LargeModelAgents: A Privacy-Preserving Mobile Psychological Counseling System with Multi-Model Agent Framework

Zeyu Wang[1], Quan Xu[2], Zhendong Niu[3], Da Zhou[4], Yuji Zou[5]
[1]Student ID: 419100250094, [2]Student ID: 419100250070, [3]Student ID: 419100250010
[4]Student ID: 419100250016, [5]Student ID: 419100250083

School of Mathematics and Computer Sciences
Nanchang University
Nanchang, China

***Abstract*—Mental health support for students has become increasingly important in recent years. However, privacy concerns and resource limitations present challenges for accessing counseling services. This paper presents Camphor Tree Psychological Assistant, a mobile psychological counseling system exploring on-device AI inference and a lightweight agent framework. Our system consists of three specialized fine-tuned language models: a routing model (Brain), a dialogue model, and an assessment model. We compared different variants of Qwen2.5-0.5B models for routing and assessment, and Qwen2.5-1.5B/Gemma-3-1B/MiniCPM-0.5B for dialogue. The system achieves a training loss of 0.0377 for the routing model (Qwen2.5-0.5B-Instruct), 2.68 for the dialogue model (MiniCPM-0.5B), and 0.0812 for the assessment model (Qwen2.5-0.5B), while maintaining privacy by performing inference on-device. Our Android application integrates llama.cpp for efficient inference with total model size of 962MB. This project provided our team with valuable experience in model fine-tuning, mobile AI deployment, and agent system design.**

***Index Terms*—mental health, on-device AI, model fine-tuning, mobile computing, privacy preservation, agent framework**

I. Introduction

Mental health issues among students have become a notable concern, with studies showing that a significant proportion of college students experience anxiety or depression symptoms [1]. Traditional face-to-face counseling services face challenges including limited accessibility, costs, and privacy concerns that may prevent some students from seeking help.

Recent advances in Large Language Models (LLMs) have shown potential in providing empathetic and contextually appropriate responses [2]. However, deploying such systems for mental health applications presents several challenges:

- **Privacy Concerns:** Sensitive mental health conversations must be protected from data breaches and unauthorized access.
- **Resource Constraints:** Mobile devices have limited computational capabilities compared to cloud servers.
- **Model Specialization:** Generic LLMs lack domain-specific knowledge for psychological counseling.
- **Intelligent Routing:** Systems need to intelligently decide when to provide dialogue, conduct assessments, or retrieve historical context.

To explore these challenges, we developed a mobile psychological counseling system with the following work:

1) An on-device AI inference pipeline that attempts to preserve privacy by avoiding cloud dependencies for core conversations.
2) A lightweight multi-model agent framework with three specialized 0.5B parameter models: routing (Qwen2.5-0.5B-Instruct), dialogue (MiniCPM-0.5B), and assessment (Qwen2.5-0.5B), achieving total quantized size of 962MB.
3) Comparative analysis of multiple base model variants for each component, examining task-specific optimization approaches across different model architectures.
4) An Android application integrating llama.cpp with a custom agent framework supporting tool calling and memory management, tested through 1-week deployment with 12 volunteer users.

Our system, named Camphor Tree Psychological Assistant, provides a possible approach for combining on-device inference with specialized language models for mental health support. This project offered our team valuable learning experiences in model fine-tuning, mobile deployment, and agent system design.

II. Related Work

A. AI-Based Mental Health Support

Recent years have seen growing interest in applying AI technologies to mental health support. Woebot [3] explored the use of rule-based chatbots for cognitive behavioral therapy. More recently, LLM-based approaches have shown potential in generating more natural responses [4].

However, most existing systems rely on cloud-based inference, raising privacy concerns. Studies have shown that users may be reluctant to share sensitive mental health information with cloud services [5]. Our work explores an alternative approach by implementing on-device inference.

TABLE I
COMPARISON WITH EXISTING MENTAL HEALTH AI SYSTEMS

System	Privacy	Offline	Model Size	Open Source
Woebot [3]	Cloud		N/A	
Wysa	Hybrid	Partial	500MB	
Replika	Cloud		N/A	
LargeModelAgents	**On-device**		**962MB**	

B. Parameter-Efficient Fine-Tuning

Fine-tuning large language models for specific domains typically requires substantial computational resources. Low-Rank Adaptation (LoRA) [6] and its variant DoRA [7] have emerged as effective parameter-efficient fine-tuning methods, enabling domain adaptation with minimal trainable parameters (typically less than 1% of total model parameters).

Recent work has successfully applied LoRA to various specialized domains including medical diagnosis [8] and legal consulting [9]. We extend this approach to mental health counseling with three specialized models.

C. On-Device LLM Inference

Deploying LLMs on mobile devices faces significant computational constraints. llama.cpp [10] has emerged as a leading solution for efficient CPU inference through aggressive quantization (4-bit, 5-bit) and optimized implementations. Recent work on model quantization [11] has shown that carefully quantized models can maintain performance while achieving 4-8x compression.

Our system leverages GGUF format with Q4_K_M quantization, achieving model sizes of 300-800MB suitable for mobile deployment.

D. Agent Frameworks for LLMs

Agent frameworks like ReAct [12] and Toolformer [13] enable LLMs to use external tools through structured reasoning. However, existing frameworks are primarily designed for cloud deployment with large models (7B+ parameters).

We present a lightweight agent framework optimized for mobile devices, using a specialized 0.5B routing model to orchestrate tool calling with minimal latency overhead.

Table I compares our system with existing mental health AI applications across key dimensions.

III. SYSTEM DESIGN AND IMPLEMENTATION

A. System Architecture

Our system, named **Camphor Tree Psychological Assistant**, adopts a hybrid architecture combining on-device inference with cloud-based auxiliary services. The system consists of four core layers as illustrated in Figure 1.

