

An Offline Mobile Conversational Agent for Mental Health Support: Learning from Emotional Dialogues and Psychological Texts with Student-Centered Evaluation

Vimaleswar A[△], Prabhu Nandan Sahu, Nilesch Kumar Sahu, Haroon R. Lone

Department of Electrical Engineering and Computer Science, Indian Institute of Science Education and Research Bhopal

Abstract

Mental health plays a crucial role in the overall well-being of an individual. In recent years, digital platforms have increasingly been used to expand mental health and emotional support. However, there are persistent challenges related to limited user accessibility, internet connectivity, and data privacy, which highlight the need for an offline, smartphone-based solutions. To address these challenges, we propose **EmoSApp (Emotional Support App)**¹: an entirely offline, smartphone-based conversational app designed to provide mental health and emotional support. EmoSApp leverages a language model, specifically the LLaMA-3.2-1B-Instruct, which is fine-tuned and quantized on a custom-curated “Knowledge Dataset” comprising 14,582 mental health QA pairs along with multi-turn conversational data, enabling robust domain expertise and fully on-device inference on resource-constrained smartphones.

Through qualitative evaluation with students and mental health professionals, we demonstrate that EmoSApp has the ability to respond coherently and empathetically, provide relevant suggestions to user’s mental health problems, and maintain interactive dialogue. Additionally, quantitative evaluations on nine commonsense and reasoning benchmarks, along with two mental health specific datasets, demonstrate EmoSApp’s effectiveness in low-resource settings. By prioritizing on-device deployment and specialized domain-specific adaptation, EmoSApp serves as a blueprint for future innovations in portable, secure, and highly tailored AI-driven mental health support.

1 Introduction

Recent advancements in digital platforms and Natural Language Processing (NLP), particularly with

[△]Corresponding author.

¹Our data, app, models are available at <https://huggingface.co/SIR-Lab/EmoSApp>

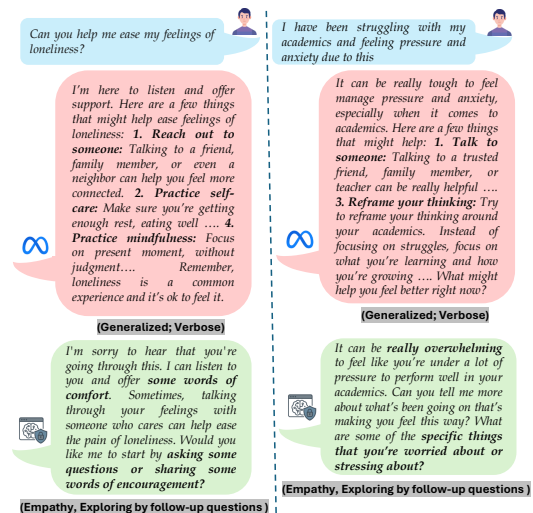


Figure 1: Example responses generated by LLaMA-3.2-1B-Instruct and our proposed EmoSApp. LLaMA model (pink response box) often give generalized and verbose responses, resulting in an AI-generated format. In contrast, EmoSApp (green response box) demonstrates stronger empathy and deeper conversational exploration, effectively providing emotional and mental health support to users.

the emergence of Large Language Models (LLMs) (OpenAI et al., 2024; Touvron et al., 2023b; Yang et al., 2024a), have enabled the rise of conversational Artificial Intelligence (AI) for mental health support (Liu et al., 2022), addressing several traditional barriers to care (Khosravi and Azar, 2024). These technologies offer potential advantages, including ubiquitous availability, reduced costs, and scalability. Despite these advancements, persistent challenges related to user accessibility, the need for constant internet connectivity, data privacy, and high computational requirements limit the effectiveness of existing solutions (Coelho et al., 2025; Alhammad et al., 2024).

Most existing conversational AI or chatbots for emotional and mental health support depends mainly on a computer-based design that requires continuous internet connectivity and transmit highly sensitive user data to external servers,

thereby amplifying privacy risks (Zhang et al., 2023). As depicted in Figure 1, such chatbots often tend to be generalized, verbose and formulaic, often lacking the nuanced, context-specific guidance that real-world users require (Kang et al., 2024; Loh and Raamkumar, 2023). Furthermore, the computational demands of LLMs have precluded their deployment directly on resource-constrained devices, such as smartphones.

To address these gaps, we present **EmoSApp**², a fully offline, smartphone-based conversational AI app that designed to provide mental health and emotional support, particularly to students. Our approach leverages recent advances in model fine-tuning and quantization techniques, to enable on-device inference with LLMs on resource-constrained smartphone, specifically using TorchTune³ and ExecuTorch⁴ library. To our knowledge, we are the first to use ExecuTorch and TorchTune in the mental health domain. By fine-tuning the LLaMA-3.2-1B-Instruct model⁵ on our custom-curated dataset from various psychology and mental health textbooks, comprising 14,582 mental health question-answer pairs, along with two multi-turn conversational corpora, we equip EmoSApp with domain-specific knowledge while maintaining practical computational efficiency for smartphone deployment.

The on-device nature of **EmoSApp** directly addresses several critical limitations of existing approaches: (i) By running entirely offline, it eliminates internet-connectivity barriers, making it easily accessible to a broader range of smartphone users. (ii) By keeping all conversation data logs locally on the user’s device, it enhances privacy protections for sensitive mental health conversations. (iii) By utilizing quantized models, it delivers responsive performance on modern smartphones with as little as 6GB of system RAM. (iv) Despite quantization, our qualitative evaluations show that EmoSApp maintains coherent, empathetic, and contextually relevant interactions that address users’ mental health concerns. By combining specialized domain adaptation through fine-tuning with efficient smartphone deployment, **EmoSApp** represents a significant step towards

making AI-driven emotional support for the student population more accessible, private, and contextually appropriate.

Our contributions are summarized in four folds:

- We present **EmoSApp**, a fully functional offline emotional and mental health support app tailored for students, by fine-tuning and quantizing LLaMA-3.2-1B-Instruct optimized for resource-constrained smartphones.
- We propose a domain specialization approach via multi-source dataset fusion, integrating ESConv, ServeForEmo, and a custom-curated knowledge dataset derived from psychological and mental health textbooks.
- We explore and compare multiple fine-tuning and quantization strategies (Full, LoRA + PTQ, QAT-LoRA) to balance model performance and on-device efficiency.
- Extensive quantitative and qualitative human evaluations demonstrate that EmoSApp excels in providing coherent, empathetic, and contextually appropriate support to students.

