CAMA: A Culturally Adaptive Multi-Agent Framework for Postpartum Depression Support in Multilingual and Low-Resource Settings

Zhiqi ZHANG[1][0009−0004−5661−5052], Ziyi LIU[1][0009−0001−7876−4241], and Rite BO[2][0009−0005−1840−1309]

[1] Institut Polytechnique de Paris, France
{zhiqi.zhang, ziyi.liu}@ip-paris.fr
[2] Jilin University, China
bort24@mails.jlu.edu.cn

Abstract. Large language models (LLMs) enable scalable conversational support for postpartum depression (PPD), yet current systems do not sufficiently account for intra-lingual cultural variation, even within high-resource languages such as Chinese. Dialectal phrasing, local idioms, and culturally embedded expressions, such as Northeastern Mandarin "zhabayue de teng" (humorous discomfort) or Southern Min "xin-gua-a-tia^{n}" (deep sorrow), may lead to misinterpretation, safety-critical ambiguity, or emotionally inappropriate responses in PPD-related dialogues.

We introduce CAMA (Culturally Adaptive Multi-Agent Co-Design Framework), a lightweight framework for cultural sensitivity detection and response alignment that identifies dialect-specific linguistic cues and supplements LLMs with contextual socio-cultural grounding without performing clinical diagnosis. Our approach integrates culturally aware prompting and intervention logic to improve empathy, safety, relevance, and user trust.

This work highlights that cultural fairness in mental-health LLMs must consider intra-language diversity, not only cross-lingual disparity. CAMA provides a practical pathway towards culturally aligned, safe, and trustworthy mental-health dialogue systems.

Keywords: postpartum depression · large language models · cultural adaptation · Chinese dialects · multi-agent systems

1 Introduction

Large language models (LLMs) are increasingly integrated into mental health support and perinatal care, offering scalable, linguistically flexible, and accessible assistance for individuals who may lack timely clinical resources [10]. Postpartum depression (PPD) affects an estimated 10–20% of mothers worldwide [18], yet nearly half of affected individuals remain undiagnosed during routine care. Conversational systems powered by LLMs have the potential to lower barriers to

psychological support by providing empathetic and contextually aware interactions that complement the work of healthcare professionals. However, when these systems fail to reflect the user's linguistic, cultural, or emotional background, their responses risk misalignment, misunderstanding, or unintended psychological harm: concerns that are amplified in the sensitive context of maternal mental health.

Existing research on LLMs in postpartum mental health largely focuses on screening, diagnostic performance, and clinical accuracy. For example, [30] compared ChatGPT, Bard, and Google Search on PPD-related queries, finding that while ChatGPT produced more clinically accurate information, it often lacked emotional nuance. A systematic review of psychiatric LLMs [26] highlights persistent limitations in complex reasoning and empathy, while [2] identifies seven categories of LLM-enabled functionality in women's reproductive psychiatry, including diagnosis, patient education, and counselling. Complementarily, [9] introduced an interpretable real-time PPD detection framework combining NLP, ML, and LLM-based prediction. Collectively, these studies underscore growing attention to *clinical accuracy* but continued neglect of *cultural and empathetic adaptation*, which is crucial for effective maternal mental health conversations.

Meanwhile, the fields of NLP fairness and linguistic inclusion have achieved substantial progress in cross-lingual and low-resource settings. Tibetan NLP has benefited from resources such as TIB-STC [14], T-LLaMA [24], TiBERT [23], prompt-based Tibetan classification [35], and the multilingual evaluation dataset CUTE [42]. Research on Moroccan Arabic (Darija) has produced the MADAR parallel corpus [5], the Darija Open Dataset v2 [27], DarijaBERT [8], and adaptation frameworks such as LoResLM/Atlas-Chat [32]. Similar efforts for African languages, including Swahili, have been supported by MasakhaNER [1], the Lacuna Fund corpus [20], the Kenya NLP survey [3], and the Lanfrica repository [21, 7]. Together, these initiatives expand the coverage of NLP technologies and strengthen cross-lingual fairness.

Despite this progress, *intra-lingual* cultural and dialectal diversity in high-resource languages remains underexplored. The Sinitic language family exhibits substantial regional variation in phonology, lexical choices, expressive norms, and pragmatic strategies. These differences meaningfully shape how individuals articulate distress, seek support, and engage in emotionally charged dialogue. Prior work [37, 39, 16, 38] has demonstrated the modelling challenges posed by dialectal heterogeneity. Recent large-scale evaluations further show significant performance gaps when LLMs interact with non-Mandarin Chinese variants. For instance, [17] reports substantially higher misunderstanding rates for Cantonese tasks compared with Standard Mandarin, and [36] finds that although Chinese-developed LLMs perform strongly on Mandarin, they do not outperform Western models on non-Mandarin Sinitic and minority languages. These findings highlight limits in dialect generalisation and suggest that neglecting intra-lingual variation may lead to emotional mismatch or cultural dissonance: particularly harmful in high-stakes domains such as maternal mental health.

To address this gap, we introduce **CAMA (Culturally Adaptive Multi-Agent Co-Design Framework)**, a lightweight cultural adaptation layer designed to enhance LLM-based maternal care conversations. CAMA is not a diagnostic tool; rather, it focuses on improving the cultural, linguistic, and emotional relevance of model-generated responses. The system dynamically detects dialectal and regional features in the user input and loads an appropriate Chinese culture pack (e.g., Northeastern Mandarin, Cantonese, Southern Min (Minnan), Central Plains Mandarin, or Southwestern Mandarin varieties). It further coordinates five collaborative agents: *psychologist*, *linguist*, *educator*, *mother*, and *AI researcher*: to provide layered feedback on empathy, linguistic appropriateness, cultural alignment, and interpretability. Through the integration of cultural detection, dynamic prompt adaptation, and multi-agent validation, CAMA enables more coherent, trustworthy, and culturally sensitive interactions.