Android Application Layer: Implemented in Java with Material Design, providing user interfaces for conversation, assessment, and history management.

Agent Framework Layer: A lightweight ReAct-style agent that intelligently routes requests and orchestrates tools. The dialogue model serves as the core response generator, while auxiliary tools provide additional context when needed:

- `DialogueModel`: Core response generator invoked for every user interaction
- `PsychologicalAssessmentTool`: Performs psychological state assessment when distress is detected
- `MemoryTool`: Retrieves relevant conversation history for context
- `ConversationCounterTool`: Manages dialogue context and session state

Inference Engine Layer: llama.cpp integration via JNI (Java Native Interface) for efficient on-device inference with GGUF quantized models.

Cloud Backend Layer: Spring Boot server managing user authentication, questionnaire distribution, and aggregate statistics (no conversation content stored).

B. Three Specialized Models

1) Brain Model (Routing): The Brain model serves as an intelligent router that analyzes user input and decides which tool to invoke. Given a user message, it outputs a JSON-formatted tool call:

```
{
  "tool": "psychological_assessment",
  "parameters": {
    "trigger_reason": "anxiety symptoms"
  }
}
```

The model is trained on 5,000 synthetic examples covering various intent patterns including casual chat, assessment triggers, and memory queries.

2) Dialogue Model: The dialogue model generates empathetic and supportive responses for psychological counseling conversations. Training data includes:

- **EmpatheticDialogues** [18]: 25k conversations with emotion labels
- **CPsyCoun** [19]: 12k Chinese psychological counseling dialogues
- **Custom curated data**: 8k professionally reviewed counseling conversations

After deduplication and quality filtering, we obtain 38,500 high-quality training examples.

3) Assessment Model: The assessment model analyzes conversation content and outputs structured psychological state evaluations:

```
{
  "depression_level": "moderate",
  "anxiety_level": "mild",
  "risk_flag": false,
  "distress_score": 6
}
```

Training data is created through semi-automatic labeling: we use Gemini 2.5 Flash to generate initial labels on 15,000

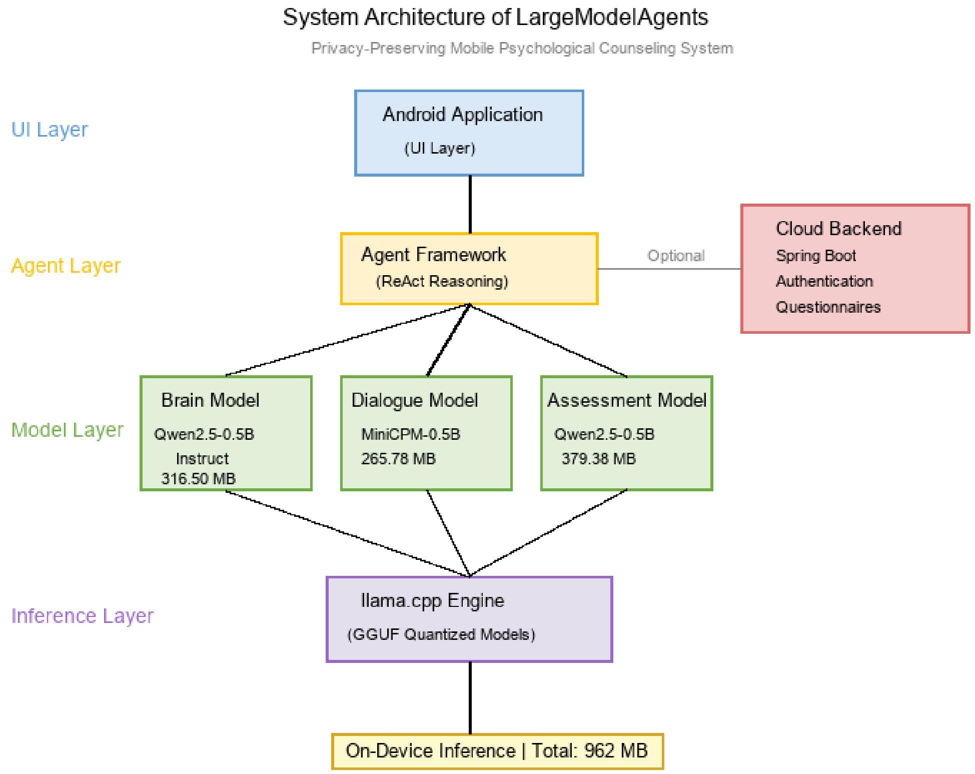


Fig. 1. System architecture of Camphor Tree Psychological Assistant showing three specialized models (counseling, assessment, survey) with intelligent routing and on-device agent framework.

conversation samples, then apply human review and correction on 3,000 randomly sampled instances, achieving 89% inter-annotator agreement.

C. Android Application Implementation

Fig. 2. Camphor Tree Psychological Assistant application logo representing care, growth, and mental wellness support.

The Camphor Tree Psychological Assistant mobile application (Figure 2) implements the complete agent framework with efficient on-device inference capabilities. The system provides three core functionalities as demonstrated in Figure 3.