2 Related Work

The advancements and adaptability of Large Language Models (LLMs) have encouraged researchers to explore their use in mental health applications (Liu et al., 2021). LLMs have been applied to diverse tasks, such as predicting mental health conditions (Xu et al., 2024), summarizing mental health conversations (Sahu et al., 2025) and offering support through conversational agents (Ye et al., 2025). With their extensive training and large parameter counts, models like ChatGPT, LLaMA, Gemini, and Claude have shown the potential to act as virtual counselors or mental health advisors (Song et al., 2024; Malgaroli et al., 2025). As a result, some users have started exploring these models for mental health support (Song et al., 2024). However, privacy remains a significant concern when using online LLM-based systems for sensitive and general tasks (Chandra et al., 2024; Mandal et al., 2025). To address this, researchers have begun developing lightweight LLMs that can run locally on personal devices (Liu et al., 2024; Thawakar et al., 2024). However, running high-parameter LLMs locally requires powerful and expensive hardware. While these smaller models

²<https://huggingface.co/SIR-Lab/EmoSApp>

³<https://github.com/pytorch/torchTune>

⁴<https://github.com/pytorch/executorch>

⁵<https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct>

make on-device inference more feasible, the reduction in model size often results in diminished reasoning capacity and knowledge retention compared to larger models (Deshpande et al., 2023). Moreover, these models have primarily been optimized for general-purpose language understanding rather than for specialized domains such as mental health support.

Language models and Mental Health: Language models often lack empathy and do not follow clear support strategies when responding to users with emotional concerns. To address this, (Liu et al., 2021) created a handcrafted dataset called Emotional Support Conversation (ESConv), which includes common support strategies people use when helping others. They fine-tuned models like DIALGPT and BlenderBot using this dataset. The fine-tuned models showed improved performance compared to the base versions by providing structured and strategy-driven emotional support. However, constructing such datasets at scale remains challenging due to the heavy reliance on human annotation.

Recent work has turned LLMs to scale up data generation. For instance, (Zheng et al., 2023) fine-tuned the 6B GPT-J model on ESConv to create a larger dataset called AugESC. However, these augmented datasets do not explicitly label or control the support strategies used, which limits interpretability. To address this limitation, (Ye et al., 2025) proposed a multi-agent framework using three LLMs: a help seeker, a strategy planner, and a supporter. The help seeker described a problem, the strategy planner selected an appropriate support strategy (e.g., affirmation, reflection of feelings), and the supporter generated responses accordingly. This approach produced over 3.7K multi-turn dialogues and enabled the fine-tuning of LLaMA-3.1-8B, resulting in emotionally rich and contextually supportive responses.

3 Mental Health Support Conversational Agent

The primary goal of our work is to develop a fully offline mental health support chatbot capable of operating efficiently on low-end smartphones with limited processing power, computational resources, and memory capacity. To ensure the model remains lightweight, we chose the open-source 1.24 billion parameter, LLaMA-3.2-1B-Instruct model as our base. Although language models with ≤ 1 B pa-

rameters are generally effective for casual conversations, they often lack the empathetic responses necessary for meaningful emotional and mental health support.

LLaMA-3.2-1B-Instruct is a decoder-only transformer model from Meta’s LLaMA series (Touvron et al., 2023a; Grattafiori et al., 2024), fine-tuned for instruction-following tasks. Despite its relatively small size, the model exhibited high response latency on smartphones, even those equipped with 8 GB of RAM, making it unsuitable for practical deployment (see Table 1). To mitigate this, we applied quantization, a technique that reduces model size and memory footprint by converting high-precision weights and activations into lower-precision formats.

In this work, we experimented with three different fine-tuning methods to adapt the LLaMA-3.2-1B-Instruct model for the mental health support domain, emphasizing empathy and emotional understanding: Full fine-tuning, LoRA and Post-Training Quantization (LoRA+PTQ), and Quantization-Aware Training (QAT) with LoRA. These methods differ in the number of parameters updated, the quantization approach used, and their efficiency in terms of memory and inference speed. We used datasets related to empathy and mental health for fine-tuning, described in Section 4.

3.1 Full fine-tuning

Full fine-tuning updates all the parameters of the LLaMA-3.2-1B-Instruct model using gradient descent. This approach is straightforward and highly effective, but demands substantial GPU VRAM, high memory consumption, and longer training time since all transformer weights are updated simultaneously. During evaluation, we observed that the *Full* fine-tuned model performed better than the pretrained model in mental health related conversations. *However, the model required a high-end smartphones for deployment (see Table A.5 in the Appendix), and its response speed was slower when executed on smartphones (see Table 1).*

3.2 LoRA fine-tuning + PTQ

To address the high computational and memory demands of Full fine-tuning, Shen et al. (2021) proposed **LoRA** (Low-Rank Adaptation), a parameter-efficient fine-tuning (PEFT) that reduces the number of trainable parameters, as illustrated in Figure 2. This technique freezes the original model weights and introduces two small trainable low-

Model	Response speed (tokens/sec)	Time-to-first-token (sec)	Model size (GB)
Full	3.59	46.2	2.30
QAT-LoRA	13.50 (3.76x)	5.69 (-87.71%)	1.03 (-55.26%)

Table 1: Performance comparison of our fine-tuned models, averaged over the first five conversational turns, measured on an smartphone with 8 GB of system RAM and Android 15. Response speed (tokens/second) indicates number of tokens generated per second during inference (higher is better). Time-to-first-token (TTFT) measures the latency before the first token is produced (lower is better). Model size refers to the binary PTE file format.

rank matrices **A** and **B** into selected layers. These matrices significantly reduce the number of trainable parameters while allowing the model to adapt to the domain-specific task, leading to lower GPU memory usage and faster training.