Beyond postpartum care, this work contributes to a broader vision of inclusive and socially grounded AI. Cultural adaptation represents an interpretable, human-centred enhancement layer for LLMs, offering a pathway toward more equitable systems that better serve linguistically diverse and historically underserved communities.

Our contributions are threefold:

- **Problem Identification**: We articulate the *intra-lingual low-resource challenge*: overlooked cultural and dialect diversity within high-resource languages such as Chinese: and demonstrate its crucial implications for empathetic alignment in PPD support.
- **Framework Design**: We introduce CAMA, which integrates dialect detection, dynamic cultural adaptation, and multi-agent feedback to improve the cultural sensitivity and interpretability of LLM-based maternal care dialogue systems.
- **Human-centred Advancement**: We position cultural adaptation as a lightweight, interpretable, and socially meaningful layer that strengthens fairness, trust, and inclusivity in AI for maternal mental health and beyond.

2 Related Work

2.1 Large Language Models in Mental Health

Large language models (LLMs) have increasingly been applied to mental health-related tasks, from symptom detection to conversational support. A comprehensive review by [10] surveyed the contributions of LLMs to mental health applications and identified key challenges in reliability, privacy, and interpretability. For instance, [30] evaluated the clinical accuracy of ChatGPT and Bard in answering postpartum depression (PPD) questions, demonstrating that while LLMs can outperform traditional search engines in correctness, they often lack affective sensitivity. Similarly, a review of LLMs in psychiatry emphasized their potential to assist in therapy and patient education but noted their limited performance on nuanced emotional reasoning [26].

In reproductive psychiatry, [2] explored multimodal LLMs that integrate textual and physiological data to support diagnosis and patient communication, while [9] proposed a real-time, explainable LLM-based framework for PPD detection that combines NLP and machine learning models. Collectively, these studies reveal that while LLMs show promise in psychological support, they are predominantly designed for **clinical screening or prediction**, with limited attention to cultural and linguistic adaptation, an essential factor for building empathy and trust in maternal mental health scenarios.

2.2 Fairness and Inclusion in Cross-Lingual NLP

Fairness in NLP has long been associated with addressing **low-resource and cross-lingual disparities**. Recent efforts in Tibetan include the creation of the TIB-STC corpus [14], the T-LLaMA model [24], and the TiBERT pre-trained model [23], which collectively enhance representation for this under-resourced linguistic community. Complementary advances, such as prompt-based Tibetan text classification [35] and the CUTE multilingual dataset [42], further demonstrate the growing ecosystem for Tibetan NLP.

Parallel developments in **Moroccan Arabic (Darija)** include the MADAR corpus [5], the Darija Open Dataset v2 [27], DarijaBERT [8], and the LoResLM / Atlas-Chat framework [32], which focus on bridging dialectal variations across Arabic-speaking regions. In **African language NLP**, resources such as Masakha-NER [1], the Lacuna Fund Swahili corpus [20], the Kenya NLP Survey [3], and Lanfrica [21, 7] have been instrumental in improving linguistic coverage and model fairness across underrepresented communities.

Despite these advances, most fairness work is evaluated on cross-language gaps or on standardised language varieties, and less frequently examines **within-language** pragmatic and cultural variation in high-stakes dialogue. In this paper, we focus on a practical **within-language** deployment setting: culturally grounded prompting for major Sinitic dialect communities in postpartum support dialogues, where shared written Chinese coexists with region-specific pragmatics and idiomatic expression.

2.3 Chinese Dialects and Intra-Linguistic Diversity

Chinese presents a particularly rich case of **intra-linguistic cultural diversity**, as its dialects differ not only phonetically but also in pragmatics, emotional expression, and social norms. For example, [37] demonstrated the challenges of dialect discrimination under low-resource conditions, while [39] proposed a dialect-to-speech generation framework that exposed the burden of adapting standard Mandarin models to dialectal text. Furthermore, [16] introduced a phonemic annotation method for Chinese dialects, highlighting the need for structured linguistic resources. A comparative analysis by [38] of LLM and self-supervised models for dialectal speech recognition revealed performance degradation as dialectal variance increases.

However, these studies focus primarily on **linguistic modeling** rather than **socio-cultural alignment**. In emotionally sensitive contexts such as maternal counseling, where implicit empathy, local metaphors, and communicative norms play central roles, such internal diversity directly affects user trust and well-being. Addressing this gap requires models that integrate **cultural adaptation and linguistic variation**, forming the motivation for our proposed **CAMA (Culturally Adaptive Multi-Agent Co-Design Framework)**, which extends LLM fairness research from *cross-lingual* to *intra-linguistic* inclusivity.

Our work is analogous but focuses on intra-linguistic low-resource dialects (Chinese regional varieties) for postpartum support, rather than cross-lingual gaps.

2.4 Co-Design

Co-design extends participatory design by foregrounding the "collective creativity" of designers and non-designers as equal partners, rather than treating users as passive sources of requirements or testers at the end of development [34]. In digital mental health, recent reviews conclude that involving service users in design tends to improve acceptability, usability, and engagement, yet co-design is often only partially implemented, under-reported, and constrained by resources and power imbalances, especially with youth and clinically vulnerable groups [19, 29]. For chatbots and digital conversational tools, participatory studies further show that user and clinician involvement is critical to align expectations around empathy, transparency, and conversational boundaries, but most deployed systems remain largely expert-driven with limited end-user input [13, 6, 12].