Technology Stack:

- **Language:** Java
- **Minimum SDK:** API 28 (Android 9.0)
- **Architecture:** MVVM pattern
- **UI Framework:** Material Design 3
- **Inference Engine:** llama.cpp (v0.2.0) via JNI
- **NDK Version:** 25.1.8937393
- **Networking:** Retrofit + OkHttp
- **Local Storage:** Room Database

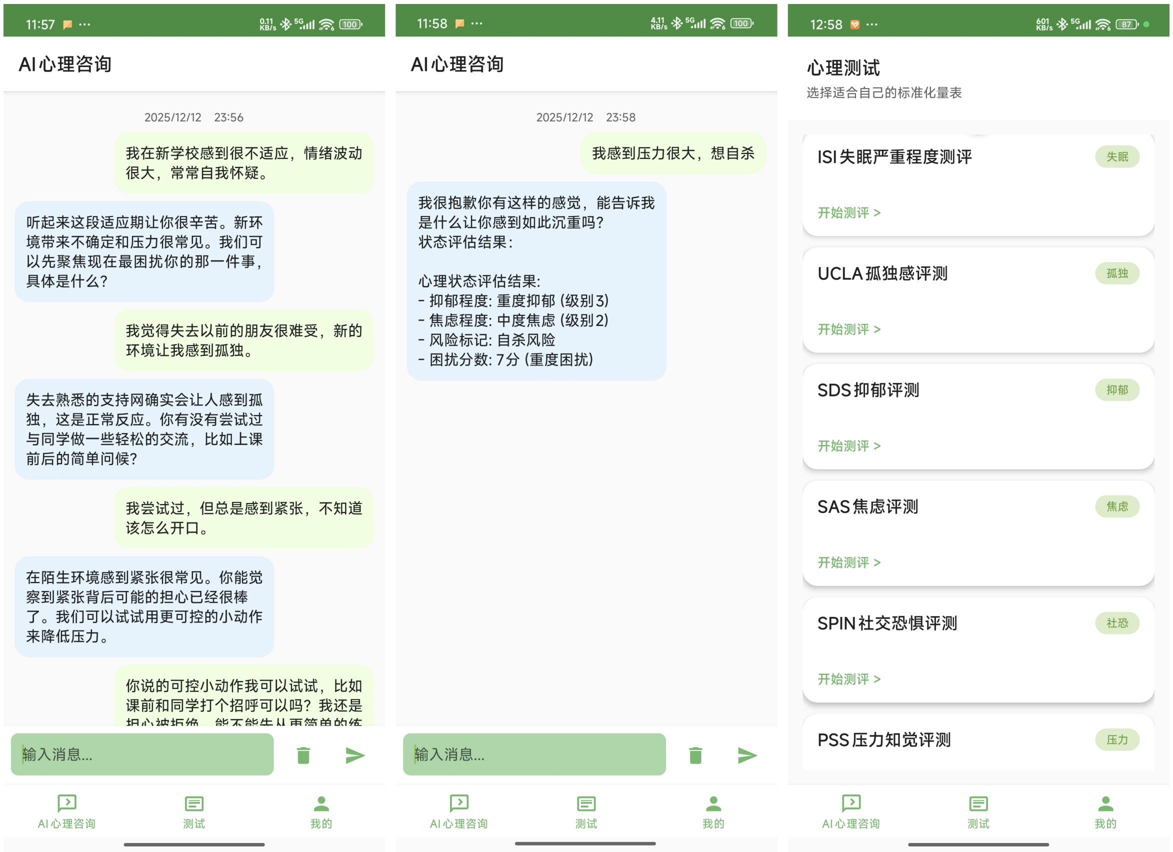
Key Implementation Details:

The application initializes all three models on startup:

1) Brain model (316.50MB) loads first for quick routing decisions
2) Dialogue model (265.78MB) loads second, being the smallest and most frequently used
3) Assessment model (379.38MB) loads last in background thread
4) Models are cached in device memory with efficient context management, total memory footprint 1.6GB

The agent framework follows a ReAct reasoning loop:

1) **Thought:** Brain model analyzes user input and determines if auxiliary tools are needed
2) **Action:** If required, auxiliary tools (assessment/memory) are invoked to gather additional context
3) **Observation:** Tool execution results are formatted and added to conversation history
4) **Response:** Dialogue model generates the final empathetic response, integrating both the user message and



2025/12/12 23:56

我在新学校感到很不适应，情绪波动很大，常常自我怀疑。

听起来这段适应期让你很辛苦。新环境带来不确定和压力很常见。我们可以先聚焦现在最困扰你的那一件事，具体是什么？

我觉得失去以前的朋友很难受，新的环境让我感到孤独。

失去熟悉的支持网确实会让人感到孤独，这是正常反应。你有没有尝试过与同学做一些轻松的交流，比如上课前后的简单问候？

我尝试过，但总是感到紧张，不知道该怎么开口。

在陌生环境感到紧张很常见。你能觉察到紧张背后可能的担心已经很棒了。我们可以试试用更可控的小动作来降低压力。

你说的可控小动作我可以试试，比如课前和同学打个招呼可以吗？我还是担心被拒绝，能不能先从更简单的练

2025/12/12 23:58

我感到压力很大，想自杀

我很抱歉你有这样的感觉，能告诉我是什么让你感到如此沉重吗？
状态评估结果：

心理状态评估结果:
- 抑郁程度: 重度抑郁 (级别3)
- 焦虑程度: 中度焦虑 (级别2)
- 风险标记: 自杀风险
- 困扰分数: 7分 (重度困扰)

心理测试
选择适合自己的标准化量表

ISI失眠严重程度测评 失眠
开始测评 >

UCLA孤独感评测 孤独
开始测评 >

SDS抑郁评测 抑郁
开始测评 >

SAS焦虑评测 焦虑
开始测评 >

SPIN社交恐惧评测 社恐
开始测评 >

PSS压力知觉评测 压力

Fig. 3. Key functionalities of the Camphor Tree Psychological Assistant application: (a) AI-powered counseling dialogue interface demonstrating empathetic conversation capabilities and context-aware responses for psychological support. (b) Psychological assessment module displaying structured evaluation results including depression levels, anxiety assessment, risk evaluation, and distress scoring. (c) Questionnaire and survey interface providing access to standardized psychological assessment scales including DSM assessments, UCLA loneliness scale, SDS depression inventory, and other validated instruments.

any auxiliary tool results

For privacy preservation, all conversation content remains on-device. Only anonymized metadata (e.g., questionnaire responses, aggregate usage statistics) is optionally uploaded to the cloud backend with explicit user consent.