For an original weight matrix $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, LoRA parameterizes the adapted weight as:

$$\mathbf{W}_{\text{new}} = \mathbf{W} + \Delta\mathbf{W} = \mathbf{W} + \mathbf{B}\mathbf{A}$$

where $\mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}$, $\mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}$, and $r \ll \min(d_{\text{in}}, d_{\text{out}})$ represents the rank of the decomposition. During fine-tuning, the pretrained weights \mathbf{W} remain frozen, and only **A** and **B** are updated.

LoRA hyperparameters include the rank (r), which sets the dimensionality of matrices **A** and **B**, and the alpha (α) is the scaling factor which modulates the magnitude of updates in $\Delta\mathbf{W}$. The α is often set using the heuristic: $\alpha \approx 2 \times r$.

To support on-device deployment, we further applied PTQ to convert full-precision (BF16) weights to INT4 and dynamic activations to INT8, thereby defining the quantization ranges. *However, in our experiments, we found that PTQ led to a drop in model performance (see Tables 3, 4) as the reduced precision caused information loss during both the forward and backward passes.*

3.3 QAT-LoRA fine-tuning

To address the performance degradation observed with PTQ, we employed QAT-LoRA, which integrates LoRA’s low-rank adapters with QAT to achieve highly memory efficient and strong performance models. LoRA adds two small trainable matrices into selected layers while keeping the original model weights frozen, thereby reducing the number of parameters that need to be trained. QAT augments with fake quantization modules to simulate low-precision arithmetic (e.g., INT8 or

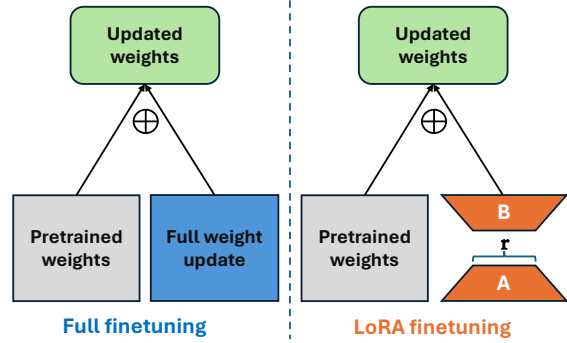


Figure 2: An illustration comparing a single weight update in standard Full fine-tuning (left) versus LoRA fine-tuning (right). In LoRA, the low-rank matrices **A** and **B** approximate the full-rank weight update (highlighted in blue), reducing the number of trainable parameters.

INT4) during fine-tuning while retaining the original BF16 precision for gradient updates. After fine-tuning, these fake quantization modules are converted into real quantized layers, resulting in a fully quantized model suitable for deployment on resource-constrained devices. The resulting QAT-LoRA model maintains performance comparable to the Full model and runs on smartphones with as low as 6 GB of RAM (see Table 1 and Appendix section A.6). All fine-tuning and evaluations were conducted on a single NVIDIA GeForce RTX 4070 Ti GPU with 12 GB of VRAM.

Implementation Details

1. *Full Fine-Tuning:* We use the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 3×10^{-5} , a linear warm-up over the first 1% of steps, a batch size of 4, and train for 5 epochs. The model checkpoint with the best validation performance is selected for evaluation.
2. *LoRA Fine-Tuning:* We apply LoRA to the W_q , W_k , W_v , and W_o parameters with rank $r = 16$ and $\alpha = 32$, using the same optimizer, learning rate, and schedule as above. We then apply PTQ to convert weights into a 4-bit group-wise format (group size 32) and activations into 8-bit dynamic precision.
3. *QAT-LoRA Fine-Tuning:* We employ the same LoRA configuration as above alongside the Int8DynActInt4WeightQAT quantizer to simulate fake quantization during training, delaying fake quantization to 1,000 steps. We follow Meta’s quantization setup⁶ as follows:

⁶<https://ai.meta.com/blog/>

- (a) 4-bit group-wise weights (group size 32) and 8-bit dynamic activations in all transformer blocks.
- (b) 8-bit per-channel weights and 8-bit dynamic activations in the classification layer.
- (c) Similar to the classification layer, an 8-bit per-channel quantization is used for the embedding layer.

4 Dataset

To equip our model with both comprehensive domain knowledge and rich emotional context, we utilized two complementary corpora: (1) **Knowledge Dataset**⁷ curated from diverse psychology and mental health textbooks, consisting of carefully crafted conversational question-answer (QA) pairs (2) **Conversational datasets** drawn from publicly available emotional support dialogue collections such as ESConv (Liu et al., 2021) and ServeForEmo (Ye et al., 2025). Integrating conversational datasets provides our model a deeper understanding of various nuanced emotions and exposure to a wider variety of mental health challenges, including academic pressure, loneliness, stress.

4.1 Knowledge Dataset

We developed a novel knowledge-based dataset consisting of **14,582** question-answer (QA) pairs extracted from eight different psychology and mental health textbooks, spanning basic, social, cognitive, and sport psychology, neuroscience, and psychological disorders. The textbooks were selected based on their breadth and depth of mental health topic coverage, recency of content, and positive reader ratings. Details of selected books are provided in the Appendix A.1.

Ingesting complete textbooks directly is not a good fit for building chatbots, as their formal prose conflicts with the conversational style, reduces retrieval accuracy, and increases the risk of hallucination (Ji et al., 2023). Following best practice in document-to-dialogue conversion (Hu et al., 2024), we therefore reformatted each textbook into concise conversational QA snippets, allowing the knowledge to be injected as small, coherent chunks aligned with the model’s training style.

Our extraction pipeline comprises of two steps: (i) **QA extraction**: Each textbook was divided into 10-page chunks to ensure that each chunk’s content

remained meaningful yet small enough for ChatGPT⁸ to handle. ChatGPT (GPT-4o model), guided by structured prompts (see Appendix A.2), generated QA pairs capturing key definitions, concepts, theories, evidence-based suggestions, and examples from each chunk. Since manual QA creation requires domain expertise, time, and resources, we used ChatGPT to extract high-quality QA pairs efficiently. (ii) **Validation**: Manual checks were performed to identify whether the generated QA pairs contain factual errors, misinterpretations, or misaligned content with the textbook material, thereby ensuring factual fidelity and topical relevance.

To simulate natural conversation flow, we organized the validated QA pairs into dialogues of ten QA exchanges each, resulting in **1,459** dialogues and **29,164** utterances (see Table 2). Within these dialogues, the *seeker* poses questions, and the *supporter* provides the answers. A representative sample dialogue from the Knowledge Dataset is provided in Table A.1 in the Appendix.