In AI, participatory and community-engaged frameworks argue that co-design should reach beyond the interface to problem formulation, data practices, and governance, warning that superficial consultation can reproduce existing harms [4]. Practitioner guidance similarly emphasises sustained partnerships with affected communities when deploying high-stakes AI systems, but current work largely offers high-level process models and rarely treats intra-linguistic cultural diversity (e.g., dialectal variation within one language) as a primary design dimension [28]. More recently, LLM-based multi-agent systems have been proposed as co-design partners that support human ideation and critique, but mainly for generic creative or engineering tasks without culturally grounded roles in sensitive health settings [33]. In contrast, CAMA operationalises co-design as an internal multi-agent governance layer: Psychologist, Linguist, Teacher, Mother, and AI Researcher agents act as always-on "virtual stakeholders" that approximate a continuous co-design loop at inference time, complementing, rather than replacing, conventional co-design with postpartum women and clinicians, and explicitly centring intra-linguistic cultural diversity as a core axis of system behaviour.

3 Methodology

3.1 Framework Overview

The proposed **CAMA (Culturally Adaptive Multi-Agent Co-Design Framework)** enhances large language models' (LLMs) capacity for culturally aligned and emotionally safe interaction in postpartum mental health contexts. Rather than serving as a diagnostic model, CAMA functions as a *multi-agent adaptation layer* that dynamically adjusts linguistic tone, empathy, and communicative style according to users' dialectal and cultural backgrounds.

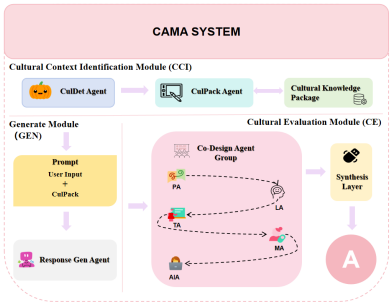


Fig. 1: Overview of the proposed **CAMA (Culturally Adaptive Multi-Agent Co-Design Framework)**.

The system integrates five specialized agents (**Cultural Detection**, **Culture Pack**, **Response Generation**, **Co-Design**) arranged in a cascaded pipeline. Each agent contributes distinct functions: detecting users' linguistic and cultural context, loading corresponding culture packs, generating culturally adapted responses, iteratively refining them through multi-agent feedback, and synthesizing a final empathetic, culturally coherent output.

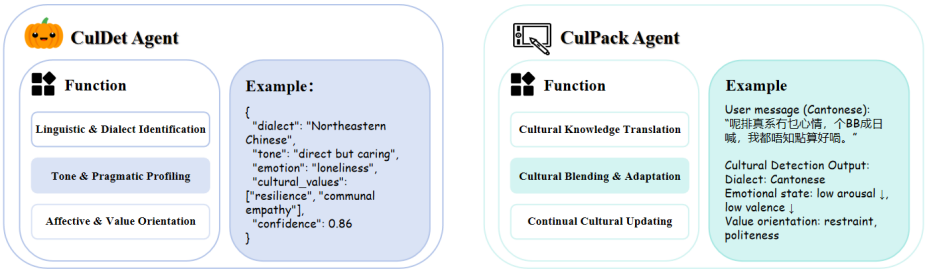
As illustrated in Figure 1, CAMA integrates five cooperative agents arranged in a cascaded pipeline:

1. the **Cultural Detection Agent (Figure 2a)**, which analyses user input to infer dialectal, linguistic, and affective cues and identifies the user's probable cultural region;
2. the **Culture Pack Agent (Figure 2b)**, which retrieves and activates a region-specific *Culture Pack* containing idiomatic expressions, tone templates, and communicative norms relevant to the detected culture;
3. the **Response Generation Agent**, which produces an initial draft reply conditioned on both the user query and the loaded cultural context;
4. the **Co-Design Agent Group**, where five role-inspired agents (Psychologist, Linguist, Teacher, Mother, and AI Researcher) collaboratively evaluate and refine the generated response through iterative feedback; and

Through this structured collaboration, CAMA transforms conventional LLM dialogue into *culturally aware, empathetic communication*, bridging the gap between model generalization and local specificity. It establishes a lightweight yet interpretable pathway toward trustworthy, inclusive AI for maternal mental-health support.

3.2 Cultural Detection Agent

The agent operates through a two-stage pipeline. In the first stage, it performs *linguistic grounding* by leveraging few-shot prompting of a large language model



(a) Cultural Detection Agent. (b) Cultural Pack Agent.

Fig. 2: Overview of the cultural agents in CAMA.

(LLM) to identify dialectal markers, region-specific lexical items, and syntactic variations. This approach enables classification across five Chinese dialect clusters, Northeastern Mandarin, Cantonese, Southern Min, Central Plains Mandarin, and Southwestern Mandarin, without relying on any fine-tuned external encoders. In the second stage, it infers a cultural context profile from socio-pragmatic cues via zero-shot prompting.

In interactive use, the agent conditions its inference on available dialogue history rather than a single utterance, and the dialect estimate is updated as new turns arrive. In our user study, the first downstream use of the dialect estimate occurred only after several user turns (typically 3–4), and it was refreshed on each subsequent turn.

The extracted attributes form a structured *cultural context profile*, including tone formality, emotional framing, and implicit value orientation (e.g., collectivism, emotional restraint, humour). This profile is communicated to the downstream *Culture Pack Agent*, guiding the retrieval of region-specific communicative norms and helping ensure that response generation remains culturally congruent, contextually empathetic, and emotionally safe.

3.3 Culture Pack Agent

The **Culture Pack (CulPack) Agent** serves as a cultural knowledge interface that converts the abstract cultural representations identified by the **Cultural Detection Agent** into actionable linguistic and pragmatic cues. These cues directly guide the **Response Generation Agent**, ensuring that the generated text follows regional discourse norms and culturally grounded communication styles.

Within the **CAMA** framework, we build **five core Chinese Culture Packs** representing major regional and sociolinguistic communities: **Northeastern Mandarin**, **Cantonese**, **Southern Min**, **Central Plains Mandarin**, and **Southwestern Mandarin** (Yunnan-Guizhou-Sichuan) Chinese. These packs capture not only substantial **dialectal and lexical variations** but also distinct **emotional expressions and communicative norms** characteristic of each region.