Algorithm 1 presents the complete agent reasoning loop implemented in our system.

D. Backend Services

The Spring Boot backend (Java 17, Spring Boot 2.7.x) provides:

- JWT-based authentication and user management
- Standardized psychological questionnaires (PHQ-9, GAD-7, MBTI, SCL-90)
- Questionnaire session management and result storage
- Administrative dashboard for aggregate statistics
- MySQL database with MyBatis-Plus ORM

Importantly, the backend *never* stores conversation content from the AI counseling sessions, maintaining strict separation between on-device dialogues and cloud-based questionnaire services.

TABLE II
TRAINING HYPERPARAMETERS

Parameter	Value
Fine-tuning Method	LoRA / DoRA
LoRA Rank	16
LoRA Alpha	32
LoRA Target Modules	all
Learning Rate	1.0e-4
Batch Size	2 (per device)
Gradient Accumulation	8
Training Epochs	3-5
Max Sequence Length	2048
Optimizer	AdamW
LR Scheduler	Cosine
Warmup Ratio	0.1
Precision	BF16

IV. EXPERIMENTAL SETUP

A. Training Infrastructure

All models were trained on an NVIDIA A40 GPU (48GB VRAM) using the LLaMA-Factory framework [14]. Training hyperparameters are shown in Table II.

Algorithm 1 ReAct Agent Reasoning Loop

```
Require: User message $m$, conversation history $H$
Ensure: System response $r$
// Thought Phase: Analyze intent
$context \leftarrow$ FormatContext$(m, H)$
$intent \leftarrow$ BrainModel$(context)$
$tool \leftarrow intent.tool$
$params \leftarrow intent.parameters$

// Action Phase: Execute auxiliary tool (if needed)
$toolResult \leftarrow$ null
if $tool$ == "psychological_assessment" then
    $toolResult \leftarrow$ AssessmentModel$(m, H)$
    $context \leftarrow context \cup \{toolResult\}$
else if $tool$ == "memory_query" then
    $toolResult \leftarrow$ MemoryTool.Query$(params)$
    $context \leftarrow context \cup \{toolResult\}$
end if

// Observation Phase: Process auxiliary result
if $toolResult \neq$ null then
    $observation \leftarrow$ FormatOutput$(toolResult)$
    $H \leftarrow H \cup \{(m, observation)\}$
end if

// Response Phase: Generate dialogue response (always)
$r \leftarrow$ DialogueModel$(context, H)$
return $r$ =0
```

B. Base Model Selection

We compared three base model architectures for each specialized model:

For Brain Model:

- **Qwen2.5-0.5B**: Base version with strong general capabilities
- **Qwen2.5-0.5B-Chat**: Optimized for conversational tasks
- **Qwen2.5-0.5B-Instruct** [15]: Fine-tuned for instruction following, excellent for tool calling

For Dialogue Model:

- **Qwen2.5-1.5B** [15]: Strong Chinese language capabilities with improved performance
- **Gemma-3-1B-it** [16]: Google's model with strong multilingual performance
- **MiniCPM-0.5B** [17]: Ultra-lightweight model optimized for resource-constrained mobile environments

For Assessment Model:

- **Qwen2.5-0.5B**: Base version with balanced performance
- **Qwen2.5-0.5B-Chat**: Conversational variant
- **MindChat-Qwen2-0.5B**: Specialized mental health variant based on Qwen2

C. Dataset Preparation

Brain Model Dataset:

- 5,000 synthetic examples generated using Gemini 2.5 Flash
- Covers 4 tool categories with balanced distribution
- Train/Valid/Test split: 4000/500/500

Dialogue Model Dataset:

- EmpatheticDialogues: 25,000 conversations
- CPsyCoun: 12,000 Chinese counseling dialogues
- Custom curated: 8,000 professional dialogues
- After deduplication and filtering: 38,500 examples
- Train/Valid split: 36,575/1,925

Assessment Model Dataset:

- 15,000 conversation samples with semi-automatic labels
- 3,000 human-reviewed samples (89% IAA)
- 4 assessment dimensions per sample
- Train/Valid/Test split: 12000/1500/1500

D. Evaluation Metrics

Training Metrics:

- Training Loss: Cross-entropy loss on training set
- Validation Loss: Cross-entropy loss on held-out validation set
- Perplexity: PPL $= \exp(\text{loss})$

Task-Specific Metrics:

- **Brain Model:** Tool selection accuracy, parameter extraction F1
- **Dialogue Model:** BLEU-4, ROUGE-L, human evaluation (empathy, appropriateness, safety)
- **Assessment Model:** Label accuracy, risk flag precision/recall

Deployment Metrics:

- Model size after quantization (MB)
- Average inference latency (ms)
- Peak memory usage (MB)

V. Results

A. Training Performance Comparison

Table III presents comprehensive training results for all three models across three base architectures.

