Cognitive Inference as the Main Predictor of AI Reliability in Automated Behavioral Coding of Parent–Child Interactions

José Amorocho[1*], Ana Balsa[2†], Juan José Giraldo-Huertas[1†], Juanita Bloomfield[2†], Paula Patrone[2†], Alejandro Cid[2†]

[1*]Facultad de Ciencias del Comportamiento, Universidad de La Sabana, Campus del Puente del Común, Km 7, Autopista Norte de Bogotá, Chía, 250001, Cundinamarca, Colombia.

[2]Department of Economics, Universidad de Montevideo, Saldún de Rodriguez 2097, Montevideo, 11500, Uruguay.

*Corresponding author(s). E-mail(s): joseammo@unisabana.edu.co;
Contributing authors: abalsa@um.edu.uy; juangh@unisabana.edu.co; j.bloomfield@um.edu.uy; ppatrone@um.edu.uy; acid@um.edu.uy;
†These authors contributed equally to this work.

Abstract

Observational coding of parent–child interactions is a gold standard in developmental science but remains unscalable. Multimodal generative AI could help, yet its reliability and failure modes are not well characterized. We benchmarked a multi-agent pipeline (GABRIEL) against a multi-rater expert consensus when scoring 22 PICCOLO items on 156 ten-minute free-play interactions from Uruguay. Agreement was summarized with Percent Agreement (PA) and Cohen's $\boldsymbol{\kappa}$, and disagreement with a unit-free normalized mean squared error (nMSE = MSE/Var$(\boldsymbol{Y}_{\text{item}})$). A priori item classes indexed *cognitive inference* (Low/Medium/High). Final calibration yielded modest agreement (PA = **50.7%**, $\boldsymbol{\kappa}$ = **.216**). Disagreement was chiefly structured by inference (Kruskal–Wallis $\boldsymbol{H}$ = **308.70**, $p<.001$), a pattern that persisted in late iterations. The model also overused the middle category (1) and underused "2." No systematic differences in nMSE emerged by sex, age quartile, or maternal education. We conclude that generative AI is promising for scalable detection of concrete, low-inference behaviors, whereas high-inference judgments still require expert adjudication. A human-in-the-loop, *co-intelligence* workflow aligns current strengths with ethical oversight and supports equitable deployment at scale.

Keywords: Generative AI, Parent-Child Interaction, Observational Coding, Inter-Rater Reliability, PICCOLO, Multimodal AI

1 Introduction

The quality of interactions between a parent and child during the first five years of life is a cornerstone of developmental science, establishing the very foundation for a child's future cognitive, social, and emotional outcomes (Jeong, Franchett, Ramos de Oliveira, Rehmani, & Yousafzai, 2021; Vilaseca et al., 2019). Central to these formative relational experiences is the concept of scaffolding, a process in which a caregiver provides contingent and interactive support that enables a child to operate successfully within their Zone of Proximal Development (ZPD), the critical space that exists between independent achievement and guided success (Spadafora & Downes, 2020; Vygotsky, 1978; Wood, Bruner, & Ross, 1976; Xi & Lantolf, 2021). This perspective, rooted in sociocultural theory, posits that learning is fundamentally a social process, where the role of parents as a child's first and most influential educators is paramount (Landry, Smith, & Swank, 2008; Morris & Smith, 2022).

The ability to accurately capture and understand these complex, dynamic, and often fleeting behaviors as they unfold in their natural context is essential for researchers to advance theory and for clinicians to identify opportunities for early intervention that can effectively support parental practices (Huang, Sun, Lau, & Zhou, 2023; Ismai et al., 2024; Myruski & Dennis-Tiwary, 2022). Specifically, domains such as Affection and Responsiveness are foundational for secure attachment and socio-emotional regulation, while Encouragement and Teaching directly scaffold the cognitive and linguistic skills crucial for academic readiness (Roggman et al., 2013; Vilaseca et al., 2019). However, the endeavor of translating the rich, nuanced tapestry of these real-world interactions into structured, reliable data presents a formidable and long-standing methodological challenge (Horowitz-Kraus & Gashri, 2023)For decades, observational methods have been rightfully considered the gold standard in developmental research, offering an unparalleled and direct window into the authentic dynamics of family life (Aprilia & Neisya, 2024; Fitria & Nurcholis, 2025). Despite their theoretical preeminence, these methods are fraught with significant practical limitations that curtail their widespread use. The process of human coding, which involves trained individuals watching recorded interactions and systematically scoring them against a predefined rubric, is notoriously labor-intensive, time-consuming, and costly. These constraints severely restrict the scalability of observational research and impede its application in the kind of large-scale developmental screening programs that are urgently needed, particularly in low- and middle-income countries where access to specialized evaluators and resources is often scarce (Cuartas, Rey-Guerra, McCoy, & Hanno, 2020; Giraldo-Huertas, 2023; Rey-Guerra, Maldonado-Carreño, Ponguta, Nieto, & Yoshikawa, 2022; Rubio-Codina & Grantham-McGregor, 2020; Ward, 2022). This fundamental problem of scale creates a critical and persistent gap between our sophisticated scientific understanding of early development and our collective ability

to deliver timely, evidence-based support to the vast number of families who could benefit from it (Desimone, 2009; Kaminski, Valle, Filene, & Boyle, 2008).

The recent advent of generative artificial intelligence (AI), particularly multimodal large language models, presents a transformative opportunity to overcome this scalability challenge (Chen et al., 2024; Mollick, 2024). By leveraging multimodal models trained on culturally diverse datasets and integrating signals such as dyadic synchrony, affective tone, and linguistic features, such systems can generate first-pass behavioral profiles in minutes rather than weeks. This capability supports the early detection of developmental vulnerabilities and the efficient triage of cases for targeted, expert-led follow-up, while optimizing the allocation of scarce clinical resources, particularly in LMICs where developmental risks are elevated but specialist availability remains critically constrained. Crucially, this approach is not intended to replace expert interpretation but to expand equitable access to evidence-based screening by automating high-volume, low-inference assessments with demonstrated reliability. In doing so, it establishes a scalable, equity-driven foundation for adaptive intervention pipelines that can bridge the gap between advanced developmental science and real-world service delivery in under-resourced settings.

In this paper, we present a rigorous benchmark of a multimodal generative AI pipeline (GABRIEL), designed to identify patterns of interactions that reliably screen developmental strengths and risks in a recorded play activity between mothers and children. The pipeline is prompted to assess the quality of the interactions using the PICCOLO scale (Parenting Interactions with Children: Checklist of Observations Linked to Outcomes) (Roggman, 2009), an instrument that measures positive parenting behaviors as parents interact with their children. Based on prior human-based coding of the same video-recordings with the PICCOLO scale, we use several measures of inter-rater reliability to assess the quality of the AI-based prediction and, more critically, explain when and why machine–expert alignment strengthens or degrades. Our objectives were: (a) to document iterative optimization of the pipeline; (b) to compare the final AI performance against expert consensus with Percent of Agreement (PA), Kappa (K), and unit-free Mean Square Error (nMSE); and (c) to test whether disagreement is organized by a priori inference levels and is not explained by observed demographics, as further illustrated in the comprehensive pipeline overview in Figure 1. We hypothesized that overall alignment would be *systematically* higher for low-inference items and lower for high-inference items, reflecting the cognitive demands embedded in the instrument.

A fundamental conceptual challenge in this entire endeavor lies in translating complex, nuanced psychological theory into a format that a machine can parse and process, a problem that lies at the intersection of psychology and computer science (Al-Gharib & Gouda, 2024; Boden, 1988; Carroll, 1997; Hu & Twidale, 2023; Tejeda, Kumar, Smyth, & Steyvers, 2022). This study tackles this challenge directly by systematically operationalizing the PICCOLO's theoretical constructs for the AI pipeline. This is not simply a matter of data input; it is a meticulous procedure that is highly analogous to the methods used to train a human rater, embodying principles of instructional design and scaffolding for the machine itself (Lee, 2022; Mindigulova, Vikhman, & Romm, 2024; Winkler, Hobert, Salovaara, Söllner, & Leimeister, 2020). It is achieved through

two primary and complementary techniques rooted in the architecture of modern foundation models like transformers (Tunstall, von Werra, & Wolf, 2022; Vaswani et al., 2017). The first is prompt engineering, a discipline wherein the abstract definitions, decision rules, and scoring criteria from the PICCOLO manual are carefully translated into a set of detailed, explicit, and logically structured instructions for the AI. The second technique is few-shot learning, a process through which the AI is provided with a curated library of concrete, pre-scored examples of behavior. These annotated examples serve as critical precedents or case studies for the AI, grounding the abstract rules in real-world scenarios. This allows the model to learn how to apply the criteria for highly interpretive anchors such as "Somewhat" or "Consistently" in a manner that closely aligns with the sophisticated judgment of human experts, moving the technology toward a form of computational intelligence (Brock, 2024; Edwards, 2024; Strachan et al., 2024).

Our endeavor is built upon a robust foundation of prior work in computationally intensive fields like psychotherapy and counseling. For over a decade, researchers have successfully used statistical text classification and natural language processing to automate the coding of complex clinical behaviors, such as assessing therapeutic fidelity in motivational interviewing, with performance metrics approaching those of human experts (Atkins, Steyvers, Imel, & Smyth, 2014; Flemotomos et al., 2022; Tanana, Hallgren, Imel, Atkins, & Srikumar, 2016). These studies provided a crucial proof-of-concept, demonstrating that machine learning models could move beyond simple tasks and reliably quantify nuanced, theory-driven human interactions from raw data (Idalski Carcone et al., 2019; Pérez-Rosas, Mihalcea, & Resnicow, 2017). This established success has paved the way for applying similar—and now far more powerful—technologies to the equally complex domain of early childhood developmental science.

