Bikia, V.; Daneshjou, R.; and Koyejo, S. 2025. Syceval: Evaluating llm sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 893–900.

Gao, Y.; Lee, D.; Burtch, G.; and Fazelpour, S. 2025. Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24): e2501660122.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2): 155–170.

Gentner, D. 2016. Language as cognitive tool kit: How language supports relational thought. *American psychologist*, 71(8): 650.

Liu, R.; Sumers, T. R.; Dasgupta, I.; and Griffiths, T. L. 2024. How do large language models navigate conflicts between honesty and helpfulness? *arXiv preprint arXiv:2402.07282*.

Reece, A.; Cooney, G.; Bull, P.; Chung, C.; Dawson, B.; Fitzpatrick, C.; Glazer, T.; Knox, D.; Liebscher, A.; and Marin, S. 2023. The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13): eadf3197.

Wang, A.; Morgenstern, J.; and Dickerson, J. P. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 1–12.

Weng, Z.; Bravo-Sánchez, L.; Wang, Z.; Howard, C.; Xenochristou, M.; Meister, N.; Kanazawa, A.; Milstein, A.; Bergelson, E.; Humphreys, K. L.; et al. 2025. Artificial intelligence–powered 3D analysis of video-based caregiver-child interactions. *Science Advances*, 11(8): eadp4422.

Wu, R.; and Papyan, V. 2024. Linguistic collapse: Neural collapse in (large) language models. *Advances in Neural Information Processing Systems*, 37: 137432–137473.

Xiao, G.; Tian, Y.; Chen, B.; Han, S.; and Lewis, M. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Zhou, L.; Schellaert, W.; Martínez-Plumed, F.; Moros-Daval, Y.; Ferri, C.; and Hernández-Orallo, J. 2024. Larger and more instructable language models become less reliable. *Nature*, 634(8032): 61–68.