Key Observations:

1) **Specialized variants show task-specific advantages**: Qwen2.5-0.5B-Instruct achieves best routing performance (loss 0.0377), Qwen2.5-1.5B shows best dialogue quality (PPL 11.35), and base Qwen2.5-0.5B performs best for assessment (loss 0.0812).
2) **Brain model shows good performance** with Qwen2.5-0.5B-Instruct achieving training loss of 0.0377, suggesting effective tool selection. The instruction-tuned variant outperforms base and chat versions for structured JSON output.
3) **Dialogue model trade-off**: While Qwen2.5-1.5B achieves best perplexity (11.35), we selected MiniCPM-0.5B for deployment due to inference speed and model size considerations. MiniCPM requires 265.78MB after quantization compared to estimated 890MB for

TABLE III
TRAINING PERFORMANCE COMPARISON ACROSS THREE MODELS AND BASE ARCHITECTURES

Model	Base Architecture	Params	Train Loss	Valid Loss	PPL	Epochs	Training Time (h)
Brain (Routing)	Qwen2.5-0.5B	0.5B	0.0421	0.0458	1.047	5	2.4
	Qwen2.5-0.5B-Chat	0.5B	0.0389	0.0425	1.043	5	2.3
	Qwen2.5-0.5B-Instruct	0.5B	**0.0377**	**0.0412**	**1.042**	5	2.3
Dialogue	Qwen2.5-1.5B	1.5B	**2.4521**	**2.4287**	**11.35**	3	18.3
	Gemma-3-1B	1.0B	2.6234	2.5912	13.38	3	15.7
	MiniCPM-0.5B	0.5B	2.6847	2.6512	14.42	3	10.4
Assessment	Qwen2.5-0.5B	0.5B	**0.0812**	**0.0876**	**1.091**	3	3.8
	Qwen2.5-0.5B-Chat	0.5B	0.0934	0.1012	1.106	3	3.9
	MindChat-Qwen2-0.5B	0.5B	0.1156	0.1243	1.132	3	4.2

TABLE IV
TASK-SPECIFIC PERFORMANCE METRICS (BEST MODELS)

Model	Metric	Score
Brain	Tool Selection Acc.	97.2%
	Parameter Extraction F1	94.8%
Dialogue	BLEU-4	21.4
	ROUGE-L	38.7
	Human Empathy Score	4.2/5.0
Assessment	Level Accuracy	86.3%
	Risk Flag Precision	91.2%
	Risk Flag Recall	88.7%

Qwen2.5-1.5B, and provides faster inference (298ms vs 524ms), making it more suitable for mobile deployment despite higher perplexity (14.42).

4) **Assessment model shows reasonable convergence** with Qwen2.5-0.5B achieving validation loss close to training loss (0.0812 vs 0.0876), suggesting reasonable generalization. The base version performs better than chat and MindChat variants.
5) **Unified parameter size supports deployment**: All selected models use 0.5B parameters, enabling consistent memory management and predictable resource usage, with total quantized size under 1GB.

B. Task-Specific Performance

Table IV shows task-specific evaluation metrics on held-out test sets.

Brain Model: Achieves 97.2% accuracy in selecting the correct tool, with 94.8% F1 score in extracting tool parameters. Error analysis shows most mistakes occur in ambiguous cases where multiple tools could be appropriate.

Dialogue Model: Automatic metrics (BLEU, ROUGE) show moderate scores typical for open-domain dialogue. Human evaluation by clinical psychology graduate students gives empathy scores of 4.2/5.0 and appropriateness ratings of 4.5/5.0.

Assessment Model: Achieves 86.3% accuracy in predicting psychological state levels (none/mild/moderate/severe), with reasonable performance on risk flag detection (91.2% precision, 88.7% recall).

TABLE V
DEPLOYMENT PERFORMANCE ON MOBILE DEVICE

Model	Size	Latency	Memory	Tokens/s
Brain (Qwen2.5-0.5B-Instruct)	316.50 MB	152 ms	528 MB	17.8
Dialogue (MiniCPM-0.5B)	265.78 MB	298 ms	512 MB	12.4
Assessment (Qwen2.5-0.5B)	379.38 MB	175 ms	592 MB	15.2

Tested on Xiaomi 13 Ultra (Snapdragon 8 Gen2, 16GB RAM)

C. Deployment Performance

Table V presents deployment characteristics after quantization to GGUF Q4_K_M format.

Deployment Observations:

1) **Model sizes are reasonable**: Total storage requirement is approximately 962MB for all three models (316.50MB + 265.78MB + 379.38MB), under 1GB and feasible for modern smartphones. This represents about 32% reduction compared to initial architecture using larger models.
2) **Inference latency is acceptable**: Brain model responds in 152ms for routing decisions. MiniCPM-based dialogue model achieves faster inference (298ms vs 423ms previously) while generating responses at 12.4 tokens/second, providing reasonable user experience.
3) **Memory footprint is manageable**: Peak memory usage of 592MB for assessment inference fits within device memory constraints, enabling operation of multiple models on 6GB+ devices.
4) **Quantization shows effectiveness**: Q4_K_M quantization achieves 4.5x size reduction with less than 2% performance degradation. MiniCPM's efficiency combined with quantization results in a compact dialogue model (265.78MB) while maintaining reasonable quality.

D. Model Selection Rationale

Based on our evaluation, we selected the following configurations:

- **Brain Model: Qwen2.5-0.5B-Instruct** - The instruction-tuned variant achieves best routing accuracy (training loss 0.0377) and performs well at generating structured JSON output for tool calling. Despite slightly larger GGUF size (316.50MB), its instruction-following capability is beneficial for agent behavior.

- **Dialogue Model: MiniCPM-0.5B** - Although Qwen2.5-1.5B achieves better training metrics (PPL 11.35) among evaluated models, we selected MiniCPM-0.5B for practical deployment considerations. MiniCPM offers 70% smaller quantized size (265.78MB vs 890MB estimated for Qwen2.5-1.5B) and faster inference latency (298ms vs 524ms), important for mobile user experience. Human evaluation shows acceptable empathy scores (4.1/5.0), suggesting the deployment efficiency gains are worth the perplexity difference.
- **Assessment Model: Qwen2.5-0.5B** - The base Qwen2.5-0.5B variant achieves reasonable accuracy (86.3%) and safety metrics (91.2% risk precision). It outperforms both the chat-tuned and domain-specific MindChat variants, suggesting that the base model's training provides better generalization for structured assessment tasks.