Our paper is among a growing body of work that has begun to target the unique complexities of parent-child interactions using multimodal AI. Researchers have successfully prototyped pipelines that quantify interaction quality from raw audio and video signals in home environments (Jebeli et al., 2024), automatically classify specific observational codes from dyadic dialogue (Lin, Huang, & Lu, 2023), and extract objective motion features that correlate significantly with expert judgments of interactional synchrony (Egmose et al., 2017; Leclère et al., 2016). While these pioneering studies validate the promise of automated analysis, they also highlight a critical gap. Many approaches focus on extracting isolated behavioral signals rather than scoring a holistic, validated clinical instrument; furthermore, they rarely employ the advanced inferential capabilities of the latest generation of multimodal foundation models. Our work is unique in attempting to conduct a rigorous, transparent evaluation of a state-of-the-art generative AI pipeline against a robust, multi-rater expert consensus on a widely used instrument like PICCOLO. We believe studies as ours are a necessary next step to truly benchmark the technology's readiness for real-world application.

At the heart of this principle lies inter-rater reliability (IRR), the statistical measure of consistency, or agreement, of scores among different independent raters (Goodwin, 2001). It is crucial to distinguish between inter-rater agreement, which

refers to the extent to which raters assign identical scores, and inter-rater reliability, which reflects the consistency in how raters differentiate among subjects, though the former is often used as a practical proxy for the latter (Vanbelle, Engelhart, & Blix, 2024). Without the assurance of robust IRR, it becomes impossible to ascertain whether observed differences in scores reflect true variations in behavior or are merely artifacts of noise, bias, and inconsistency inherent in the measurement process itself. For an assessment tool to be considered a valid and trustworthy instrument, different raters must be able to use it to produce the same results under the same conditions.

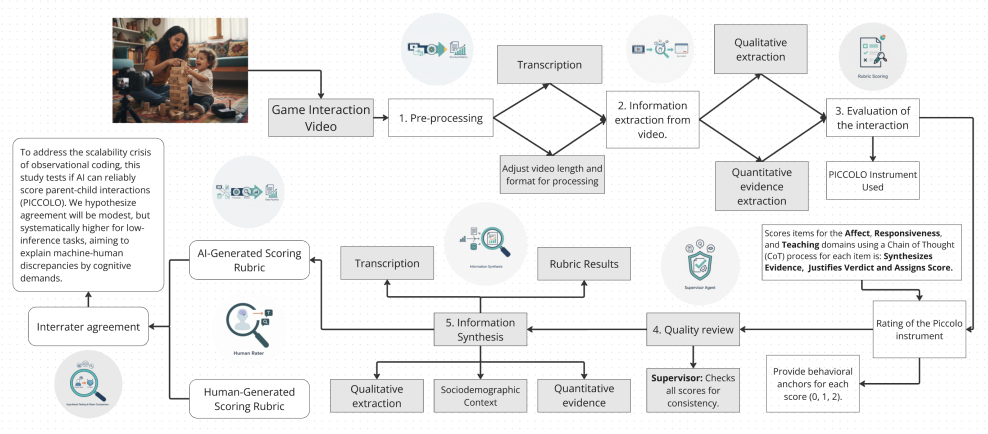

Fig. 1 AI-powered behavioral coding pipeline for parent-child interactions. This diagram illustrates the multi-stage process, from video ingestion and qualitative/quantitative evidence extraction to AI and human rubric scoring, culminating in interrater agreement analysis to benchmark AI reliability against expert consensus, thereby explaining machine-human discrepancies by cognitive demands.

2 Method

2.1 Participants

The sample for this study consisted of 442 parent-child dyads recruited in Uruguay in 2017 for a longitudinal study (Balsa et al., 2024). For the present reliability study, a random subsample of 156 unique interaction videos was analyzed, representing interactions for which scores from both the final AI pipeline and the human expert consensus were available. All children in the sample were attending early childhood centers at the time of recruitment, 44% were female, and their average age was 24 months old. 8% of children were born premature and 39% lived in a household with no other children. A large fraction of families had low socioeconomic status: almost two out of three caregivers reported receiving government aid, only one out of four mothers had a high school degree, and 23% reported at least one unmet basic need (material problems in the house, no supply of drinking water, or no access to sanitation). Table 1 summarizes these baseline characteristics. In what follows, we describe the methodology to build the pipeline for the AI-based multi-agent scoring of the interaction quality,

and explain the extent to which and why this scoring may differ from those of human experts.

Table 1 Descriptive Statistics

Panel A. Child	Mean	Std. Dev.	Min	Max
Sex=female	0.442	0.498	0	1
Preterm	0.083	0.277	0	1
Age (months)	24.126	6.811	5.233	43.867
Total number of other children living in the hhld	0.966	1.074	0	6
Panel B. Caregiver				
Lower secondary education	0.355	0.480	0	1
Upper secondary education	0.245	0.432	0	1
Mother works full-time	0.364	0.483	0	1
Mother works part-time	0.199	0.400	0	1
Father works full-time	0.781	0.415	0	1
Panel C. Household				
At least one unmet basic need	0.230	0.422	0	1
Government assistance	0.641	0.481	0	1

Nota. En los Paneles B y C, las variables son *proxy* binarias (1 = característica presente; 0 = ausente); por ello, los valores mínimo y máximo son 0 y 1.

The core of the data consists of video recordings of the full ten-minute, semi-structured free-play interaction between each parent and their child. The dyads were provided with a standardized set of age-appropriate toys designed to elicit a range of interactive behaviors. Both the interaction data and information on family's demographics and socioeconomic status were collected following ethical guidelines approved by the Ethics Committee of the Universidad Católica del Uruguay, and informed consent was obtained from all adult participants prior to their inclusion in the study.

2.2 Measures

2.2.1 Observational Coding: The PICCOLO Instrument

Parent-child interactions were coded using the *Parenting Interactions with Children: Checklist of Observations Linked to Outcomes* (PICCOLO) (Roggman, 2009). The PICCOLO is a validated observational tool designed to assess 29 specific positive parenting behaviors, which are grouped into four developmental domains: Affection, Responsiveness, Encouragement, and Teaching. Each of the 29 behavioral items is scored on a three-point Likert-type scale: '0' (Absent: the behavior is not observed), '1' (Barely: the behavior is observed briefly, is of minor quality, or is emergent), or '2' (Clearly: the behavior occurs more than once, is clear and appropriate, and of sufficient quality).

2.2.2 Human Expert Scoring

A team of nine human raters, all certified in the reliable use of the PICCOLO instrument, scored the video recordings. Most video recordings were rated by only one professional, while a random sample of 10% were scored by two independent raters to assess inter-rater reliability. The process was considered successful if Cohen's Kappa exceeded 0.90 ($\kappa > 0.90$) for each of the PICCOLO dimensions in this 10% subsample. This high threshold was met for the Affection, Responsiveness, and Teaching domains. However, the Encouragement domain failed to meet the minimum reliability criterion among the human raters. For this reason, the Encouragement scale was excluded from the present study. The human ratings on the other three dimensions serve as the gold standard against which the AI's performance is benchmarked in this paper.

2.2.3 Item Classification by Inference Level

To investigate the hypothesis that the AI's performance is contingent on the cognitive complexity of the behavior being assessed, each PICCOLO item was classified into one of three a priori categories based on the required level of rater inference:

- **Low Inference:** Items requiring the detection of concrete, objective behaviors with minimal interpretation of quality or intent (e.g., *is physically close to the child*).
- **Medium Inference:** Items requiring a qualitative judgment of a directly observable behavior (e.g., *speaks with a warm tone of voice*).
- **High Inference:** Items requiring the rater to infer pedagogical or emotional intent, context, and the purpose behind an action (e.g., *explains the reasons for things*, *is flexible to the child's change in activities*).

A complete mapping of the 22 analyzed PICCOLO items to these three inference levels is provided in Appendix Table A1.

2.3 Automated Scoring: The GABRIEL AI Pipeline

To automate the scoring process, we developed and employed a multi-agent AI pipeline named GABRIEL (Generative AI for Behavioral Research and Language-based Evaluation). The pipeline's architecture was designed to systematically analyze and score the interactions by emulating a structured, multi-stage human judgment process.

2.3.1 System Architecture and Data Preprocessing

The pipeline was built in Python and utilized several key technologies. For multimodal data processing, we used Google's Gemini family of models, which are capable of analyzing both video and text inputs simultaneously. Video files (in formats including .mp4, .mts, and .mov) were first processed using the `moviepy` library to extract the audio track. The audio was then transcribed and diarized (i.e., assigned speech to speakers) using the Google Cloud Speech-to-Text API, configured to identify two speakers (an adult and a child) in Spanish. This process yielded a full text transcription of the dialogue, along with timestamps and speaker tags for each word. Additionally, linguistic features such as dialogue coherence were calculated from the transcription using the `spaCy` library.