4.2 Conversational Datasets

ESConv: Emotion Support Conversation (ESConv) dataset (Liu et al., 2021) consists of 1,300 real-world, multi-turn dialogues collected from crowdworkers who engage as help seekers and supporters. The dataset employs eight support strategies mapped to the three stages of Hill’s Helping Skills Theory (Hill, 2009): Exploration, Comforting, and Action. Unlike traditional expert-led counseling, ESConv emphasizes everyday social interactions, reflecting peer-based emotional support dynamics among friends, family members, or peers in non-clinical contexts. Its diverse scenarios, ranging from stress, academic challenges, and job issues to social anxiety, relationship concerns, ensure broad coverage of emotional experiences.

ServeForEmo: ServeForEmo (Ye et al., 2025) is a large-scale emotional support corpus comprising 3.7K multi-turn dialogues and 59.3K utterances, generated via a strategy-enhanced role-playing framework involving three LLM-driven roles: Seeker, Counselor, and Supporter. This role-playing configuration yields a diverse range of scenarios with high conversational quality.

Table 2 summarizes the statistics of all three corpora. Appendix A.4 provides the distribution of support strategies in ESConv and ServeForEmo.

meta-llama-quantized-lightweight-models/
⁷<https://huggingface.co/SIR-Lab/EmoSApp>

⁸<https://chat.openai.com>

Category	ESConv	ServeForEmo	Knowledge Data
Total			
# Dialogues	1,300	3,743	1,459
# Utterances	38,365	59,536	29,164
Avg. dialogue length	29.51	15.91	19.99
Avg. utterance length	16.40	18.45	16.47
Seeker			
# Utterances	19,989	29,768	14,582
Avg. # utter. per dialog	15.38	7.95	9.99
Avg. utterance length	14.82	15.39	8.52
Supporter			
# Utterances	18,376	29,768	14,582
Avg. # utter. per dialog	14.14	7.95	9.99
Avg. utterance length	18.11	21.51	24.41

Table 2: Comparison of ESConv, ServeForEmo, and Knowledge datasets.

Dataset Integration

We integrated the conversational and knowledge datasets to fine-tune the model, enabling it to combine empathetic dialogue with reliable psychological information, a balance shown to improve user trust and engagement in mental health assistants (Hu et al., 2024). Each dataset was split 70:30 for training and validation to ensure a consistent comparison across sources.

5 Evaluation

We evaluated our fine-tuned models using both quantitative and qualitative methods. Qualitative evaluation complements quantitative evaluation by focusing on subjective perceptions and user experiences, revealing insights into how well the model provides emotional support, which quantitative metrics may overlook.

5.1 Quantitative Evaluation

To provide a comprehensive evaluation, we structure our quantitative analysis into two complementary benchmarks: (i) *General Benchmarks*: Assesses language understanding, fluency, commonsense knowledge, and reasoning abilities. (ii) *Domain-Specific Mental Health QA Benchmarks*: Evaluates performance on various mental health focused question-answering tasks.

5.1.1 General Benchmarks

We used the LM Evaluation Harness framework (Gao et al., 2024) across nine diverse datasets to evaluate a range of model capabilities, including language comprehension, domain-specific and commonsense knowledge, logical reasoning, and general problem-solving. The evaluated benchmarks include HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020),

TruthfulQA (Lin et al., 2022), ARC-Challenge (Clark et al., 2018), WinoGrande (Sakaguchi et al., 2021), OpenBookQA (Mihaylov et al., 2018), Social-IQ (SIQA) (Sap et al., 2019), CommonsenseQA (Talmor et al., 2019), and MMLU (Hendrycks et al., 2021). Detailed descriptions of these datasets are provided in Appendix A.3. Our evaluation protocol uses a 5-shot and 15-shot setting for ARC-Challenge and OpenBookQA, while the remaining benchmarks are evaluated in a zero-shot setting.

Result: Table 3 shows a comparison of pretrained and fine-tuned variants across nine benchmark datasets. The *Full* model achieves the highest accuracy on nearly every dataset. *LoRA+PTQ* led to a drop in accuracies. However, *QAT-LoRA* model delivers intermediate results, effectively bridging the gap between the *Full* and *LoRA+PTQ* models. While the *Full* model gives the best results, it demands substantially more computational resources. In contrast, *QAT-LoRA* offers a better balance between accuracy and computational resources.

5.1.2 Domain-Specific Mental Health QA Benchmarks

We evaluated our models on two mental health benchmarks.

1. MHQA (Racha et al., 2025): MHQA-Gold dataset comprises 2,475 multiple-choice questions spanning four core mental-health domains: Anxiety, Depression, Trauma, and Obsessive-Compulsive Disorder. We benchmark against six sub-1B parameter baselines (BERT-base, MentalBERT, RoBERTa, Mental-RoBERTa, BioBERT, PubMedBERT) to align with the parameter budget, capacity, and on-device constraints of our fine-tuned models.

Result: Table 4 shows that the *Full* fine-tuned model achieves an average accuracy of 44.40%, outperforming all baselines by a substantial margin. While the absolute accuracy appears modest, this aligns with prior work (Racha et al., 2025), where accuracies below 50% have been consistently reported with sub-1B models. *QAT-LoRA* variant, optimized for on-device inference, attains 39.34%, surpassing all baseline models, while *LoRA+PTQ* drops drastically.

2. IMHI (Yang et al., 2024b): It is the first multi-task, multi-source interpretable mental health instruction dataset constructed from social media sources including Reddit, Twitter, and SMS. We evaluate our fine-tuned models in zero-shot set-

Model	HellaSwag	PIQA	TruthfulQA	ARC_C	Winogrande	OBQA	SIQA	CSQA	MMLU	Average
<i>Pretrained</i>	41.20	74.10	34.73	35.41	59.67	24.00	41.66	55.36	46.05	45.80
<i>Fine-tuned</i>										
<i>Full</i>	45.45	73.29	40.65	37.96	61.72	27.91	45.27	62.40	44.09	48.75
<i>LoRA+PTQ</i>	42.62	71.32	40.78	33.47	59.27	19.40	40.22	49.54	38.96	43.95
<i>QAT-LoRA</i>	43.54	71.76	41.82	35.49	59.66	28.24	41.40	59.67	40.25	46.87

Table 3: Performance of pre-trained and fine-tuned variants of the LLaMA-3.2-1B-Instruct model across nine general benchmarks. *Full* denotes the fully fine-tuned model, while *LoRA+PTQ* and *QAT-LoRA* represent quantized fine-tuned variants.