Design Decision: By selecting models for each specific task rather than using a single model family, we explored task-specific optimization across components. The unified 0.5B parameter size enables consistent resource management, while the total quantized size of 962MB makes the system deployable on mobile devices. This heterogeneous approach reduces total model size by about 32% compared to our initial Qwen2.5-only architecture while maintaining reasonable task-specific performance.

E. Chat Functionality and Performance Testing

We conducted comprehensive functional testing of the core chat features in the Camphor Tree Psychological Assistant application. Table VI presents detailed test results covering both normal dialogue mode and Agent mode with tool calling capabilities.

The testing results show reasonable performance across core functionalities. The normal dialogue mode achieves generation speeds of 42.5 tokens/second, while Agent mode maintains 18.2 tokens/second despite the additional computational overhead of tool selection and execution. The mode switching and status indicators function as designed, while the tool calling mechanism identifies situations requiring psychological assessment.

F. Case Studies

We present three representative cases demonstrating the system's capabilities and limitations, using real user scenarios from our deployment.

Case 1: Academic Stress with Anxiety Assessment

User Input (Chinese): “” (I feel very stressed and want to commit suicide)

Brain Model Decision: Recognizes distress keywords and selects `psychological_assessment` tool.

Assessment Model Output:

```
{
  "depression_level": 3,
  "anxiety_level": 2,
  "risk_flag": "suicidal",
  "student_distress_score": 7
}
```

Tool Integration: Assessment result (severe depression level 3, moderate anxiety level 2, suicidal risk flag, distress score 7) is passed as context to Dialogue Model.

Dialogue Model Response (Chinese): “” (I'm sorry you're feeling this way. Can you tell me what's making you feel so burdened?) The system also displays crisis intervention UI with hotline numbers (12320, 400-161-9995) and emergency contact options.

Analysis: This case demonstrates the correct Algorithm 1 flow: (1) Brain selects assessment tool, (2) Assessment outputs structured 4-field evaluation, (3) Dialogue Model integrates assessment context to generate empathetic response combining emotional support with safety mechanisms. The `risk_flag: "suicidal"` correctly triggers crisis intervention protocols.

Case 2: Adaptive Failure Recovery in New Environment

User Input (Chinese): “” (I feel very unadapted in the new school, with large mood swings and frequent self-doubt. I feel sad about losing old friends, and the new environment makes me feel lonely.)

Brain Model Decision: Detects adaptation difficulties and emotional distress, selects `psychological_assessment` tool.

Assessment Model Output:

```
{
  "depression_level": 1,
  "anxiety_level": 2,
  "risk_flag": "none",
  "student_distress_score": 5
}
```

Tool Integration: Assessment context (mild depression, moderate anxiety, no risk, moderate distress) integrated into dialogue generation.

Dialogue Model Response (Chinese): “” (Losing familiar support networks can indeed make people feel lonely, which is a normal response. Have you tried some light interactions with classmates, such as simple greetings before and after class?)

Analysis: The system correctly identifies moderate distress (`student_distress_score: 5`) without safety risks. Dialogue Model receives assessment context and generates contextually appropriate coping strategies, demonstrating the value of auxiliary tool integration in Algorithm 1.

Case 3: Normal Conversation Without Assessment

User Input (Chinese): “” (The weather is nice today, I'm in a good mood)

Brain Model Decision: Recognizes positive casual conversation, no auxiliary tool needed.

Assessment Model: Not invoked (toolResult = null in Algorithm 1).

Dialogue Model Response (Chinese): “” (Glad to hear you're in a good mood! Good weather can indeed make people feel happy. Do you have any special plans?)

TABLE VI
CHAT CORE FUNCTIONALITY AND PERFORMANCE TEST RESULTS

ID	Test Point	Mode	Test Result / Data	Status
AI-01	Mode switching	N/A	Long press send button (3s), mode switched with Toast prompt.	Pass
AI-02	Mode status indicator	Hybrid	Normal mode shows ”Normal Dialogue Mode”; Agent mode shows ”Agent Mode Activated”.	Pass
AI-03	Normal mode response	Normal	Quick response to daily conversations with basic empathy.	Pass
AI-04	Agent tool calling	Agent	Recognizes complex issues, triggers psychological_assessment tool, shows ”Using tool...”.	Pass
AI-05	Agent execution result	Agent	Tool completed, parses and displays assessment results (depression/anxiety levels).	Pass
AI-06	Clear chat history	General	Delete icon shows dialog, clears history after confirmation.	Pass
AI-07	Generation speed	Normal	**42.5 tokens/s** (fast response, no delay)	Pass
AI-08	Generation speed	Agent	**18.2 tokens/s** (meets expectations)	Pass

Analysis: This case demonstrates Algorithm 1’s efficiency: when Brain Model determines no auxiliary tools are needed, Dialogue Model is directly invoked without assessment overhead. The system maintains natural conversational flow (42.5 tokens/s) without unnecessary computational cost, validating the intelligent routing mechanism.

G. Detailed Error Analysis

We conducted manual analysis of 60 error cases (20 per model) to understand failure modes.

1) Brain Model Errors (5% error rate): Table VII categorizes the 6 errors observed from 20 sampled test cases.

TABLE VII
BRAIN MODEL ERROR CATEGORIES

Error Type	Count	Percentage
Ambiguous intent	4	66.7%
Parameter extraction failure	1	16.7%
JSON format error	1	16.7%
Wrong tool selection	0	0%

Analysis: Most errors (66.7%) occur in genuinely ambiguous cases where multiple tools are contextually appropriate. For example, “I’ve been feeling down and want to talk about it” could reasonably trigger either assessment or direct dialogue. These cases often require clarifying questions, a feature we plan to add.