2.3.2 Multi-Agent Scoring and Iterative Optimization

The core of the pipeline is a multi-agent system where different AI agents, each with a specialized role, collaborate to produce the final PICCOLO scores. This Chain-of-Thought (CoT) approach was refined through several iterative cycles to improve the AI's alignment with human expert scoring. The final, optimized pipeline consisted of the following agents:

1. **The Evidence Agent**: This initial agent analyzes the video and transcription to produce a factual report. Its task is to detect the presence or absence of each of the 29 PICCOLO behaviors and to quantify their frequency (e.g., "smiled at child: Present, 5 times").
2. **The Qualitative Analyst Agent**: This agent receives the evidence report and re-analyzes the video to interpret the *quality* of the observed behaviors. It provides a narrative analysis of the interaction's emotional tone and dynamics within each PICCOLO domain.
3. **The Scorer Agent**: This agent acts as the final judge. It receives the outputs from the previous two agents, along with the sociodemographic context of the dyad. Its prompt is heavily engineered with the specific theoretical definitions of each PICCOLO item, concrete scoring criteria (e.g., the difference between a score of 1 and 2), and examples to guide its judgment. It is tasked with assigning a final score of 0, 1, or 2 to each item.
4. **The Social Desirability Agent**: As part of the developmental process, this agent was included to evaluate interactions against a set of predefined criteria for behaviors that might appear artificial or overly stylized. Its purpose was to explore the feasibility of automatically flagging interactions where social desirability could be a confounding factor, thereby helping to establish concrete criteria to ensure the reliability of the observations.
5. **The Supervisor Agent**: As a final quality control step, this agent reviews the entire output (evidence, analysis, and scores) to assess its internal consistency. For instance, it checks if a high score in "smiles at child" is logically consistent with the qualitative description of the interaction's affect.

This iterative process of refining the pipeline's architecture and the prompts for each agent was central to our methodology, allowing us to test different machine learning configurations to better approximate the nuanced judgment of human experts.

2.3.3 Ethical Considerations in AI Data Handling

To address the significant ethical considerations involved in processing video data of minors, several stringent safeguards were implemented in the AI's data handling protocol. The pipeline accessed Google's Gemini models exclusively via their Application Programming Interface (API), utilizing a configuration that explicitly opts out of data logging and model training. This technical measure guarantees that the video files and their content are not retained by the service provider or used to train future AI models. Furthermore, both the system architecture and the prompts for each agent were engineered to uphold data privacy. The generated outputs were designed to be

fully anonymized, with explicit instructions embedded in the prompts to prevent the extraction or reporting of any personally identifiable information. As a final layer of protection, all outputs were subject to human review to ensure no sensitive or identifying details concerning the children or their families were inadvertently included in the final dataset.

2.4 Data Analysis Plan

The data analysis was designed to comprehensively evaluate the reliability of the AI-generated scores against the human expert consensus and to understand the nature of any observed discrepancies.

2.4.1 Inter-Rater Reliability and Error Analysis

The scientific rigor and, ultimately, the practical utility of any observational assessment depend fundamentally on its reliability (Kane, 2013). We therefore employ a suite of complementary indices that capture distinct facets of rater alignment. Agreement is summarized with **Percent Agreement (PA)** and **Cohen's** κ (unweighted, three-category exact agreement on {0,1,2}). PA offers an intuitive starting point but can be inflated by chance agreement; κ corrects for chance by subtracting the agreement expected from the marginals (Bryington, Palmer, & Watkins, 2002; McHugh, 2012). Although chance-corrected indices can sometimes underestimate reliability (Zhao, Feng, Ao, & Liu, 2022), they remain a cornerstone for IRR assessment.

Neither PA nor κ captures the *magnitude* of disagreements. To quantify discrepancy on a common, unit-free scale that is comparable across items, we use the **normalized mean squared error** (*nMSE*). For item k with N_k video observations $i = 1, \ldots, N_k$, let Y_k^{exp} denote the random variable for the expert score and $y_{i,k}^{\text{exp}}$ and $\hat{y}_{i,k}^{\text{AI}}$ the observed expert and AI scores, respectively. We define

$$\text{nMSE}_k \;=\; \frac{\text{MSE}_k}{\text{Var}(Y_k^{\text{exp}})} \;=\; \frac{1}{N_k}\sum_{i=1}^{N_k} \frac{\left(\hat{y}_{i,k}^{\text{AI}} - y_{i,k}^{\text{exp}}\right)^2}{\text{Var}(Y_k^{\text{exp}})}\,.$$

Overall (and domain-level) *nMSE* is obtained by averaging nMSE_k across items (and across items within domain). Unless stated otherwise, indices are computed on the analysis subset specified for each result.

Notation: i indexes *videos*, k indexes *items*; uppercase symbols (e.g., Y_k^{exp}) denote random variables and lowercase (e.g., $y_{i,k}^{\text{exp}}$) denote realized observations.

To compare groups, we use nonparametric tests on *nMSE*: **Mann–Whitney** U for binary factors (e.g., child sex, maternal education) and **Kruskal–Wallis** H for multi-level factors (e.g., child age quartiles, inference level). As a complementary parametric check, we fit one-way **ANOVA** models with *nMSE* as the outcome and **inference level** as the factor, followed by **Tukey's HSD** for pairwise contrasts (adjusted p-values). All tests are two-sided with $\alpha = .05$. For κ, expected agreement is computed from the empirical marginals within each analysis stratum (overall or domain-pooled) and we report point estimates unless otherwise noted.

Across **Attempts** we summarize reliability (i) for the final, optimized pipeline, (ii) by PICCOLO domain, (iii) by item, and (iv) by a priori inference level (Low/Medium/High). Here, *Attempts* refer to *successive, improved versions of the pipeline* (prompting, exemplars, information extraction, scoring, and supervisory agents), not multiple runs of the same version. When aggregating across late iterations, we report simple means across the specified attempts.

2.4.2 Analysis of Systematic Error Patterns

To move beyond single agreement indices and characterize the machine–expert discrepancy on a common scale, we use the per-observation form implied by the definition above:

$$\mathrm{nMSE}_{i,k} \;=\; \frac{\left(\hat{y}^{\mathrm{AI}}_{i,k} - y^{\mathrm{exp}}_{i,k}\right)^2}{\mathrm{Var}(Y^{\mathrm{exp}}_{k})}\,.$$

This normalization makes error magnitudes comparable across items with different dispersions: $\mathrm{nMSE} \approx 0$ indicates negligible disagreement, values near 1 indicate disagreement on the order of the experts' dispersion, and larger values reflect greater relative error.

First, to test the effect of cognitive complexity, we group observation-level $\mathrm{nMSE}_{i,k}$ by the a priori inference categories (Low, Medium, High) and use Kruskal–Wallis to assess omnibus differences, with one-way ANOVA + Tukey's HSD as a parametric check.

Second, to probe potential systematic biases, we construct a *per-video* mean nMSE by averaging $\mathrm{nMSE}_{i,k}$ across items and compare this summary across sociodemographic groups. Binary factors (e.g., child's sex, only-child status, maternal education) are tested with two-sided Mann–Whitney U tests; child age (quartiles) is tested with Kruskal–Wallis. Throughout, we report exact test statistics and p-values and retain PA and Cohen's κ as complementary agreement indices for context.

2.4.3 Bayesian Modeling of Scoring Discrepancies

To formally model sources of discrepancy at the item level, we fit separate Bayesian ordered logistic regressions that predict the *AI's* ordinal score for each item conditional on the expert's score and observed covariates. This directly addresses the question, "What is the probability that the AI assigns a score $\leq j$ given the expert's score and context?"

For each PICCOLO item k we estimate

$$\mathrm{logit}\,\Pr\!\left(Y^{\mathrm{AI}}_{i,k} \leq j\right) \;=\; \kappa_{j,k} - \Big(\alpha_{0,k} + \alpha_{1,k}\, y^{\mathrm{exp}}_{i,k} + \mathbf{X}^{\top}_{i,k}\boldsymbol{\gamma}_k\Big), \quad j \in \{0,1\},$$

where:

- $Y^{\mathrm{AI}}_{i,k}$ is the AI's ordinal score (0, 1, 2) for video i on item k; $y^{\mathrm{exp}}_{i,k}$ is the expert's observed score.
- $\kappa_{j,k}$ are the ordered cutpoints for item k.
- $\alpha_{0,k}$ is an item-specific intercept; $\alpha_{1,k}$ captures how the AI aligns with the expert's score.

- $\mathbf{X}_{i,k}$ includes sociodemographic variables (child's sex, age, maternal education) and technical features (e.g., audio duration, dialogue coherence).

We use weakly informative priors to regularize estimation: $\alpha_{0,k}$, $\kappa_{j,k} \sim \mathcal{N}(0, 2)$ and $\alpha_{1,k}$, $\gamma_k \sim \mathcal{N}(0, 1)$. Continuous predictors are standardized. Models are implemented in Python with `PyMC`, and diagnostics are summarized with `ArviZ`.

3 Results

This section presents the results of the study, beginning with descriptive statistics of the dataset and the performance of the AI pipeline's internal quality control. Subsequently, we report the main findings regarding the inter-rater reliability between the AI model and human experts, followed by a detailed analysis of scoring discrepancies. Finally, we present the results from the machine learning and Bayesian models used to understand the nature of the systematic error between human and AI raters. Given that PICCOLO items are scored on {0,1,2}, we report a unit-free *normalized mean squared error* (**nMSE**). By construction, nMSE ≈ 0 indicates negligible error; nMSE ≈ 1 implies that the average squared disagreement is on the order of the dispersion of the expert scores; values $\gtrsim 2$ denote large error relative to item variance.

3.1 Descriptive Statistics and Pipeline Performance

The original cohort comprised 442 parent–child dyads. For the reliability analyses reported here, we used a random subsample of 156 ten-minute interactions for which both expert consensus and final-pipeline AI scores were available. The analysis focused on 22 specific PICCOLO items across the domains of Affection, Responsiveness, and Teaching, for which complete data were available.