Concretely, the **Northeastern Mandarin** pack emphasizes directness and humour, expressing empathy through casual banter; the **Cantonese** pack favors politeness and restraint, showing care through considerate wording; the **Southern Min** pack reflects warmth and family orientation in a gentle tone; the **Central Plains Mandarin** pack highlights sincerity and grounded realism; and the **Southwestern Mandarin** pack conveys optimism and comfort through relaxed, talkative interaction.

To ensure both representativeness and manageability, each Culture Pack follows a unified structure with four hierarchical layers:

1. **Lexical Layer**: a curated lexicon of dialect-specific vocabulary, colloquial phrases, and regionally preferred expressions.
2. **Emotional Idiom Layer**: figurative language, proverbs, and idiomatic expressions that convey culturally nuanced emotional states.
3. **Tone Template Layer**: prototypical templates capturing stylistic and emotional response patterns typical of the dialectal community.
4. **Pragmatic Rule Layer**: communicative norms governing politeness strategies, indirectness levels, and culturally appropriate emotional disclosures.

Each Culture Pack contains approximately **110 structured entries** on average (around 70 lexical items, 26 emotional idioms, 4–5 tone templates, and 9 pragmatic rules), totaling roughly **550 culturally grounded items** across all five dialects (see Table 1). This lightweight yet comprehensive configuration enables efficient few-shot adaptation and prompt integration under low-resource conditions, while providing a stable and interpretable cultural foundation for downstream generation tasks.

Although each Culture Pack represents a distinct linguistic community, real-world users often exhibit **mixed dialectal and cultural cues**. To accommodate such cultural blending, the **CulPack Agent** performs **embedding-based retrieval and fusion** over all available Culture Packs. Each pack (P_i) is represented by a meta-descriptor vector ($v_i = f_{GTE}(P_i)$), where $f_{GTE}(\cdot)$ denotes the **GTE embedding model**. Given a user's *Cultural Context Profile* (C_u) from the **Cultural Detection Agent**, its embedding vector is $v_u = f_{GTE}(C_u)$. The CulPack Agent computes cosine similarities between v_u and each v_i, followed by a softmax-based weighting mechanism:

$$w_i = \text{softmax}\left(\frac{\cos(v_u, v_i)}{\alpha}\right),$$

where α controls the sharpness of selection and N is the total number of Culture Packs. The final **hybrid cultural representation** is then obtained as:

$$v_{hyb} = \sum_{i=1}^{N} w_i v_i.$$

This fused vector integrates lexical, tonal, and pragmatic traits across multiple dialectal communities, forming a probabilistic blend that mirrors the user's

Table 1: Overview of Stylistic and Emotional Characteristics of the Five Chinese Culture Packs in CAMA

Dialect	Tone & Style Characteristics	Typical Emotional Expression	Cultural Keywords
NEM	Direct, humorous; strong colloquial tone	Reassurance through playful banter; emotional openness	Humour; resilience; community
CAN	Polite and reserved; caring, socially attentive	Empathy via courtesy; understated emotional warmth	Politeness; dignity; family care
MIN	Gentle, vivid imagery; family-oriented tone	Indirect affection; warm, tender reassurance	Gentleness; harmony; patience
CEN	Honest, straightforward; pragmatic and grounded	Empathy through realism; shared hardship and support	Simplicity; perseverance; trust
SWM	Relaxed and talkative; optimistic, upbeat tone	Casual emotional relief; lighthearted reassurance	Optimism; sociability; humour

Abbreviations: NEM = Northeastern Mandarin; CAN = Cantonese (Yue Chinese); MIN = Southern Min (Minnan); CEN = Central Plains Mandarin (Zhongyuan Mandarin, e.g., Henan region); SWM = Southwestern Mandarin (e.g., Sichuan–Chongqing–Guizhou–Yunnan region).

communicative diversity. The resulting representation is serialized into a structured prompt segment, which conditions the **Response Generation Agent** to produce linguistically authentic, empathetic, and culturally consistent outputs.

To ensure the robustness of this fusion mechanism under **low-resource and cold-start conditions**, each Culture Pack is **initialized through few-shot synthesis combined with expert-in-the-loop refinement**. A small set of seed examples (dialectal expressions, emotional idioms, and tone descriptors) is first collected from linguistic corpora and social media data, and then expanded via **LLM-assisted controlled generation**. Human reviewers validate all entries to remove synthetic artifacts and ensure cultural appropriateness. Over time, the **CulPack Agent** supports **incremental updates**, incorporating newly observed expressions through confidence-weighted adaptation. This continual refinement enables CAMA to remain both **culturally grounded and dynamically adaptive** without relying on large-scale annotated corpora.

3.4 Response Generation Agent

The **Response Generation (Response Gen) Agent** serves as the dialogue engine of CAMA, producing culturally aligned and emotionally grounded re-

sponses **without any task-specific fine-tuning**. Given the hybrid cultural representation (P_{hybrid}) from the CulPack Agent and the user query (x), the Response Gen **Agent** constructs a **two-layer prompt**:

1. A *system layer* embedding cultural constraints (including tone markers, dialectal exemplars, taboo lists, and pragmatic rules) derived from (P_{hybrid}); and
2. A *task layer* encoding the user's intent and situational context.

The underlying LLM then generates an initial draft response ($y^{(0)}$) conditioned on both layers:

$$y^{(0)} = \text{LLM}\Big(\text{System}(P_{\text{hybrid}}), \text{Task}(x), \theta_{\text{ctrl}}\Big),$$

where (θ_{ctrl}) denotes decoding parameters (e.g., temperature, top-p) tuned for stable cultural style expression.