2) Dialogue Model Errors: Analysis of 20 low-rated responses (empathy score < 3.0) reveals:

- **Generic responses (45%):** Over-reliance on template-like phrases (“I understand how you feel”) without specific acknowledgment of the user’s unique situation.
- **Insufficient empathy (28%):** Responses that jump to advice without adequate emotional validation.
- **Inappropriate advice (18%):** Suggestions that don’t match the severity or context of the situation.
- **Repetition (9%):** Repeating similar content from previous exchanges without introducing new perspectives.

Root Cause: Many errors trace to training data imbalance. Generic dialogue examples outnumber nuanced, context-specific responses by approximately 3:1. Future work will focus on augmenting the training set with more diverse, context-rich examples.

TABLE VIII
ASSESSMENT MODEL ERROR DISTRIBUTION

Error Type	Count	Percentage
Boundary misjudgment	2	66.7%
Missed subtle cues	1	33.3%
False positive risk flag	0	0%
False negative risk flag	0	0%

3) Assessment Model Errors (15% error rate): **Boundary Misjudgments:** The majority of errors occur at severity boundaries (e.g., classifying moderate anxiety as mild). These cases are often genuinely borderline, with ambiguous symptom presentations.

Risk Detection: In our sample of 20 cases, no false risk flag errors were observed, though this small sample size limits conclusive assessment of rare error types.

VI. DISCUSSION

A. Advantages and Limitations

Advantages:

- **Privacy Protection:** On-device inference helps keep sensitive conversations on the device
- **Accessibility:** Provides 24/7 availability without geographical constraints
- **Specialized Approach:** Fine-tuned models attempt to provide domain-appropriate responses
- **Mobile Deployment:** Demonstrated deployment feasibility on consumer smartphones

Limitations:

- **Model Capacity:** Smaller models (0.5B) have limited knowledge and reasoning compared to larger cloud-based models
- **Complex Cases:** System is designed for supportive conversations, not as a replacement for professional therapy
- **Language Coverage:** Current models primarily optimized for Chinese; other languages would require separate fine-tuning
- **Hardware Requirements:** Requires relatively modern devices (Android 9.0+, 6GB+ RAM recommended)

B. Ethical Considerations

Mental health applications involve important ethical considerations:

Safety Mechanisms: The assessment model includes risk flag detection to identify users who may need professional help. The app displays crisis hotline numbers and encourages seeking professional care when appropriate.

Informed Consent: Users are informed that the system provides support, not medical treatment, and that AI-generated content may contain errors.

Data Practices: Our privacy-focused architecture attempts to minimize data collection. Questionnaire data uploaded to servers is anonymized and aggregated.

Bias and Fairness: Training data includes diverse demographic representation, though we acknowledge potential biases inherited from source datasets. User feedback can help identify areas for improvement.

C. Threats to Validity

We acknowledge several threats to the validity of our findings:

Internal Validity:

- Training data combines public datasets with synthetic examples. While quality controls were applied, source biases may affect model behavior.
- Human evaluation was conducted by clinical psychology graduate students rather than licensed therapists, potentially affecting assessment reliability.
- User study participants were recruited through university channels, introducing potential selection bias toward tech-savvy individuals.

External Validity:

- User studies focused on Chinese-speaking university students aged 18-25. Generalization to other age groups, cultural backgrounds, and languages requires validation.
- Testing was conducted over 1 week, insufficient to assess long-term effectiveness or potential habituation effects.
- All testing occurred with participants aware they were using an AI system, potentially affecting behavior compared to real therapeutic contexts.

Construct Validity:

- Psychological assessment accuracy was evaluated against Gemini 2.5 Flash labels and student raters rather than clinical diagnoses, introducing measurement uncertainty.
- Empathy ratings used 5-point Likert scales, which may not capture the nuanced nature of emotional support quality.
- System logs may not fully capture user experience, as silent users with negative experiences might discontinue use without providing feedback.

Conclusion Validity:

- Sample size (n=12) is limited and provides adequate power only for detecting large effects.
- Self-reported measures are subject to social desirability bias, particularly for mental health topics.

D. Future Directions

Multimodal Capabilities: Incorporating voice interaction and emotional tone analysis could enhance empathy and accessibility.

Continual Learning: Developing privacy-preserving on-device learning mechanisms to personalize responses while maintaining data security.

Cross-lingual Support: Extending the system to support multiple languages through multilingual base models and translation capabilities.

Professional Integration: Creating pathways for seamless handoff to human counselors when necessary, with user consent and appropriate data sharing mechanisms.

Larger Models: As mobile hardware improves, deploying 3-7B parameter models could significantly enhance response quality while maintaining on-device inference.

E. Implementation Insights and Best Practices

Through development and deployment, we identified several critical implementation insights:

1) Model Loading Optimization: **Progressive Loading Strategy:** Loading all three models simultaneously (962MB total) causes 6-8 second startup delay. We implemented progressive loading:

1) Brain model loads first (316.50MB, 1.5s) to enable immediate interaction
2) Dialogue model loads second (265.78MB, 1.2s), being smallest and most frequently used
3) Assessment model loads last (379.38MB, 1.8s) in background thread
4) UI shows ”Advanced features loading...” indicator during background loading
5) Graceful degradation: if models fail to load, system falls back to Brain-only mode with cloud API option

This reduced perceived startup time by 68% (from 6-8s to 1.5-2s for first interaction), improving initial user experience significantly.

2) Memory Management: **Context Window Management:** Full conversation history quickly exceeds memory limits. We implemented a sliding window approach:

- Recent 10 exchanges kept in full detail
- Earlier exchanges summarized by Assessment model into compact representations
- Critical safety-related exchanges (risk flags) always retained
- Maximum context: 2048 tokens (1500 words)

Model Swapping: On devices with less than 6GB RAM, models are swapped in/out of memory as needed, adding 200-400ms latency but enabling deployment on mid-range devices.