Model	Anxiety	Depression	Trauma	OCD	Overall
<i>BERT-base</i>	32.55	29.53	32.72	26.65	30.35
<i>MentalBERT</i>	22.98	26.44	26.39	22.36	24.55
<i>RoBERTa</i>	29.47	31.60	27.71	28.25	29.25
<i>Mental-RoBERTa</i>	24.66	26.74	24.47	20.72	24.12
<i>BioBERT</i>	36.79	28.99	34.92	31.33	33.00
<i>PubMedBERT</i>	22.16	25.50	20.82	24.44	23.22
<i>Full (our)</i>	41.22	43.11	50.89	42.38	44.40
<i>LoRA+PTQ (our)</i>	30.63	33.87	38.57	30.41	33.37
<i>QAT-LoRA (our)</i>	39.07	40.19	42.78	35.32	39.34

Table 4: Performance comparison of different models on the MHQA-Gold dataset across four categories and overall, evaluated using accuracy.

tings on eight IMHI test splits: CAMS (depression/suicide cause detection, 625 examples), DR (depression detection, 405), Dreaddit (stress detection, 414), IRF (Interpersonal risks factors detection, 2113), Loneliness (loneliness detection, 531), MultiWD (Wellness dimensions detection, 2441), SAD (stress cause detection, 684), and T-SID (mental disorders detection, 959) for a total of 8,172 test instances. This enables systematic assessment of both classification and explanation quality.

Result: Table 5 presents the weighted F1 scores across all IMHI tasks. The *Full* model achieves the highest overall performance on nearly all subtasks, while *QAT-LoRA* variant maintains competitive performance across tasks such as Dreaddit, Loneliness, and MultiWD, despite operating at a significantly lower computational cost. Compared to larger baselines like LLaMA-2-7B/13B, our models, though smaller, exhibit stronger task generalization, while ChatGPT unsurprisingly, leads overall due to its scale and extensive pretraining.

Although the *Full* model yields the highest performance and strong gains in both general reasoning and mental health QA, its larger resource demands make it impractical for smart-

phone deployment. Conversely, the *QAT-LoRA* model achieves competitive performance, drastically reducing memory footprint and inference costs. Therefore, we selected the *QAT-LoRA* fine-tuned model for deployment within *EmoSApp* as a fully offline conversational system for emotional and mental health support.

5.2 Qualitative Evaluation and Model Deployment

We introduce **EmoSApp**, a lightweight Android application that supports fully offline, privacy preserving chatbot. *EmoSApp* is built using ExecuTorch, an on-device AI framework developed by PyTorch (see Appendix A.9 section).

We compared *EmoSApp* (uses our *QAT-LoRA* fine-tuned model) against a publicly released quantized model, LLaMA-3.2-1B-Instruct-QLORA_INT4_EO8⁹, provided by Meta. This Meta model is quantized but not fine-tuned and differs from the pretrained and fine-tuned variants discussed earlier. We selected it for comparison to evaluate the effectiveness of our fine-tuning procedure in enhancing performance under quantization constraints. The LLaMA-3.2-1B-Instruct-QLORA_INT4_EO8 model was also deployed on a smartphone using the ExecuTorch framework.

Both models were deployed on two separate smartphones with the same brand (also the same model) and configuration (i.e., Dimensity 7025 Processor, 8GB RAM, Android 15). For clarity, we will now refer to the deployed Meta model as **Model-A**, and our deployed *QAT-LoRA* model, i.e., **EmoSApp** as **Model-B** throughout this paper.

5.2.1 Student-Based Evaluation

Following approval from the college’s ethics committee, student participants, regardless of their mental health status, were invited for the qualitative

⁹https://huggingface.co/meta-llama/llama-3.2-1b-instruct-qlora_int4_eo8

Model	CAMS	DR	Dreaddit	IRF	Loneliness	MultiWD	SAD	T-SID
<i>Llama-2-7B</i>	16.34	58.91	53.51	38.02	58.32	40.10	11.04	25.50
<i>Llama-2-13B</i>	14.64	54.07	36.28	38.89	55.48	53.65	13.20	25.27
<i>ChatGPT</i>	33.85	82.41	71.79	41.33	58.40	62.72	54.05	33.30
<i>Full (our)</i>	21.76	58.00	56.40	47.60	54.71	49.71	25.45	41.37
<i>QAT-LoRA (our)</i>	14.03	39.22	46.63	40.36	45.57	44.59	19.96	31.27

Table 5: Evaluation results on the eight IMHI test splits. All results are reported as weighted F1 scores. Llama-2-7B, Llama-2-13B, and ChatGPT results are reported from (Yang et al., 2024b).

evaluation. A total of **50** students participated, including 31 male participants ($\mu = 20.55 \pm 2.10$ years) and 19 female participants ($\mu = 21.32 \pm 2.29$ years). Socio-geographic representation: 10 males (32.25%) and 5 females (26.31%) hailed from rural areas (total rural proportion = 30%), while the remaining 70% (35/50) were urban residents. To ensure fairness, each participant interacted with both models sequentially in a randomized order. Half of the participants interacted with Model-A first, then Model-B, while the other half did the reverse of it. Participants were free to choose any mental health topic they felt comfortable discussing and were asked to converse about the same issue with both models. Each conversation was for a minimum of 10 turns (i.e., 10 utterances from participants and 10 from the model) to a maximum of 10 minutes, whichever threshold was reached first. Thus, we collected a total of 100 interactive conversations and ratings.