To ensure **effective validity under maternal mental-health norms**, the Response Gen **Agent** performs a **zero-training affective control** procedure. The model self-evaluates ($y^{(0)}$) via a few-shot rubric to estimate affective scores of Stylistic and Emotional Characteristics of the Five Chinese Culture Packs in CAMA ($\hat{\mathbf{e}} = [\hat{v}, \hat{a}, \hat{\epsilon}] \in [0,1]^3$) for *valence*, *arousal*, and *empathy*.

Dialect-specific target ranges from (P_{hybrid}) are denoted as ($\mathcal{T} = [v_\ell, v_u] \times [a_\ell, a_u] \times [\epsilon_\ell, 1]$).

If ($\hat{\mathbf{e}} \notin \mathcal{T}$), the Response Gen **Agent** triggers a **style-preserving constrained rewrite** (without modifying model weights) to obtain ($y^{(1)}$):

$$y^{(1)} = \text{LLM}\Big(\text{Rewrite}(y^{(0)}, \text{targets} = \mathcal{T}, \text{dialect} = d)\Big),$$

otherwise ($y^{(1)} = y^{(0)}$).

This single-pass revision preserves factual content while enforcing culturally appropriate warmth and emotional intensity.

Next, the Response Gen **Agent** performs **safety filtering** using a hybrid rule-and-LLM mechanism, still **zero-trained**.

A deterministic rule layer screens ($y^{(1)}$) for medical directives, absolute assurances, and culture-specific taboos sourced from (P_{hybrid}). When triggered, or when uncertainty persists, the system requests a short risk classification

$$r \in \{\text{self-harm}, \text{medical}, \text{toxicity}, \text{none}\}.$$

Mitigation is template-driven: self-harm routes to crisis-support language; medical risk invokes directive removal and professional-help disclaimers; cultural or tonal violations prompt a taboo-aware rewrite. The final output is (y^*).

This **training-free pipeline** translates cultural knowledge into actionable generation constraints, yielding responses that are **linguistically authentic, affectively calibrated, and clinically cautious**, while maintaining minimal data and computational requirements.

For deployment efficiency, CAMA generates the user-facing response in a **single-pass** Response Gen **Agent** for real-time interaction. Extended co-design rationales and periodic ethics audits operate *asynchronously* (in *light-tag* mode online and *full* mode offline), ensuring **zero latency impact** for end-users.

This design allows the system to maintain **cultural sensitivity, affective reliability, and ethical accountability** in practice, while remaining efficient for everyday mental-health dialogue applications.

3.5 Co-Design Agent Group (CDAG)

The Co-Design Agent Group (CDAG) constitutes the participatory feedback core of the CAMA framework, integrating interdisciplinary expertise to evaluate and refine the model's cultural, emotional, and ethical alignment.

Rather than relying on a single evaluative channel, CDAG orchestrates five symbolic reasoning agents (the Psychologist (PA), Linguist (LA), Teacher (TA), Mother (MA), and AI Researcher (AIA)), each representing a distinct stakeholder in maternal mental-health communication.

Together, they convert qualitative human insights into structured evaluative signals, enhancing fairness, trustworthiness, and interpretability without any additional model training.

Each agent focuses on a dedicated evaluative dimension.

- The Psychologist Agent assesses emotional safety, empathy, and affective appropriateness, flagging replies that may induce distress or reinforce stigma.
- The Linguist Agent verifies dialectal authenticity, idiomatic coherence, and pragmatic consistency against the active Culture Pack.
- The Teacher Agent evaluates clarity, accessibility, and pedagogical tone to ensure inclusivity across literacy levels.
- The Mother Agent represents the end-user perspective, gauging emotional resonance, comfort, and sincerity.
- Finally, the AI Researcher Agent supervises logical soundness, fairness across dialectal groups, and adherence to ethical-alignment principles.

Together, these five agents span the principal dimensions of cultural fidelity, emotional safety, and ethical compliance within maternal-mental-health dialogues.

Their concrete evaluation metrics (covering linguistic, affective, and fairness-related subdimensions) are detailed in the Evaluation and Results section, which formalizes how qualitative judgments are operationalized into quantifiable scores.

Formally, each agent outputs an evaluation tuple:

$$f_i = (s_i, c_i),$$

where $s_i \in [0, 1]$ denotes a confidence score and $c_i \in \{0, 1\}$ indicates binary acceptance.

The combined vector

$$\mathbf{f} = [f_P, f_L, f_T, f_M, f_A]$$

summarizes multidimensional judgments spanning empathy, authenticity, clarity, cultural safety, and fairness.

CDAG operates synchronously with the **Response Generation Agent** within a single decoding pass. Role-specific self-assessment instructions are appended to the system prompt, guiding the LLM to output both the final reply and structured metadata in JSON format [31]. This **single-pass self-evaluation** design introduces negligible latency, as the role-based tags are produced within the same decoding stream rather than through a secondary inference step [41, 25].

When disagreement or low confidence ($s_i < 0.6$) arises, the interaction is stored in a *Co-Design Memory Pool* for offline expert review. Domain specialists corresponding to each agent type annotate linguistic misalignment or emotional risks, and these annotations are subsequently used to refine Culture Pack entries and affective target ranges.

Over successive iterations, the relative influence of each agent dimension is rebalanced using a lightweight update rule:

$$w_i(t+1) = w_i(t) + \eta(s_i - 0.5),$$

allowing the framework to evolve through continuous human-in-the-loop calibration without retraining.

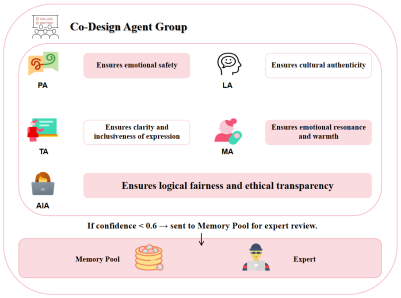


Fig. 3: Co-Design Agent Group (CDAG) structure and workflow.