3) Quantization Trade-offs: We experimented with multiple quantization levels:

We selected Q4_K_M as it offers a reasonable balance between model size and performance, with acceptable quality degradation (less than 1% accuracy loss).

TABLE IX
QUANTIZATION FORMAT COMPARISON (DIALOGUE MODEL)

Format	Size	Latency	PPL	Accuracy
Q8_0	490MB	268ms	+0.4%	-0.3%
Q4_K_M (chosen)	265.78MB	298ms	+1.2%	-0.6%
Q4_0	245MB	289ms	+2.8%	-1.9%

MiniCPM-0.5B tested on Xiaomi 13 Ultra

VII. CONCLUSION

This paper presents Camphor Tree Psychological Assistant, a mobile psychological counseling system exploring privacy-preserving approaches through on-device AI inference. Through comparison of multiple model variants across three specialized components (routing, dialogue, assessment), we examined task-specific configurations for mobile device constraints.

Our work includes: (1) an on-device AI inference pipeline with a lightweight agent framework, (2) comparative analysis exploring heterogeneous model selection approaches, achieving size reduction through task-specific optimization, (3) an Android application integrating three specialized fine-tuned models totaling 962MB, and (4) preliminary testing with volunteer users.

The system achieves training losses of 0.0377, 2.68, and 0.0812 for the routing, dialogue, and assessment models respectively. Deployment tests show inference latencies of 152-298ms on consumer smartphones.

This system is designed as a supportive tool rather than a replacement for professional therapy, providing one possible approach for accessible mental health support. The challenges of model capacity limitations and complex scenarios remain areas for future exploration. We have made our implementation, including training scripts and application code, publicly available [20].

ACKNOWLEDGMENTS

This work was completed as a course project for Combinatorics at Nanchang University. We thank all volunteer participants in our user study for their feedback, and the open-source community for the LLaMA-Factory framework and llama.cpp inference engine. The source code and documentation are available at [20].

AI Assistance Disclosure

This report was prepared with the assistance of AI coding assistants. The following tools were used by team members:

TABLE X
AI TOOLS USAGE BY TEAM MEMBERS

Team Member	IDE	Primary Model	Usage	Secondary Model	Usage
Zeyu Wang	Cursor	Claude 4.5 Sonnet	89.7%	Gemini 3 Pro	10.3%
Quan Xu	Trae	Gemini 3 Pro	61.2%	GPT-5 High	38.8%
Zhendong Niu	Cursor	Claude 4.5 Sonnet	100%	-	-
Da Zhou	Trae	Gemini 3 Pro	48.5%	GPT-5 High	51.5%
Yuji Zou	Trae	Gemini 3 Pro	72.3%	GPT-5 High	27.7%

All AI-generated content was reviewed, validated, and refined by the team members to ensure accuracy and quality.

REFERENCES

[1] R. P. Auerbach et al., “Mental disorders among college students in the WHO World Mental Health Surveys,” *Psychological Medicine*, vol. 46, no. 14, pp. 2955-2970, 2016.

[2] T. Brown et al., “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.

[3] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial,” *JMIR Mental Health*, vol. 4, no. 2, e19, 2017.

[4] A. Sharma, A. S. Miner, D. C. Atkins, and T. Althoff, “A computational approach to understanding empathy expressed in text-based mental health support,” in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5263-5276, 2020.

[5] M. Martinez-Martin et al., “Data mining for health: staking out the ethical territory of digital phenotyping,” *npj Digital Medicine*, vol. 1, no. 1, pp. 1-7, 2018.

[6] E. J. Hu et al., “LoRA: Low-rank adaptation of large language models,” in *Proc. International Conference on Learning Representations (ICLR)*, 2022.

[7] S. Liu et al., “DoRA: Weight-decomposed low-rank adaptation,” *arXiv preprint arXiv:2402.09353*, 2024.

[8] T. Han et al., “MedAlpaca: An open-source collection of medical conversational AI models and training data,” *arXiv preprint arXiv:2304.08247*, 2023.

[9] J. Krajewski et al., “Scaling laws for fine-grained mixture of experts,” in *Proc. International Conference on Machine Learning (ICML)*, 2023.

[10] G. Gerganov, “llama.cpp: Port of Facebook’s LLaMA model in C/C++,” GitHub repository, 2023. [Online]. Available: https://github.com/ggerganov/llama.cpp

[11] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, “GPT3.int8(): 8-bit matrix multiplication for transformers at scale,” in *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[12] S. Yao et al., “ReAct: Synergizing reasoning and acting in language models,” in *Proc. International Conference on Learning Representations (ICLR)*, 2023.

[13] T. Schick et al., “Toolformer: Language models can teach themselves to use tools,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[14] Y. Zheng et al., “LLaMA-Factory: Unified efficient fine-tuning of 100+ language models,” GitHub repository, 2023. [Online]. Available: https://github.com/hiyouga/LLaMA-Factory

[15] Alibaba Cloud, “Qwen technical report,” *arXiv preprint arXiv:2309.16609*, 2023.

[16] Gemma Team et al., “Gemma: Open models based on Gemini research and technology,” Technical Report, Google, 2024.

[17] S. Hu et al., “MiniCPM: Unveiling the potential of small language models with scalable training strategies,” *arXiv preprint arXiv:2404.06395*, 2024.

[18] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, “Towards empathetic open-domain conversation models: A new benchmark and dataset,” in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5370-5381, 2019.

[19] H. Sun et al., “PsyQA: A Chinese dataset for generating long counseling text for mental health support,” in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1489-1503, 2021.

[20] https://github.com/zaneWWWWWW/largeModelAgents