Across these 156 interactions and 22 items, there are $22 \times 156 = 3{,}432$ potential item–video observations. Of these, $N = 2{,}399$ pairs passed the pipeline's internal supervisor acceptance and are used for machine–expert comparisons unless otherwise noted.

Table 2 Expert Item Scores by PICCOLO Domain (pooled)

Domain	Mean	Std. Dev.	Min	Max
Affection	1.448	0.717	0.000	7.000
Responsiveness	1.655	0.690	0.000	7.000
Teaching	1.363	0.762	0.000	2.000

Note. Summaries are pooled across the analysis corpus to contextualize central tendency and dispersion; observed ranges reflect domain-specific distributions.

The AI pipeline included an internal supervisor agent designed to assess the logical consistency of the outputs from the other agents. Across all runs, this supervisor agent approved approximately 69.9% of the AI-generated scoring instances as internally

consistent. Of these approved instances, the scoring agent reported a ‘High‘ level of confidence in its own judgments 93.7% of the time, with ‘Medium‘ confidence in 6.2% of cases and ‘Low‘ confidence in only 0.1% of cases. This indicates that for the majority of outputs considered valid by the pipeline's quality control, the model had a high degree of internal certainty in its generated scores.

3.2 AI Pipeline Optimization and Inter-Rater Reliability

The development of the AI pipeline involved an iterative process of architectural and prompt engineering adjustments aimed at improving its alignment with human expert scores. The performance of thirteen distinct pipeline versions was tracked. While the improvement was not strictly linear, a clear trend of increasing reliability emerged in later iterations. The final optimized version of the pipeline demonstrated the most robust performance.

Across Attempts 1–13, successive refinements included: (i) prompt clarifications of item definitions and quality anchors; (ii) addition of few-shot, domain-specific exemplars; (iii) improved multimodal evidence extraction and speaker diarization; (iv) explicit, item-level decision rules distinguishing scores 1 versus 2; and (v) supervisory checks for internal consistency, rejecting outputs that violate rubric–evidence coherence.

When comparing the final aggregated AI scores against the human expert consensus, the machine–expert agreement was modest (overall PA = 56.87%; $\kappa = .128$ across $N = 3{,}102$ item–video pairs). In a subsequent refinement focused on calibrating for the AI's scoring tendencies (Attempt #13), agreement reached PA = 50.68% with $\kappa = .216$ ($N = 880$). This $\kappa = .216$ is among the highest kappas observed; the maximum in this batch was $\kappa = .221$ (Attempt 11; $N = 1{,}260$).

To contextualize the final calibration, inter-rater reliability (IRR) averaged across late iterations (*Attempts 10–13*) yielded an overall percent agreement of *50.72%* and an overall Cohen's κ of *.197* (simple means across attempts). For the final calibrated model (*Attempt #13*), overall % PA was *50.68%* with $\kappa = .216$ (N=880 item–video pairs). By domain in *Attempt #13*, % PA was *60.00%* (Teaching), *57.50%* (Affection), and *35.00%* (Responsiveness).

3.3 Analysis of Scoring Discrepancies

3.3.1 Systematic Distributional Bias

A primary source of disagreement was a systematic bias in the AI's scoring distribution compared to the human experts. We define *overuse* (or *underuse*) of a category as the percentage-point difference in its usage between AI and expert (AI minus Expert). As shown in Figure 2, the AI tended to *overuse* the middle score ‘1‘ (Barely Present) and *underuse* the highest score ‘2‘ (Clearly Present). For instance, for the item "speaks with a warm tone of voice", the AI assigned a score of ‘1‘ in 93.3% of cases, whereas experts did so only 24.1% of the time. Conversely, experts assigned a score of ‘2‘ in 71.5% of cases, while the AI did so in only 6.7%. This pattern suggests the AI favored a conservative “middle-ground” rating and systematically underestimated the presence of high-quality parenting behaviors.

3.3.2 Item-Level Error Analysis

We computed item-wise normalized error (nMSE) to identify which items were most and least problematic for the AI. The highest-error items concentrated in the Teaching domain (e.g., *explains characteristics of objects*, *carries out activities in a sequence of steps*, *explains the reasons for things*), whereas lower-error items were concrete and directly observable (e.g., *is physically close to the child*, *responds to child's words or sounds*, *observes the child when speaking or making sounds*). This pattern is consistent with the notion that items requiring greater cognitive inference generate larger machine–expert discrepancies.

3.3.3 Discrepancy by Cognitive Inference Level

To quantify the role of cognitive inference using a unit-free scale, we analyzed normalized error ($nMSE = \text{MSE}/\text{Var}(Y_{\text{item}})$). In the pooled analysis across all processed files, nMSE differed markedly by inference level (Kruskal–Wallis $H = 308.70$, $p<.001$), with larger errors for more inferentially demanding judgments. For reference, an early pooled dataset showed average nMSE of 2.51 (Low), 1.42 (Medium), and 1.96 (High), whereas a consolidated final dataset yielded a lower, monotonic gradient of 1.16 (Low), 1.50 (Medium), and 1.63 (High). Focusing on the most refined configurations (Attempts 10–13), the inference effect persisted (Kruskal–Wallis $H = 75.78$, $p<.001$), with average nMSE of 1.78 (Low), 2.01 (Medium), and 1.76 (High). As a complementary parametric check, a one-way ANOVA on nMSE also indicated significant differences across inference levels ($p<.001$), and Tukey's HSD pairwise comparisons were significant for all contrasts (Table 3).

Table 3 Tukey's HSD Post-Hoc Comparisons for **nMSE** by Inference Level

Inference level group 1	Inference level group 2	Mean Diff. (nMSE)	p-adj	Lower CI	Upper CI	Reject H_0
High	Low	0.5091	<.001	0.2762	0.7420	True
High	Medium	0.1454	0.030	0.0109	0.2799	True
Low	Medium	-0.3637	<.001	-0.5966	-0.1308	True

Values are expressed in normalized units ($\text{nMSE} = \text{MSE}/\text{Var}(Y_{\text{item}})$). Omnibus one-way ANOVA on nMSE was significant ($p<.001$).

3.3.4 Analysis of Heterogeneous Effects

In the pooled analysis across attempts, nonparametric tests indicated no systematic differences in error by child sex (Mann–Whitney $U = 2{,}843.00$, $p=.888$) and no differences by child age quartiles (Kruskal–Wallis $H = 0.80$, $p=.850$). Error magnitudes continued to track the a priori cognitive inference required by each item, with larger normalized error for more inferentially demanding judgments (see Appendix Table A2 for per-attempt nMSE summaries by inference).

Table 4 Nonparametric Tests for Heterogeneous Effects on Normalized MSE (nMSE)

Variable	Test	Statistic	p-value
Panel A: All processed files			
Maternal education	Mann–Whitney U	117.00	.106
Child's sex	Mann–Whitney U	29,045.00	.712
Child age (quartiles)	Kruskal–Wallis H	4.67	.197
Item inference level	Kruskal–Wallis H	308.70	$<.001$
Panel B: Late iterations (Attempts 10–13)			
Maternal education	Mann–Whitney U	3.00	.200
Child's sex	Mann–Whitney U	2,331.00	.159
Child age (quartiles)	Kruskal–Wallis H	3.81	.282
Item inference level	Kruskal–Wallis H	75.78	$<.001$

nMSE = MSE/Var(Y_{item}); lower values indicate smaller disagreement relative to item variance. Tests are two-sided.

Focusing on the most refined configurations (Attempts 10–13), the inference effect persisted (Kruskal–Wallis $H = 75.78$, $p<.001$), with average nMSE of 1.78 (Low), 2.01 (Medium), and 1.76 (High). Heterogeneity checks remained null for child sex (Mann–Whitney $U = 2{,}331.00$, $p=.159$) and maternal education (Mann–Whitney $U = 3.00$, $p=.200$). Regarding child age, mean nMSE by quartile was 1.64 (Q1), 2.24 (Q2), 1.57 (Q3), and 1.79 (Q4), with no overall differences detected (Kruskal–Wallis $H = 3.81$, $p=.282$). These patterns reinforce that cognitive inference demands are the primary driver of machine–expert discrepancy, rather than demographic covariates.

3.4 Further Probes for Systematic Bias

To reconcile the observed distributional tendency (overuse of the middle category 1 and underuse of 2) with the Bayesian *no global bias* finding, note that the ordered-logit intercept tests a *uniform level shift* in log-odds after adjusting for covariates. The absence of an intercept effect indicates no single upward or downward displacement across items. By contrast, the empirical pattern reflects differences in *category usage* and *item-specific thresholds* that are strongly structured by cognitive inference. Thus, there is no contradiction: the intercept addresses a global shift, whereas our discrepancies arise from how categories are deployed and from difficulty/threshold parameters at the item level.

As a formal check, item-wise Bayesian ordered-logit models yielded intercept posteriors centered near zero on average; the across-item mean 95% credible interval included zero, indicating no uniform upward or downward displacement once covariates were controlled. This harmonizes with the distributional evidence: the intercept tests a global level shift, whereas the observed discrepancies reflect *category-usage* differences and *item-specific thresholds* structured by inference demands.

4 Discussion

The iterative process of developing and refining the GABRIEL pipeline was more than a technical exercise in model optimization; it was a pedagogical endeavor that provided a unique window into the promises and perils of applying generative AI to the nuanced task of psychological assessment. This journey of "teaching" a machine to observe parenting interactions reveals less about achieving a perfect reliability score and more about the fundamental nature of human judgment, the current frontiers of artificial intelligence, and the path forward for its responsible implementation. This discussion uses our developmental process as a heuristic to first, deconstruct what the AI's successes and failures teach us about its capabilities; second, to explore the profound implications for early childhood intervention in resource-constrained contexts like Latin America; and third, to outline a future centered on human-AI collaboration.