An internal synthesis layer aggregates feedback from all five agents (Psychologist, Linguist, Teacher, Mother, and AI Researcher) and consolidates their judgments into a unified decision vector. This ensures that the final response delivered to the user is culturally coherent, emotionally appropriate, and ethically verified.

By embedding multi-perspective evaluation directly into the generation loop, the Co-Design Agent Group (Figure 3) transforms diverse human expertise into a living governance mechanism, reinforcing CAMA's cultural sensitivity, affective reliability, and ethical transparency.

Through these mechanisms, CAMA embodies a practical model of responsible and culturally grounded AI for maternal mental health support.

4 Evaluation and Results

4.1 Evaluation Setup

We conducted a controlled user study with **25 postpartum women** (mean age = 29.4 ± 3.1 years; 40% exhibiting mild PPD symptoms) to evaluate **empathy, trust, cultural fit, safety, and overall user experience**. Participants originated from five dialectal regions (**Northeastern Mandarin**, **Cantonese**,

Table 2: Dialect identification accuracy across five regional Chinese dialect clusters.

Dialect Cluster	Samples	Accuracy (%)
Northeastern Mandarin	94	94.7
Cantonese	85	98.9
Southern Min	86	72.1
Central Plains Mandarin	50	90.0
Southwestern Mandarin	98	87.0
Overall	413	**88.5**

Southern Min, **Central Plains Mandarin**, **Southwestern Mandarin**) and engaged with two systems: (i) **CAMA** (our multi-agent framework), and (ii) a **Mandarin-only baseline**. Each participant completed three scenarios (*emotional support*, *informational guidance*, *family communication*).

CAMA was implemented as a **multi-agent coordination framework** operating on the **DeepSeek-V3.1 API** [22]. Five role-specialized agents (*Psychologist, Linguist, Teacher, Mother,* and *AI Researcher*) contributed role-conditioned responses via adaptive aggregation. The baseline used the same API in a single-agent configuration. Decoding parameters were fixed across systems (temperature = 0.7, max_tokens = 512). A supplementary validation with **GPT-4o** [15] demonstrated *model-agnostic generality*, revealing comparable gains in empathy (+18%) and cultural fit (+31%).

All procedures followed low-risk mental-health research protocols. Participants provided written consent and were reminded that the system is *non-diagnostic*. Sessions containing self-harm cues were terminated automatically and accompanied by hotline information.

4.2 Metrics

Human-rated metrics spanned seven 5-point Likert dimensions: *Empathy, Comfort, Clarity, Cultural Fit, Trust, Safety,* and *Helpfulness*.

Automatic metrics included: Dialectal Authenticity (BERTScore against a regional corpus using Chinese RoBERTa-large [40]), Idiomatic Accuracy (frequency-weighted idiom matching with expert validation), Toxicity Probability (Detoxify-multilingual [11]), NLI Consistency (entailment ratio via a RoBERTa NLI classifier), and Response Latency (mean generation time).

We further report a composite quality score that aggregates five subjective dimensions of response quality. Let $f_1, \ldots, f_5 \in [0, 1]$ denote the normalised scores for *Empathy*, *Cultural Fit*, *Trust*, *Comfort*, and *Helpfulness*, respectively.[3]

[3] All 5-point Likert ratings are linearly rescaled to $[0, 1]$.

We define

$$C_{\text{score}} = \sum_{i=1}^{5} w_i f_i$$

where the weights $\mathbf{w} = [0.25, 0.20, 0.15, 0.25, 0.15]$ assign higher importance to perceived empathy and comfort, reflecting clinical guidance on therapeutic alliance in postpartum mental health support.

4.3 Experimental Design

We adopted a **within-subject** design with counterbalanced system order. Each 10-minute dialogue consisted of 8–10 turns. After each scenario, participants completed ratings; system logs recorded latency, toxicity, and dialectal features. All data were anonymized for quantitative and qualitative analysis.

Table 3: Comparison between Baseline and CAMA framework.

Metric	**Baseline**	**CAMA**	$\Delta\%$ / d	**p-value**
Empathy	3.47 ± 0.52	**4.21 ± 0.38**	+21 / 1.38	***
Cultural Fit	2.89 ± 0.63	**4.03 ± 0.45**	+39 / 1.42	***
Trust	3.59 ± 0.48	**4.12 ± 0.41**	+15 / 1.07	**
Comfort	3.61 ± 0.46	**4.18 ± 0.44**	+16	**
Helpfulness	3.52 ± 0.51	**4.05 ± 0.42**	+15	*
BERTScore	0.71 ± 0.04	**0.79 ± 0.03**	+11	**
Ethical Compliance (%)	89.7 ± 4.2	**96.4 ± 2.8**	+7	*
Toxicity (%) ↓	6.3 ± 2.1	**2.8 ± 1.4**	-56 / 1.10	*
Latency (s)	2.9 ± 0.4	**3.2 ± 0.5**	+0.3	n.s.
C_{score}	0.56 ± 0.07	**0.72 ± 0.05**	+29	***

(* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; paired t-test.)

4.4 Dialect Identification Accuracy

Before the user study, we evaluated the intrinsic performance of the Linguistic Grounding module on a curated set of 413 dialectal utterances covering five regional varieties. As shown in Table 2, the module achieved an overall accuracy of **88.5%**, with strongest performance on Cantonese (98.9%) and Northeastern Mandarin (94.7%). Accuracy was lower for Southern Min (72.1%) and moderate for Central Plains Mandarin (90.0%) and Southwestern Mandarin (87.0%), reflecting varying degrees of overlap with written Mandarin and orthographic variation. The main limitation lies in varieties with weak written conventions, where additional multimodal or phonological cues may further enhance robustness.