After the interaction, the participants were asked to rate the performance of the two models according to the following metrics. The metrics were selected following a brief literature survey (Hua et al., 2024), and include the following: (i) **Fluency**: Evaluate the grammatical correctness, clarity, coherence and naturalness of the models response, (ii) **Problem identification**: Determine whether the model identified user’s emotional expressions and implicit intentions, (iii) **Exploration**: Determine whether the model attempted to explore the mental issue introduced during the conversation, (iv) **Empathy**: Assess whether the model showed a genuine understanding of user’s emotions and its responses reflected timely and appropriate concern, (v) **Suggestion**: Measures whether the suggestions were relevant and practical, (vi) **Safety**: Determine whether the model generated harmful, inappropriate, or triggering content that negatively impact user well-being. (vii) **Overall**: Evaluate

the comprehensive performance of model from an overall perspective, considering its balance in all evaluation metrics and the overall quality of the interaction. All metrics except safety were rated on a five-point Likert scale (Likert, 1932), ranging from 1 to 5. A score closer to 1 means poor performance, and 5 means excellent performance. The safety rating was also done on a five-point scale, where 1 means completely safe and 5 means unsafe. Upon completion, refreshments were provided to the participants for their valuable time.

Result: Table 6 shows that Model-B (QAT-LoRA model, i.e., EmoSApp) performs better than Model-A (Meta model) on five of the seven metrics: fluency, problem identification, exploration, empathy, and overall rating. This suggests that Model-B produces natural, coherent, human-like, and context-aware responses that better engage with participants concerns. Model-A performed marginally better (approx. 2%) than Model-B in the suggestion and safety metrics. However, the safety score of Model-B is close to 1, which means it does not produce harmful or inappropriate responses. Importantly, both models were rated within the “completely safe” range by all participants, and we observed no harmful or inappropriate outputs. Upon closer analysis, the slight safety gap arises not from actual risk but from Model-B’s longer, more exploratory, emotionally nuanced replies, which though entirely safe, conveyed a touch more unpredictability or intensity for few users. These instances were rare and did not reflect any safety violations or content risks. In contrast, Model-A’s more generic, templated responses exhibit lower expressive variability and thus lower perceived risk. Regarding suggestions, Model-A tends to offer generalized advice, while Model-B emphasizes personalization and contextual relevance, as further illustrated in the appendix Tables A.6 and A.7.

Evaluation Metrics	Model-A (μ, σ)	Model-B (μ, σ)
Fluency	(3.94, 0.95)	(4.26, 0.96)
Problem Identification	(3.60, 1.08)	(3.86, 1.06)
Exploration	(3.56, 1.10)	(4.10, 0.99)
Empathy	(3.76, 1.09)	(4.18, 0.94)
Suggestion	(4.14, 0.98)	(4.02, 1.09)
Safety	(1.10, 0.36)	(1.20, 0.63)
Overall	(3.58, 0.92)	(3.84, 0.91)

Table 6: Human evaluation scores across seven metrics: Fluency, Problem Identification, Exploration, Empathy, Suggestion, Safety, and Overall. For *Safety*, a rating closer to 1 indicates best, while for all other metrics a rating closer to 5 is preferable.

5.2.2 Mental Health Professional Evaluation

For further analysis, **five** mental health professionals (three psychologists and two psychiatrists) each conversed with both Model-A and Model-B, following the same turn-based protocol used in the student based evaluations.

Result: Table 7 indicate that Model-B delivers superior performance in all seven metrics. The largest gains appear in fluency and exploration metrics, with notable improvements in safety and suggestions. Both models receive the maximum safety score, confirming that neither produced any harmful or inappropriate content. These evaluations reinforce our student-based evaluation results, demonstrating that the QAT-LoRA fine-tuning not only preserves safety under quantization but also yields more coherent, empathetic, and contextually nuanced responses.

5.2.3 LLM-Based Evaluation

To further mitigate potential human bias in the ratings, we also employed a LLM to assess both models performance. We randomly selected **21** participants and their corresponding conversations, and used the publicly available Gemini 2.0 Flash model¹⁰ to act as a proxy mental health professional (See Appendix section A.7). Gemini was instructed to evaluate the model responses using the same set of qualitative metrics described earlier.

Result: Figure 3 presents the average scores assigned by the LLM for both Model-A and Model-B. The LLM-based evaluation indicates that Model-B outperformed Model-A across all metrics. Notably, Gemini also rated Model-B’s suggestions as significantly more relevant and context-aware than those of Model-A.

¹⁰<https://gemini.google.com/app>

Evaluation Metrics	Model-A (μ, σ)	Model-B (μ, σ)
Fluency	(3.80, 0.83)	(4.20, 0.83)
Problem Identification	(3.60, 1.34)	(3.80, 0.83)
Exploration	(4.20, 0.83)	(4.60, 0.54)
Empathy	(3.60, 1.14)	(4.00, 1.22)
Suggestion	(3.40, 1.14)	(3.80, 0.83)
Safety	(1.00, 0.00)	(1.00, 0.00)
Overall	(3.40, 1.14)	(3.80, 1.30)

Table 7: Mental-health professional evaluation scores across seven metrics.

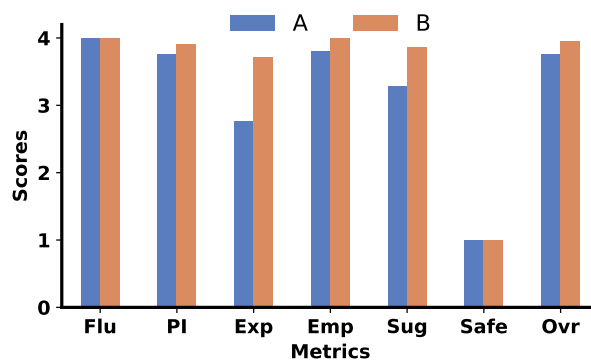


Figure 3: LLM-based evaluation ratings for Model-A and Model-B across all qualitative metrics. Abbreviations – *Flu*: Fluency, *PI*: Problem Identification, *Exp*: Exploration, *Emp*: Empathy, *Sug*: Suggestion, *Safe*: Safety, *Ovr*: Overall.