4.1 The Optimization Journey as Computational Pedagogy

Our path through thirteen distinct pipeline versions underscores that training a generative AI for a psychological task is fundamentally a pedagogical process, not merely a software engineering one. The systematic refinement of prompts, the explicit inclusion of theoretical definitions from the PICCOLO manual, and the use of few-shot learning examples functioned as instructional scaffolds for the AI, a process directly analogous to how human raters are trained and certified (de Kleijn, 2023; Winkler et al., 2020). Our findings demonstrate that these "teaching" methods have a direct and measurable impact on the model's reliability.

However, the non-linear and often unpredictable nature of the improvements suggests that this is not a straightforward process of simply providing more data. Rather, it is a delicate calibration exercise in translating abstract psychological theory into a machine-interpretable format, hinting at the complex and still poorly understood internal representations the model develops (Brock, 2024; Strachan et al., 2024). This reframes AI development in this domain as an applied "computational pedagogy," where the central challenge is to effectively scaffold the machine's learning process to align its "judgment" with the rich, context-aware cognition of human experts (Lin et al., 2023; Mindigulova et al., 2024; Sayed et al., 2023). The most critical lessons, therefore, came not from the final reliability scores, but from the nature of the disagreements that persisted throughout this process.

4.2 What AI Errors Teach Us: Deconstructing Machine–Human Disagreement

Our findings delineate a clear reliability gradient across cognitive inference levels, validating the hypothesis that inferential complexity is a primary determinant of AI performance. For low-inference items (nMSE $hickapprox1.14$), the GABRIEL pipeline achieved error margins comparable to human–human variability, enabling accurate detection of concrete behaviors such as physical proximity or shared attention. In contrast, medium-inference items showed nMSE $hickapprox1.49$, and high-inference items exhibited nMSE $hickapprox1.51$, reflecting the increased cognitive load required

to integrate temporal context, emotional nuance, and pedagogical intent. This gradient highlights the architectural limitations of current generative models and signals the need for targeted optimization. The observed performance profile strongly suggests a functional division of labor for scalable behavioral assessment: automated systems can reliably handle low-inference tasks with high sensitivity (K approaching 0.22, PA 50–70%) while flagging complex, high-inference behaviors for expert review, where interpretive fidelity remains essential. Such a tiered architecture parallels the hybrid design of early developmental screening interventions (Giraldo-Huertas, 2023) and could transform observational research into a scalable, equitable framework for real-time developmental monitoring.

This inference-driven error pattern was statistically robust and consistent. In a pooled analysis, the normalized mean squared error (nMSE) differed markedly by the a priori cognitive inference required by each item (Kruskal–Wallis $H = 308.70$, $p<.001$), and the gradient persisted across the most refined pipeline configurations (Attempts 10–13: Kruskal–Wallis $H = 75.78$, $p<.001$). Crucially, heterogeneity tests indicated no systematic differences in nMSE by child sex, age quartile, or maternal education. This confirms that cognitive inference demands—rather than observed demographics—are the primary driver of machine–expert discrepancy.

The practical implications of this gradient define both the technology's main potential and its principal peril. The potential lies in the scalable detection of concrete behaviors. The model performed best on low-inference items demanding minimal subjective judgment, such as *is physically close to the child* (MSE=0.113) and *pays attention to what the child does* (MSE=0.180). This confirms its value as a first-pass detector of objective events, capable of automating the most laborious layer of observational coding. The peril, however, is its struggle with qualitative and contextual judgment. Error rises sharply for high-inference items that require interpreting intent, like *explains characteristics of objects* (MSE=1.128). This exposes the model's difficulty with pedagogical intentionality; it can see a parent talking about a toy, yet it does not reliably understand the underlying scaffolding purpose.

This limitation is further illuminated by a systematic distributional bias: the model's tendency to overuse the middle score of 1 (Barely) and underuse the score of 2 (Clearly)—a cautious middle default when faced with inferential uncertainty. Such a bias is not merely technical; it reflects the deeper challenge of translating rich psychological constructs, laden with context and theory of mind, into machine-parsable logic (Boden, 1988; Carroll, 1997). If deployed uncritically, this pattern risks systematically underestimating parenting strengths in high-inference domains (O'Neil, 2016).

Beyond a qualitative description, item–level diagnostics reveal that middle-category inflation is large in magnitude and pervasive across domains. Using 21/22 PICCOLO items for which distributional summaries were available, the median overuse of the '1' score by AI (Machine minus Expert) was +44.7 percentage points (mean = +46.8 pp), with the underuse of '2' universal (21/21 items; median = 40.2 pp). The most extreme middle inflation appeared for Responsiveness (*"pays attention to what the child does"*: +81.7 pp), while a characteristic low-inference anchor such as Affection's *"is physically close to the child"* exhibited only modest inflation (+20.4 pp) and concentrated disagreement on the upper category (2). In particular,

'explains the reasons for things" in Teaching departs from the modal pattern: in addition to underusing '2' (33.7 pp), the model overuses '0' (+26.5 pp), suggestive of an abstention strategy when causal discourse and pedagogical intentionality are required. Figure 2 synthesizes these distributional shifts.

Two mechanisms plausibly drive these patterns. First, an *ambiguity-aversion* heuristic: when evidence is noisy or inferentially demanding, the scoring agent defaults to the ordinal middle to minimize expected loss under symmetric penalties. Second, *item-specific threshold miscalibration*: empirical category boundaries for 2 appear too conservative, particularly in items where experts apply qualitative criteria (warm prosody, sustained engagement, causal language). Three corrective levers follow: (i) **item-wise ordinal recalibration** (e.g., isotonic or Platt-style post-hoc mapping) using expert marginals to right-shift the 2 threshold where warranted; (ii) **feature augmentation** aligned with theory—prosodic contours for *"warm tone"*, temporal synchrony for *"sustained engagement"*, and causal cue phrases for *"explains reasons"*; and (iii) **cost-sensitive tuning** that penalizes false negatives on 2 more heavily for items intended to capture strengths. Together, these adjustments target the observed middle inflation without inflating spurious 2 calls.

Importantly, this observed distributional pattern does not imply a uniform directional bias. Item-wise ordered-logit models showed no evidence of a global upward or downward shift once covariates were controlled, as the intercept (baseline log-odds) was not reliably different from zero on average. This reconciles the two findings: the model's errors are not from a simple tendency to score higher or lower overall, but are structured by complex differences in category usage and item-specific thresholds that covary directly with cognitive inference demands.

Viewed through a computational-pedagogy lens, these diagnostics specify what the model still needs to be "taught." Responsiveness shows the highest median middle-category inflation (+52.5 pp across items), suggesting that instruction must make explicit the transition rules from "emergent" to "clear" quality when the evidence is prosodic, dyadic, and temporally distributed. Affection, by contrast, exhibits lower inflation on low-inference anchors (e.g., *physical proximity*), indicating that the current instruction set suffices for concrete behavioral detection. Teaching reveals a dual challenge: besides conservative thresholds for "2," certain items (e.g., *explains reasons*) require reliable detection of causal discourse and pedagogical intent beyond surface verbalisations. Targeted exemplars and threshold calibration at the item level operationalize this "last mile" of instruction.

4.3 Implications for Early Intervention targeting Low Income Populations

The practical implications extend to resource-constrained systems as well as programs targeting vulnerable populations in high-income settings. As a first-pass screening tool, a pipeline like GABRIEL can automate the detection of concrete, low-inference behaviors at scale, reducing the human time required for initial observational review while routing high-inference judgments to certified experts. This division of labor aligns observed reliability patterns: lower nMSE for concrete items suggests near-term utility for scalable screening, whereas higher nMSE for inferential judgments

underscores the need for human oversight where contextual interpretation, pedagogical intent, and quality appraisals are central. The "scalability crisis" of observational methods described in the introduction is particularly acute in low and middle income countries, where the number of trained developmental specialists is vastly insufficient to meet population needs (Cuartas et al., 2020; Giraldo-Huertas, 2023).

The baseline profile in Table 1 magnifies the policy stakes: in this cohort, 64% of caregivers reported receiving government assistance, only 24.5% of mothers had completed upper secondary education, and 23% of households faced at least one unmet basic need. In such contexts, a systematic under-calling of the highest category (2) on items that index relational strengths (e.g., warm tone, sustained engagement, responsive following) risks *attenuating observed assets* precisely where positive resources should be recognized and leveraged. A practical guardrail is to implement *selective human review* of the top decile of 1 predictions in items where experts frequently use 2 (e.g., *warm tone*, *devotes full attention*, *names objects/actions*), converting the model's conservative middle into a triage signal for strengths verification rather than a de facto downgrade.

While equity considerations are encouraging, they warrant caution. AI holds a transformative, democratizing potential; a tool like GABRIEL could enable governments and NGOs to implement developmental monitoring at an unprecedented scale, identifying at-risk children and families far more efficiently. Although we found no systematic differences in performance by child sex, age, or maternal education, the risks are magnified in these same contexts. The model's conservative tendency to favor lower ratings can *underestimate* clearly present strengths, particularly in high-inference situations. Deploying an AI with such an inherent bias, or one that fails to understand culturally-specific practices, could lead to flawed policy decisions and stigmatize vulnerable communities. For instance, a model trained on urban Uruguayan dyads may not generalize to diverse populations across the region, from indigenous communities in the Andes to rural families in Central America. Therefore, deployments should pair automated, low-inference detection with expert review for nuanced qualitative judgments and must be validated across diverse linguistic and cultural settings to ensure robust and equitable generalization.