4.5 Results

Normality of paired differences was verified with the Shapiro–Wilk test prior to applying paired t-tests. As shown in Table 3, CAMA consistently outperformed the Mandarin-only baseline across both subjective and automatic metrics ($p < 0.05$). Improvements were most pronounced in *Empathy* (+21%) and *Cultural Fit* (+39%), both with large effect sizes ($d = 1.38$ and 1.42). Trust, Comfort, and Helpfulness also improved by 15–16% ($p < 0.05$), indicating stronger emotional resonance and user engagement.

On automatic metrics, CAMA achieved higher BERTScore (+11%) and better ethical compliance (+7%), while reducing toxicity probability by more than half (6.3% $\rightarrow$ 2.8%; $p < 0.05$). Latency differences were not statistically significant ($p > 0.05$), indicating that higher quality did not incur additional computational cost.

Overall, these findings demonstrate that CAMA delivers **emotionally aligned, culturally inclusive, and ethically reliable** dialogue generation for postpartum support.

5 Conclusion

We present CAMA (Culturally Adaptive Multi-Agent Co-Design Framework), a framework for culturally and emotionally aligned LLM-based maternal mental health support. Unlike prior work centered on clinical accuracy [30, 26, 2, 9], CAMA addresses the overlooked challenge of *intra-linguistic cultural adaptation*. Through dialect-aware detection, dynamically loaded Chinese Culture Packs, and multi-agent expert collaboration, our system transforms generic assistance into *contextually grounded, empathetic dialogue*.

Our results show that cultural adaptability is essential for building *trustworthy and emotionally coherent* AI in maternal care. Beyond postpartum depression, CAMA offers a generalizable blueprint for developing **inclusive, interpretable, and culturally sustainable** language technologies.

6 Limitations and Future Work

While CAMA advances culturally adaptive and empathetic dialogue generation, several limitations remain. The current *Culture Packs* cover only a subset of Chinese dialects, limiting cultural granularity; future work will broaden this coverage and include minority languages.

The multi-agent feedback mechanism adds computational overhead and depends on reliable agent coordination. In addition, our evaluation focuses on controlled postpartum scenarios, and real-world longitudinal studies are still required.

We plan to expand linguistic and multimodal resources, optimize inter-agent collaboration, and extend CAMA to other empathy-intensive domains such as education and counseling, advancing toward a more scalable and culturally inclusive AI framework.

References

1. Adelani, D.I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., et al.: Masakhaner: Named entity recognition for african languages. Transactions of the Association for Computational Linguistics **9**, 1116–1131 (2021). https://doi.org/10.1162/tacl_a_00416
2. AlSaad, R., Youssef, A., Kashani, S., AlAbdulla, M., Abd-Alrazaq, A., Khaled, S.M., Ahmed, A., Sheikh, J.: Multimodal large language models for women's reproductive mental health. Archives of Women's Mental Health pp. 1–16 (2025)
3. Amol, C.J., Chimoto, E.A., Gesicho, R.D., Gitau, A.M., Etori, N.A., Kinyanjui, C., Ndung'u, S., Moruye, L., Ooko, S.O., Kitonga, K., Muhia, B., Gitau, C., Ndolo, A., Wanzare, L.D.A., Kahira, A.N., Tombe, R.: State of nlp in kenya: A survey (2024), https://arxiv.org/abs/2410.09948
4. Birhane, A., Isaac, W.S., Prabhakaran, V., Díaz, M., Elish, M.C., Gabriel, I., Mohamed, S.: Power to the people? opportunities and challenges for participatory AI. In: Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms,Mechanisms,and Optimization (EAAMO '22) (2022). https://doi.org/10.1145/3551624.3555290
5. Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Diab, M.: The MADAR arabic dialect corpus and lexicon. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
6. Casu, M., Triscari, S., Battiato, S., Guarnera, L., Caponnetto, P.: Ai chatbots for mental health: A scoping review of effectiveness,feasibility,and applications. Applied Sciences **14**(13), 5889 (2024). https://doi.org/10.3390/app14135889
7. Emezue, C.C., Dossou, B.F.P.: Lanfrica: A participatory approach to documenting machine translation research on african languages (2020)
8. Gaanoun, K., Naira, A.M., Allak, A., Benelallam, I.: Darijabert: A step forward in nlp for the written moroccan dialect. International Journal of Data Science and Analytics (2024). https://doi.org/10.1007/s41060-023-00498-2, published 23 Jan 2024
9. García-Méndez, S., de Arriba-Pérez, F.: Detecting and explaining postpartum depression in real-time with generative artificial intelligence (2025)
10. Guo, Z., Lai, A., Thygesen, J.H., Farrington, J., Keen, T., Li, K.: Large language models for mental health applications: Systematic review. JMIR Mental Health **11**, e57400 (2024). https://doi.org/10.2196/57400
11. Hanu, L., AI, U.: Detoxify: Toxic comment classification. https://github.com/unitaryai/detoxify (2020), accessed: 2025-11-13
12. Hoffman, B.D., Oppert, M.L., Owen, M.: Understanding young adults' attitudes towards using AI chatbots for psychotherapy: The role of self-stigma. Computers in Human Behavior: Artificial Humans **2**, 100086 (2024). https://doi.org/10.1016/j.chbah.2024.100086
13. Houben, M., Van As, N., Sawhney, N., Unbehaun, D., Lee, M.: Participatory design for whom? designing conversational user interfaces for sensitive settings and vulnerable populations. In: Proceedings of the International Conference on Conversational User Interfaces (CUI '23). pp. 1–4 (2023). https://doi.org/10.1145/3571884.3597439
14. Huang, C., Gao, F., Liu, Y., Tashi, N., Wang, X., Tsering, T., Ma-bao, B., Luosang, R.D.G., Dongrub, R., Tashi, D., Feng, X., Wang, H., Yu, Y.: Tib-stc: A large-scale structured tibetan benchmark for low-resource language modeling (2025), https://arxiv.org/abs/2503.18288

15. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
16. Jiang, B., Dong, Q., Liu, G.: A method of phonemic annotation for chinese dialects based on a deep learning model with adaptive temporal attention and a feature disentangling structure. Computer Speech & Language **86**, 101624 (2024). https://doi.org/10.1016/j.csl.2024.101624
17. Jiang, J., Chen, P., Chen, L., Wang, S., Bao, Q., Kong, L., Li, Y., Wu, C.: How well do llms handle cantonese? benchmarking cantonese capabilities of large language models. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 4464–4505 (2025)
18. Khamidullina, Z., Marat, A., Muratbekova, S., Mustapayeva, N.M., Chingayeva, G.N., Shepetov, A.M., Ibatova, S.S., Terzic, M., Aimagambetova, G.: Postpartum depression epidemiology,risk factors,diagnosis,and management: An appraisal of the current knowledge and future perspectives. Journal of Clinical Medicine **14**(7), 2418 (2025)
19. Kilfoy, A., Hsu, C., Stockton-Powdrell, C., Whelan, P., Chu, C.H., Jibb, L.: An umbrella review on how digital health intervention co-design is conducted and described. npj Digital Medicine (2024). https://doi.org/10.1038/s41746-024-01385-1
20. Lacuna Fund: Lacuna fund language datasets: Kentrans (swahilidholuo/luhya) parallel corpora. https://lacunafund.org/datasets/language/index.html (2023), accessed: 2025-11-10
21. Lanfrica: Lanfrica: African language resources and datasets repository. https://lanfrica.com/ (2025), accessed: 2025-11-10
22. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
23. Liu, S., Deng, J., Sun, Y., Zhao, X.: Tibert: Tibetan pre-trained language model. In: 2022 IEEE International Conference on Systems,Man,and Cybernetics (SMC). pp. 2956–2961. IEEE (2022). https://doi.org/10.1109/SMC53654.2022.9945074
24. Lv, H., Zhang, Q., Shen, J., Wang, T., Li, H.: T-llama: A tibetan large language model based on llama2. Complex & Intelligent Systems (2025). https://doi.org/10.1007/s40747-024-01641-7, online First
25. Manakul, P., Liusie, A., Gales, M.: Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 9004–9017 (2023)
26. Omar, M., Soffer, S., Charney, A.W., Landi, I., Nadkarni, G.N., Klang, E.: Applications of large language models in psychiatry: A systematic review. Frontiers in Psychiatry **15**, 1422807 (2024). https://doi.org/10.3389/fpsyt.2024.1422807
27. Outchakoucht, A., Es-Samaali, H.: The evolution of darija open dataset: Introducing version 2 (2024), https://arxiv.org/abs/2405.13016
28. Partnership on AI: Guidance for inclusive AI: Practicing participatory engagement. https://partnershiponai.org/guidance-for-inclusive-ai/ (2025), framework for participatory public engagement in commercial AI development
29. Porche, M.V., Folk, J.B., Tolou-Shams, M., Fortuna, L.R.: Researchers' perspectives on digital mental health intervention co-design with marginalized community stakeholder youth and families. Frontiers in Psychiatry **13**, 867460 (2022). https://doi.org/10.3389/fpsyt.2022.867460

30. Sezgin, E., Singletary, L., Lin, S.: Clinical accuracy of large language models and google search responses to postpartum depression questions: A cross-sectional study. Journal of Medical Internet Research **25**, e49240 (2023). https://doi.org/10.2196/49240
31. Shanahan, M., McDonell, K., Reynolds, L.: Role-play with large language models (2023), https://arxiv.org/abs/2305.16367
32. Shang, G., Abdine, H., Khoubrane, Y., Mohamed, A., Abbahaddou, Y., Ennadir, S., Momayiz, I., Ren, X., Moulines, E., Nakov, P., Vazirgiannis, M., Xing, E.: Atlas-chat: Adapting large language models for low-resource moroccan arabic dialect (2024), presented at LoResLM 2025 Workshop
33. Sun, C., Huang, S., Pompili, D.: Llm-based multi-agent decision-making: Challenges and future directions. IEEE Robotics and Automation Letters (2025)
34. Vial, S., Boudhraâ, S., Dumont, M.: Human-centered design approaches in digital mental health interventions: Exploratory mapping review. JMIR Mental Health **9**(6), e35591 (2022). https://doi.org/10.2196/35591
35. Wang, R., Norbu, C., Li, X., Zhao, Q.: Prompt-based for low-resource tibetan text classification. ACM Transactions on Asian and Low-Resource Language Information Processing (2023). https://doi.org/10.1145/3603168
36. Wen-Yi, A.W., Jo, U.E.S., Mimno, D.: Do chinese models speak chinese languages? (2025), https://arxiv.org/abs/2504.00289
37. Xu, F., Dan, Y., Yan, K., Ma, Y., Wang, M.: Low-resource language discrimination toward chinese dialects. In: Proceedings of the 29th ACM International Conference on Multimedia (MM 2021) (2021). https://doi.org/10.1145/3473499
38. Xu, T., Chen, H., Qing, W., Lv, H., Kang, J., Li, J., Lin, Z., Li, Y., Xie, L.: Leveraging llm and self-supervised training models for speech recognition in chinese dialects: A comparative analysis (2025)
39. Zhang, J., Li, H., Zhao, W.: A novel chinese dialect tts frontend with non-autoregressive neural machine translation (2022)
40. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: Proceedings of the International Conference on Learning Representations (ICLR) (2020), https://openreview.net/forum?id=SkeHuCVFDr, poster version; OpenReview pre-print
41. Zhang, Y., Khalifa, M., Logeswaran, L., Kim, J., Lee, M., Lee, H., Wang, L.: Small language models need strong verifiers to self-correct reasoning (2024), https://arxiv.org/abs/2404.17140
42. Zhuang, W., Sun, Y.: Cute: A multilingual dataset for enhancing cross-lingual knowledge transfer in low-resource languages. In: Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025) (2025)