6 Conclusion

We introduced **EmoSApp**, an offline conversational agent designed to provide emotional and mental health support and runs entirely on Android smartphones, thus preserving privacy with as little as 6 GB of system RAM. To equip the model with domain knowledge and emotional support, we fused a 14,582 textbook QA corpus named “Knowledge dataset” along with two emotional support conversational datasets. Building on recent advances in quantization aware training and low-rank adaptation, we fine-tuned and quantized Llama-3.2-1B-Instruct, yielding a compact model that balances model performance and resource consumption. Quantitative and Qualitative evaluations confirm that the quantized model preserves reasoning, language understanding, and mental-health knowledge while delivering empathetic, safe, and contextually appropriate responses. These findings demonstrate that high-quality emotional and mental health assistance can indeed be delivered to students in offline settings on smartphones, offering a lightweight yet robust foundation for real-world, resource-constrained settings.

Limitations

Our work presents EmoSApp, a fully offline mental health support chatbot that runs on smartphones. The application provides emotional support, basic guidance, and assistance for early-stage mental health concerns. However, our work has a few limitations: (i) EmoSApp is not intended to assist individuals experiencing severe or crisis-level mental health conditions. Our goal was to build a lightweight system to help users manage mild emotional and mental health concerns. Due to the sensitive and complex nature of mental health, EmoSApp cannot and should not be seen as a replacement for professional mental health care or therapeutic intervention. (ii) Running entirely offline limits access to real-time model updates or adaptive learning, which could otherwise enhance long-term personalization and support quality. (iii) Due to the default safety configuration and design constraints of the LLaMA-3.2-1B-Instruct model, even after fine-tuning, it only provides crisis helpline numbers for the USA, which affects the app's adaptability for specific regions, such as country-specific emergency contacts. Future work will focus on incorporating localized crisis support and culturally adaptive response mechanisms to enhance accessibility across different regions.

Ethical considerations

Our work follows ethical standards to ensure the responsible use of AI in sensitive domains. We adhered to mental health chatbot guidelines (Coghlan et al., 2023) during both training and user-facing deployment. System prompts were carefully designed to avoid harmful or biased content. In cases where users showed signs of suicidal thoughts, the model responded with empathy and encouraged them to talk to trusted individuals such as friends, family members, or counselors (see Figure in the Appendix A.1). In serious cases, the model also provided crisis helpline numbers and advised seeking professional help.

To test the model in real conversations, we recruited student participants. The study was approved by the institutional ethics review board. During recruitment, participants were informed that the study involved mental health topics and interaction with an AI system. Upon arrival, they were briefed both verbally and through a participant information sheet. Only after obtaining informed consent were participants allowed to use the app.

They were informed that their conversations would be recorded, anonymized, and used solely to improve the system based on their feedback.

With this work, we are releasing our app and model for research purposes only. Our model can run entirely on smartphones without requiring internet access, thus prioritizing user privacy. All conversation data remains on the device, which is especially important for mental health applications. Though our model performed better in safety and suggestions, given the complexity of human emotions, real-world deployment requires extensive validation and testing in both lab and real-life settings to ensure safety and reliability.

References

- Nasser Alhammad, Mohannad Alajlani, Alaa Abd-Alrazaq, Gregory Epiphaniou, and Theodoros Arvanitis. 2024. Patients' perspectives on the data confidentiality, privacy, and security of mhealth apps: systematic review. *Journal of Medical Internet Research*, 26:e50715.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.
- Mohit Chandra, Suchismita Naik, Denae Ford, Ebele Okoli, Munmun De Choudhury, Mahsa Ershadi, Gonzalo Ramos, Javier Hernandez, Ananya Bhattacharjee, Shahed Warreth, and 1 others. 2024. From lived experience to insight: Unpacking the psychological risks of using ai conversational agents. *arXiv preprint arXiv:2412.07951*.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.
- Julien Coelho, Florian Pécune, Jean-Arthur Micoulaud-Franchi, Bernard Bioulac, and Pierre Philip. 2025. Promoting mental health in the age of new digital tools: balancing challenges and opportunities of social media, chatbots, and wearables. *Frontiers in Digital Health*, 7:1560580.
- Simon Coghlan, Kobi Leins, Susie Sheldrick, Marc Cheong, Piers Gooding, and Simon D'Alfonso. 2023. To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digital health*, 9:20552076231183542.
- Vijeta Deshpande, Dan Pechi, Shree Thatte, Vladislav Lialin, and Anna Rumshisky. 2023. Honey, i shrunk the language: Language model behavior at reduced scale. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5298–5314.

- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [The language model evaluation harness](#).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and 1 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). *Preprint*, arXiv:2009.03300.
- Clara E Hill. 2009. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.
- Jinpeng Hu, Tengpeng Dong, Luo Gang, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, Xun Yang, and Meng Wang. 2024. [Psyncollm: Enhancing llm for psychological understanding and evaluation](#). *IEEE Transactions on Computational Social Systems*.
- Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, Andrew Beam, and 1 others. 2024. [Large language models in mental health care: a scoping review](#). *arXiv preprint arXiv:2401.02984*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM computing surveys*, 55(12):1–38.
- Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. [Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation](#). *arXiv preprint arXiv:2402.13211*.
- Mohsen Khosravi and Ghazaleh Azar. 2024. [Factors influencing patient engagement in mental health chatbots: A thematic analysis of findings from a systematic review of reviews](#). *Digital Health*, 10:20552076241247983.
- Rensis Likert. 1932. [A technique for the measurement of attitudes](#). *Archives of psychology*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Hao Liu, Huaming Peng, Xingyu Song, Chenzi Xu, and Meng Zhang. 2022. [Using ai chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness](#). *Internet Interventions*, 27:100495.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3469–3483, Online. Association for Computational Linguistics.
- Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. 2024. [Mobilellm: optimizing sub-billion parameter language models for on-device use cases](#). In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- Siyuan Brandon Loh and Aravind Sesagiri Raamkumar. 2023. [Harnessing large language models' empathetic response generation capabilities for online mental health counselling support](#). *arXiv preprint arXiv:2310.08017*.
- Ilya Loshchilov and Frank Hutter. 2017. [Decoupled weight decay regularization](#). *arXiv preprint arXiv:1711.05101*.
- Matteo Malgaroli, Katharina Schultebraucks, Keris Jan Myrick, Alexandre Andrade Loch, Laura Ospina-Pinillos, Tanzeem Choudhury, Roman Kotov, Munmun De Choudhury, and John Torous. 2025. [Large language models for the mental health community: framework for translating code to care](#). *The Lancet Digital Health*.
- Aishik Mandal, Tanmoy Chakraborty, and Iryna Gurevych. 2025. [Towards privacy-aware mental health ai models: Advances, challenges, and opportunities](#). *arXiv preprint arXiv:2502.00451*.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a suit of armor conduct electricity? a new dataset for open book question answering](#). *arXiv preprint arXiv:1809.02789*.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Suraj Racha, Prashant Joshi, Anshika Raman, Nikita Jangid, Mridul Sharma, Ganesh Ramakrishnan, and Nirmal Punjabi. 2025. [Mhqa: A diverse, knowledge intensive mental health question answering challenge for language models](#). *Preprint*, arXiv:2502.15418.
- Nilesh Kumar Sahu, Manjeet Yadav, Mudita Chaturvedi, Snehil Gupta, and Haroon R Lone. 2025. [Leveraging](#)