Encouragingly, while our analysis identified biases related to task complexity, our preliminary exploration of heterogeneous effects found no evidence that the AI's error was systematically biased by the child's sex or only-child status. While this requires validation on larger and more diverse samples, it is a promising initial sign for the equitable application of this technology. It suggests that the model's primary limitations are tied to the cognitive nature of the task itself, rather than the demographic characteristics of the individuals being observed. This finding strengthens the case for its potential as a fair screening tool, provided its inferential limitations are appropriately managed in a human-in-the-loop system.

4.4 Limitations

This study has several limitations that should be considered. First, the sample was drawn from a specific cultural and linguistic context (Uruguay), which may limit the generalizability of the findings. Further, while our heterogeneity analysis did not

find evidence of demographic bias, this null finding is based on a limited sample and requires replication with larger, more diverse populations to be considered conclusive. Second, the pipeline was based on a specific large multimodal model (Google's Gemini family); future or different models may exhibit different performance characteristics (Georgiev et al., 2024). Third, our gold standard was a human expert score. While this is standard practice, it is still a proxy for the latent construct of interaction quality and may embed residual human variability. Relatedly, our adoption of nMSE (MSE/Var) as the primary unit-free error metric improves comparability across items with different dispersions but does not eliminate the need for complementary agreement indices (e.g., percent agreement, Cohen's κ) when interpreting practical significance.

Privacy and data protection were central to the analysis. Video processing used API configurations that explicitly opt out of provider data logging and training, prompts instructed agents to avoid extracting personally identifiable information, outputs were anonymized by design, and a human review step screened for any sensitive content prior to analysis. These safeguards are essential when working with recordings of minors and should be considered a minimum standard for future work.

The analytic sample comprised 156 unique interactions with both AI and expert scores from an original frame of 442 dyads. Clarifying inclusion pathways and any technical exclusions is important for reproducibility and external validity and will be documented alongside versioned code and data descriptors. Finally, category-specific Cohen's κ (e.g., by domain or inference level) requires stratified confusion matrices to compute expected agreement; these estimates can be produced once those paired counts are finalized.

A diagnostic limitation is that our distributional analysis draws on 21/22 items with mapped score frequencies; therefore, summary statistics (e.g., median inflation of 1) reflect that subset. Moreover, distributional gaps are unweighted by item prevalence within domains. Future releases should report prevalence-weighted and domain-stratified calibration curves to complement unit-free nMSE and chance-corrected agreement.

Methodologically, our item-wise Bayesian ordered logistic models tested for a uniform directional bias and found none on average. A natural extension is to incorporate explicit indicators for cognitive inference (e.g., High, Medium; Low as the reference) to quantify how thresholds shift with inferential demands in a single hierarchical framework. Such models could also pool information across items to stabilize estimates where sample sizes are limited.

4.5 Conclusion: Forging Human-AI Co-Intelligence

This study provides a rigorous benchmark of a multimodal generative AI pipeline (GABRIEL) for PICCOLO scoring and, more critically, *explains* when and why machine–expert alignment degrades. We introduce a framework that classifies behaviors by their required level of cognitive inference, enabling us to identify not only whether the AI agrees with experts but *under what conditions* its reliability falters. In doing so, we directly probe the challenge of operationalizing complex psychological theory for machine learning and explore the potential of generative AI to support

more scalable, equitable, and accessible developmental systems—particularly in Latin America (Giraldo-Huertas, Sánchez, & Gutiérrez-Romero, 2023; Lin et al., 2023).

The path forward, illuminated by our findings, is not a narrow pursuit of perfect machine–human agreement—an objective that our qualitative divergence patterns suggest may be neither feasible nor desirable. Instead, the evidence supports a collaborative, *co-intelligence* pathway (Mollick, 2024). In the final calibrated configuration, agreement with experts was modest (overall % PA = 50.68%, Cohen's $\kappa = .216$), and normalized error analyses (nMSE = MSE/Var) showed that discrepancies are primarily organized by the *cognitive inference* required by items rather than by observed demographics (sex, age quartile, maternal education). This structure argues for a hybrid workflow in which AI scales the detection of concrete, low-inference behaviors and *selectively escalates* clips for expert adjudication when judgments hinge on pedagogical intent, contextual nuance, or quality appraisal.

Operationalizing this paradigm entails two complementary streams of work. First, deepen *explanatory* analyses of errors—item-level audits and qualitative case reviews—to identify where and why model and human judgments diverge, and use those insights to refine prompts (e.g., dynamic chain-of-thought) and exemplars specifically for high-inference constructs. Second, improve *decision mechanics*: incorporate structured, domain-informed features, and calibrate category thresholds to mitigate the model's conservative, middle-category overuse without inflating false positives on 2 (Clearly Present).

Deployment should follow a *human-in-the-loop* design that pairs machine efficiency with human depth and explicit oversight (Dellermann et al., 2019; Spring, Faulconbridge, & Sarwar, 2022). Practically, this means using AI for first-pass screening on large video corpora, routing high-inference segments to certified coders, and monitoring reliability with unit-free error (nMSE) alongside agreement indices (PA, κ). To ensure equity and robustness, validation must extend across diverse linguistic and cultural contexts and varied recording conditions, under strong privacy safeguards like those used here (provider logging/training opt-outs, anonymized outputs, and human review of sensitive content).

The aim is not to replace expert judgment but to *extend its reach*: combining machine scale with human discernment to deliver more equitable, efficient, and context-sensitive developmental support. Progress is likely to come from three fronts: (a) strengthening instruction and examples targeted at high-inference items, (b) integrating structured features with calibrated thresholds to address category-usage biases, and (c) broadening validation under rigorous privacy and governance. This co-intelligence approach aligns the technology's current strengths with ethical and scientific safeguards, turning AI into a scalable ally for early childhood assessment rather than a substitute for human expertise.

In conclusion, this study provides a realistic, empirically grounded benchmark for generative AI in psychological assessment. But more importantly, it reframes the ultimate goal: the aim is not to replace human expertise, but to augment and democratize it. Our findings indicate that while AI is not yet a surrogate for the seasoned judgment of a clinician, it has crossed a significant threshold to become a viable tool for

scalable screening. The GABRIEL pipeline's ability to reliably detect specific parenting behaviors, even if modestly, demonstrates a clear potential to help overcome the critical barriers to developmental support, particularly in resource-constrained regions like Latin America (Giraldo-Huertas, 2023; Jeong et al., 2021). The ultimate promise of this technology lies not in its autonomy, but in the synergistic partnership we can forge between the scale of artificial intelligence and the wisdom of human experience. By building this collaborative future, we can create more equitable, accessible, and effective pathways to support the development of children and families everywhere.

Open Practices Statement

This study adheres to the TOP Guidelines (Level 2). Analysis code, documentation, and a privacy-preserving synthetic dataset that mirrors the statistical properties of the original data are permanently available at the Open Science Framework (OSF) under DOI: https://doi.org/10.17605/OSF.IO/UYS68. Raw videos and identifiable transcripts cannot be shared due to ethics and participant privacy constraints; the synthetic dataset enables full computational reproducibility of all reported analyses.

Declarations

Funding

This work was supported by Universidad de La Sabana under Grant Agreement No GL0000000001695. Data collection was funded by Reaching U, A Foundation for Uruguay and the InterAmerican Development Bank (UR-T1190-P001).

Conflicts of interest/Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

The study was approved by the Ethics Research Committee of the Universidad Católica del Uruguay (Comité de Etica en Investigación, Universidad Católica del Uruguay) under number A-22-08-17 and final resolution on August 31, 2017. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Consent to participate

Informed consent was obtained from all individual participants included in the study.

Consent for publication

All adult participants (or legal guardians, as applicable) provided written informed consent for participation and for the publication of de-identified, aggregate results and illustrative excerpts. No identifiable information will be published.

Availability of data, materials, and code

All materials required for computational reproducibility are publicly available at OSF: https://doi.org/10.17605/OSF.IO/UYS68. The repository includes: (i) a Jupyter/Colab notebook to generate privacy-preserving synthetic data with the same structure and scale as the analysis corpus (`Analisis_videos_Uruguay (2).ipynb`); (ii) documented analysis scripts to reproduce tables and figures on the synthetic dataset; (iii) codebooks/metadata (`CODEBOOK_item_variables.csv`, `item_metadata.csv`); and (iv) study materials and prompts used by the multi-agent pipeline (GABRIEL). Due to ethics and consent restrictions, raw videos and identifiable transcripts of parent–child interactions are not shareable. The code is released under the MIT License; the synthetic data and documentation are provided to enable full computational reproducibility of the analyses presented in this manuscript.

Authors' contributions

Conceptualization: José Amorocho, Juan José Giraldo, Ana Balsa. Methodology: José Amorocho, Juan José Giraldo, Ana Balsa, Juanita Bloomfield, Paula Patrone. Software: José Amorocho. Formal Analysis: José Amorocho. Investigation: Ana Balsa, Juanita Bloomfield, Paula Patrone, Alejandro Cid. Writing – Original Draft: José Amorocho. Writing – Review & Editing: All authors reviewed and edited the manuscript. Supervision: Ana Balsa, Juan José Giraldo.

References

Al-Gharib, S.A., & Gouda, A.M.A.-G.M. (2024). User experience and its impact in the design of service robots. *Majallat al-ʿimārah wa-al-funūn wa-al-ʿulūm al-insānīyah*, *9*(45), 550–564, https://doi.org/10.21608/MJAF.2022.146760.2792

Aprilia, F., & Neisya, N. (2024). Insights into Early Language Development: An Investigation of Speech Acquisition in 25-Month-Old Toddler. *Journal of Languages and Language Teaching (Online)*, *12*(3), 1491–1502, https://doi.org/10.33394/jollt.v12i3.11123

Atkins, D.C., Steyvers, M., Imel, Z.E., Smyth, P. (2014). Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, *9*(1), 49, https://doi.org/10.1186/1748-5908-9-49

Balsa, A.I., Gómez Muzzio, E., González, M.L., Bloomfield, J., Cid, A., Valdés, R. (2024). Crianza Positiva: Combining Group Workshops and E-Messages to Strengthen Parenting Competences. *Child & Youth Care Forum*, *53*(3), 719–740, https://doi.org/10.1007/s10566-023-09772-4

Boden, M.A. (1988). *Computer models of mind: Computational approaches in theoretical psychology.* Cambridge University Press.

Brock, O. (2024). Intelligence as computation. *arXiv preprint arXiv:2405.16604*, ,

Bryington, A.A., Palmer, D.J., Watkins, M.W. (2002). The estimation of interobserver agreement in behavioral assessment. *The Behavior Analyst Today*, *3*(3), 323–328, https://doi.org/10.1037/h0099978

Carroll, J.M. (1997). Human–computer interaction: Psychology as a science of design. *International Journal of Human-Computer Studies*, *46*(4), 501–522, https://doi.org/10.1006/ijhc.1996.0101

Chen, Z., Xu, L., Zheng, H., Chen, L., Tolba, A., Zhao, L., ... Feng, H. (2024). Evolution and prospects of foundation models: From large language models to large multimodal models. *Computers, Materials and Continua*, *80*(2), 1753-1808, https://doi.org/https://doi.org/10.32604/cmc.2024.052618 Retrieved from https://www.sciencedirect.com/science/article/pii/S1546221824005472

Cuartas, J., Rey-Guerra, C., McCoy, D.C., Hanno, E. (2020). Maternal knowledge, stimulation, and early childhood development in low-income families in colombia. *Infancy*, *25*(5), 526–534, https://doi.org/10.1111/infa.12335

de Kleijn, R.A. (2023). Supporting student and teacher feedback literacy: an instructional model for student feedback processes. *Assessment & Evaluation in Higher Education*, *48*(2), 186–200,

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., Ebel, P. (2019). The future of human-ai collaboration: a taxonomy of design knowledge for intellectually partnering with cognitive assistants. *Proceedings of the 52nd hawaii international conference on system sciences.*

Desimone, L.M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199, https://doi.org/10.3102/0013189X08331140

Edwards, D.J. (2024). A functional contextual, observer-centric, quantum mechanical, and neuro-symbolic approach to solving the alignment problem of artificial general intelligence: safe AI through intersecting computational psychological neuroscience and LLM architecture for emergent theory of mind. *Frontiers in*

Computational Neuroscience, *18*, 1395901, https://doi.org/10.3389/fncom.2024.1395901

Egmose, I., Varni, G., Cordes, K., Smith-Nielsen, J., Væver, M.S., Køppe, S., ... Chetouani, M. (2017). Relations between automatically extracted motion features and the quality of mother–infant interactions at 4 and 13 months. *Frontiers in Psychology*, *8*, 2178, https://doi.org/10.3389/fpsyg.2017.02178

Fitria, Y., & Nurcholis, I.A. (2025). Observation of the relationship between play activities and language development in children aged 3-5 years and the role of parents. *Indonesian Journal Of Education*, *2*(1), 42–52,

Flemotomos, N., Martinez, V.R., Chen, Z., Singla, K., Ardulov, V., Peri, R., ... Narayanan, S. (2022). Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods*, *54*(2), 690–711, https://doi.org/10.3758/s13428-021-01623-4

Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., ... et al., J.L. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.* Retrieved from https://arxiv.org/abs/2403.05530

Giraldo-Huertas, J. (2023). Parental developmental screening with CARE: A pilot hybrid assessment and intervention with vulnerable families in Colombia. *PloS One*, *18*(6), e0287186, https://doi.org/10.1371/journal.pone.0287186

Giraldo-Huertas, J., Sánchez, D.C., Gutiérrez-Romero, M.F. (2023). Efectos en el desarrollo cognitivo de niños y niñas en condición de riesgo y pobreza multidimensional de dos intervenciones con cuidadores principales. *Revista Complutense de Educación*, *34*(1), 157–166, https://doi.org/10.5209/rced.77229

Goodwin, L.D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science*, *5*(1), 13–34,

Horowitz-Kraus, T., & Gashri, C. (2023). Multimodal approach for characterizing the quality of parent–child interaction: A single synchronization source may not tell the whole story. *Biology*, *12*(2), 241, https://doi.org/10.3390/biology12020241

Hu, X., & Twidale, M. (2023). A Scoping Review of Mental Model Research in HCI from 2010 to 2021. D.D. Schmorrow et al. (Eds.), *Hci international 2023 - late*

breaking papers (Vol. 14054, pp. 101–125). Springer.

Huang, Q., Sun, J., Lau, E.Y.H., Zhou, Y.-l. (2023). Parental scaffolding and children's math ability: The type of activities matters. *British Journal of Developmental Psychology*, *41*(3), 246–258, https://doi.org/10.1111/bjdp.12444

Idalski Carcone, A., Hasan, M., Alexander, G.L., Dong, M., Eggly, S., Hartlieb, K.E.B., ... Kotov, A. (2019). Developing machine learning models for behavioral coding. *Journal of Pediatric Psychology*, *44*(3), 289–299, https://doi.org/10.1093/jpepsy/jsy113

Ismai, N., Juhari, R., Madon, Z., Arshat, Z., Daud, M.N., Azam, N.D. (2024). Parental scaffolding and problem-solving skills in early childhood: A systematic literature review. *Library of Progress-Library Science, Information Technology & Computer*, *44*(3), ,

Jebeli, A., Chen, L.K., Guerrerio, K., Papparotto, S., Berlin, L., Jones Harden, B. (2024). Quantifying the quality of parent–child interaction through machine-learning based audio and video analysis: Towards a vision of ai-assisted coaching support for social workers. *ACM Journal on Computing and Sustainable Societies*, *2*(1), 1–21, https://doi.org/10.1145/3617693

Jeong, J., Franchett, E.E., Ramos de Oliveira, C.V., Rehmani, K., Yousafzai, A.K. (2021). Parenting interventions to promote early child development in the first three years of life: A global systematic review and meta-analysis. *PLoS medicine*, *18*(5), e1003602,

Kaminski, J.W., Valle, L.A., Filene, J.H., Boyle, C.L. (2008). A meta-analytic review of components associated with parent training program effectiveness. *Journal of Abnormal Child Psychology*, *36*(4), 567–589, https://doi.org/10.1007/s10802-007-9201-9

Kane, M.T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73, https://doi.org/10.1111/jedm.12000

Landry, S.H., Smith, K.E., Swank, P.R. (2008). The role of parents in early childhood learning. *Handbook of developmental psychology* (pp. 193–217). Sage.

Leclère, C., Avril, M., Viaux, S., Achard, C., Chetouani, M., Missonnier, S., Cohen, D. (2016). Interaction and behaviour imaging: a novel method to quantify

mother–infant interaction using synchronized audio–video and motion capture data. *Translational Psychiatry*, *6*, e816, https://doi.org/10.1038/tp.2016.82

Lee, Y.L. (2022). Scaffolding online learning in a higher education institution: a need analysis/lee yee ling. *ESTEEM Journal of Social Sciences and Humanities*, *6*(2), 71–88,

Lin, C.-C., Huang, A.Y.Q., Lu, O.H.T. (2023, 8). Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart Learning Environments*, *10*, 41, https://doi.org/10.1186/s40561-023-00260-y

McHugh, M.L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282, https://doi.org/10.11613/BM.2012.031

Mindigulova, A., Vikhman, V., Romm, M. (2024). Artificial intelligence and personalized learning: Scaffolding technology. *Professional Education in the Modern World*, *13*(4), 613–622,

Mollick, E. (2024). *Co-intelligence: Living and working with ai.* New York: Portfolio/Penguin.

Morris, A.S., & Smith, J.M. (2022). *The cambridge handbook of parenting: Interdisciplinary research and application.* Cambridge University Press.

Myruski, S., & Dennis-Tiwary, T.A. (2022). Observed parental spontaneous scaffolding predicts neurocognitive signatures of child emotion regulation. *International Journal of Psychophysiology*, *177*, 111–121, https://doi.org/10.1016/j.ijpsycho.2022.05.004

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown.

Pérez-Rosas, V., Mihalcea, R., Resnicow, K. (2017). Predicting counselor behaviors in motivational interviewing encounters. *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 1, long papers* (pp. 1128–1137). Association for Computational Linguistics.

Rey-Guerra, C., Maldonado-Carreño, C., Ponguta, L.A., Nieto, A.M., Yoshikawa, H. (2022). Family engagement in early learning opportunities at home and in early childhood education centers in colombia. *Early Childhood Research Quarterly*, *58*, 35–46, https://doi.org/https://doi.org/10.1016/j.ecresq.2021.08.002

Roggman, L.A. (2009). *Piccolo (parenting interactions with children: Checklist of observations linked to outcomes): An overview of an observational measure of developmental parenting.* Vancouver, BC. (Presentación de PowerPoint)

Roggman, L.A., Cook, G.A., Innocenti, M.S., Norman, V.J., Christiansen, K., Anderson, S. (2013). Piccolo in diverse ethnic groups of mothers and fathers: Relations to parenting and child language in toddlers. *Infant Mental Health Journal*, *34*(4), 306–321, https://doi.org/10.1002/imhj.21389

Rubio-Codina, M., & Grantham-McGregor, S. (2020). Predictive validity in middle childhood of short tests of early childhood development used in large scale studies compared to the Bayley-III, the Family Care Indicators, height-for-age, and stunting: A longitudinal study in Bogota, Colombia. *PloS One*, *15*(4), e0231317, https://doi.org/10.1371/journal.pone.0231317

Sayed, W.S., Noeman, A.M., Abdellatif, A., Abdelrazek, M., Badawy, M.G., Hamed, A., El-Tantawy, S. (2023). AI-based adaptive personalized content presentation and exercises navigation for an effective and engaging E-learning platform. *Multimedia Tools and Applications*, *82*(3), 3303–3333, https://doi.org/10.1007/s11042-022-13076-8

Spadafora, N., & Downes, T. (2020). Scaffolding in learning. *Encyclopedia of Evolutionary Psychological Science, T. Shackelford & V. Weekes-Shackelford, Ed. Switzerlang: Springer Nature Switzerland AG*, 1–4,

Spring, M., Faulconbridge, J., Sarwar, A. (2022). How information technology automates and augments processes: Insights from artificial-intelligence-based systems in professional service operations. *Journal of Operations Management*, *68*(6-7), 592–618, https://doi.org/https://doi.org/10.1002/joom.1215

Strachan, J.W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... others (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 1–11, https://doi.org/10.1038/s41562-024-01882-z

Tanana, M., Hallgren, K.A., Imel, Z.E., Atkins, D.C., Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment*, *65*, 43–50, https://doi.org/10.1016/j.jsat.2016.01.006

Tejeda, H., Kumar, A., Smyth, P., Steyvers, M. (2022). AI-Assisted Decision-making: a Cognitive Modeling Approach to Infer Latent Reliance Strategies. *Computational Brain & Behavior*, *5*(4), 491–508, https://doi.org/10.1007/s42113-022-00157-y

Tunstall, L., von Werra, L., Wolf, T. (2022). *Natural language processing with transformers: Building language applications with hugging face*. O'Reilly Media.

Vanbelle, S., Engelhart, C.H., Blix, E. (2024). A comprehensive guide to study the agreement and reliability of multi-observer ordinal data. *BMC Medical Research Methodology*, *24*, 310, https://doi.org/10.1186/s12874-024-02431-y Retrieved from https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-024-02431-y

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*, ,

Vilaseca, S., Rivero, M., Bersabe, R.M., Navarro-Pardo, E., Cantero, M.-J., Galera, N., ... Fathers, J. (2019). Spanish validation of the piccolo (parenting interactions with children: Checklist of observations linked to outcomes) with parents of 10- to 47-month-old children. *Frontiers in Psychology*, *10*, 680, https://doi.org/10.3389/fpsyg.2019.00680

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes* (1st ed.). Cambridge, MA: Harvard University Press. (Original work published 1930)

Ward, K. (2022). *Ecological influences of parental discipline behaviors and child outcomes among families in low-and middle-income countries* (Unpublished doctoral dissertation). University of Michigan.

Winkler, R., Hobert, S., Salovaara, A., Söllner, M., Leimeister, J.M. (2020). Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. *Proceedings of the 2020 chi conference on human factors in computing systems* (p. 1–14). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3313831.3376781

Wood, D., Bruner, J.S., Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100,

Xi, J., & Lantolf, J.P. (2021). Scaffolding and the zone of proximal development: A problematic relationship. *Journal for the Theory of Social Behaviour*, *51*(1), 25–48, https://doi.org/10.1111/jtsb.12260

Zhao, X., Feng, G.C., Ao, S.H., Liu, P.L. (2022). Interrater reliability estimators tested against true interrater reliabilities. *BMC Medical Research Methodology*, *22*(1), 232, https://doi.org/10.1186/s12874-022-01707-5

Appendix A PICCOLO Inference Mapping and Attempt Summaries

Table A1 PICCOLO Items by Inference Level (short English labels).

Low inference	**Medium inference**	**High inference**
Physical proximity	Warm tone	Fully engages
Watches while child speaks/-sounds	Smiles	Responds to emotions
	Praises	Flexible to changes
	Responds to words/sounds	Explains reasons
	Repeats/expands words	Suggests extensions
	Names objects/actions	Joins pretend play
	Asks for information	Guides step sequences
	Attends to child's activity	Explains features
	Follows child's lead	Adjusts pace/activity
	Shows affection	Uses positive expressions

English labels provide concise descriptors for readability; full PICCOLO definitions were used in scoring and analysis.

Table A2 IRR and nMSE by Attempt (script v7.5 with imputed values).

Attempt (CSV)	N **pairs**	**PA (%)**	κ	**nMSE (All)**	**nMSE (Low)**	**nMSE (High)**
`analisis_videos_piccolo_completo`	196	55.102	.096	1.255	0.950	1.340
`..._2`	110	49.091	.054	1.137	0.862	1.214
`..._3`	968	56.405	.143	1.276	1.149	1.365
`..._4`	510	60.196	.078	1.684	1.394	1.464
`..._5`	330	54.849	.183	2.214	1.403	2.405
`..._7`	554	48.376	.056	2.037	2.056	1.800
`..._9`	418	53.589	.193	1.770	1.188	2.208
`..._10`	352	46.023	.194	1.874	3.733	3.301
`..._11`	1,260	48.333	.221	1.445	1.210	1.382
`..._12`	286	44.755	.157	1.642	1.438	1.211
`..._13`	880	50.682	.216	1.463	1.138	1.505
final aggregate	3,102	56.867	.128	1.537	1.164	1.641

Raw MSE scaled by the aggregate normalization factor (≈ 2.46). Low/High derived from aggregate multipliers (Low $\approx 0.76\times$All; High $\approx 1.07\times$All).

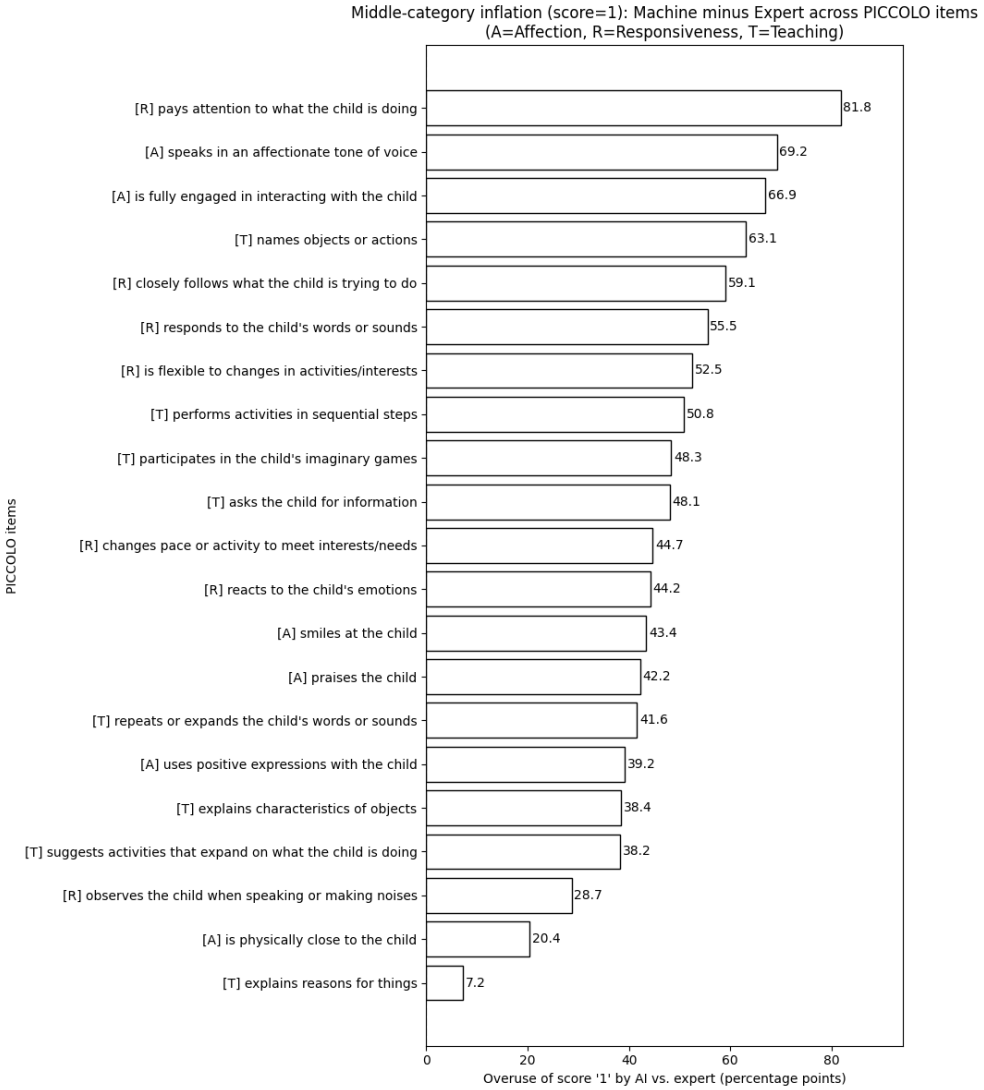

Fig. 2 Middle–category inflation across PICCOLO items. Bars show the percentage-point difference for score 1 (Barely) between the machine and the expert (Machine minus Expert) for 21/22 items. Labels prefix item names with domain initials (A: Affection; R: Responsiveness; E: Teaching). The distribution evidences systematic overuse of the middle category by the AI.