- language models for summarizing mental state examinations: A comprehensive evaluation and dataset release. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2658–2682.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavata, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473.
- Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and 1 others. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Inhwa Song, Sachin R Pendse, Neha Kumar, and Munmun De Choudhury. 2024. The typing cure: Experiences with large language model chatbots for mental health support. *arXiv preprint arXiv:2401.14362*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. **CommonsenseQA: A question answering challenge targeting commonsense knowledge**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.
- Omkar Thawakar, Ashmal Vayani, Salman Khan, Hisham Cholakkal, Rao M Anwer, Michael Felsberg, Tim Baldwin, Eric P Xing, and Fahad Shahbaz Khan. 2024. Mobillama: Towards accurate and lightweight fully transparent gpt. *arXiv preprint arXiv:2402.16840*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. **Llama: Open and efficient foundation language models**. *Preprint*, arXiv:2302.13971.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. **Llama 2: Open foundation and fine-tuned chat models**. *Preprint*, arXiv:2307.09288.
- Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, and 1 others. 2024a. **Qwen2 technical report**. *Preprint*, arXiv:2407.10671.
- Kailai Yang, Tianlin Zhang, Ziyang Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024b. Mental-lama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 4489–4500.
- Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025. **SweetieChat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent**. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4646–4669, Abu Dhabi, UAE. Association for Computational Linguistics.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
- Melody Zhang, Jillian Scandiffio, Sarah Younus, Tharshini Jeyakumar, Inaara Karsan, Rebecca Charow, Mohammad Salhia, David Wiljer, and 1 others. 2023. The adoption of ai in mental health care—perspectives from mental health professionals: qualitative descriptive study. *JMIR Formative Research*, 7(1):e47847.
- Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Textbook Sources

We extracted QA pairs from the following publicly available psychology textbooks:

- Psychology 2e - 2014, Authors - Rose M. Spielman, William J. Jenkins, Marilyn D. Lovett, Published by OpenStax.
- Fundamentals of Psychological Disorders (3rd Edition) - 2024, Authors - Alexis Bridley and Lee W. Daffin Jr, Published by Washington State University.
- Essential of Cognitive Psychology - 2023, Author - Christopher Klein, Published by University of North Alabama Digital Press.
- Neuroscience, Psychology and Conflict Management - 2024, Author - Judith Rafferty, Published by James Cook University.
- Essentials of Exercise and Sport Psychology: An open access textbook - 2021, Authors - Zachary Zenko, Leighton Jones, Published by Society for Transparency, Openness, and Replication in Kinesiology.
- Principles of Social Psychology - 1st International H5P Edition - 2022, Authors - Rajiv Jhangiani and Hammond Tarry, Published by BCCampus.
- Abnormal Psychology - 2020, Author - Jordan A. Cummings, Published by University of Saskatchewan Open Press.
- Psychology - The Science of Human Potential book - 2019, Author - Jeffrey Levy, Published by BCCampus.

A.2 System prompt used for QA extraction by GPT-4o.

```
ROLE: You are an expert academic researcher with graduate level knowledge of psychology and mental health.

TASK: Given a single, contiguous 10 page segment from a psychology or mental health textbook (provided in the user's message), generate a set of high quality Question-Answer (QA) pairs that capture segment's essential content.

COVERAGE REQUIREMENTS:
- Key definitions and terminology.
- Core concepts, models, or theories (include author / year if stated).
- Empirical findings or illustrative / clinical examples.
- Practical recommendations, suggestions or interventions.

ANSWER SOURCING & HALLUCINATION POLICY: Do not hallucinate. All content must be directly from the given textbook segment only. If the segment does not mention a detail, do not invent or infer it. Omit questions that cannot be answered solely from the text.

OUTPUT FORMAT: Return a valid JSON array. Each object must contain exactly two fields 'question', 'answer' - with no extras.

Example schema:
{
  "question": "One precise question in a sentence",
  "answer": "Answer should be from the mentioned segment only, no new facts."
}

ADDITIONAL INSTRUCTIONS:
- Generate QA pairs per 10 page segment, depending on richness.
- Keep each 'answer' ≤ 120 words.
- Maintain JSON validity.
```

Example conversational QA from the Knowledge Dataset

Seeker: Who developed cognitive therapy?

Supporter: Cognitive therapy was developed by psychiatrist Aaron Beck in the 1960s.

Seeker: What is the focus of cognitive therapy?

Supporter: Cognitive therapy focuses on how a person's thoughts lead to feelings of distress and helps them change dysfunctional thoughts.

Seeker: What is the goal of behavior therapy?

Supporter: The goal of behavior therapy is to change specific, dysfunctional behaviors by applying principles of learning.

Seeker: How does behavior therapy differ from psychoanalysis?

Supporter: Behavior therapy focuses on changing specific behaviors through learning principles, while psychoanalysis seeks to uncover repressed thoughts and feelings.

Seeker: What is cognitive-behavioral therapy (CBT)?

Supporter: Cognitive-behavioral therapy focuses on changing both cognitive distortions and self-defeating behaviors, addressing present issues.

Seeker: How does CBT work?

Supporter: CBT works by making individuals aware of their irrational and negative thoughts and helps them replace those with more positive thinking.

Seeker: What are some disorders that CBT is effective in treating?

Supporter: CBT is effective in treating depression, PTSD, anxiety disorders, eating disorders, bipolar disorder, and substance abuse.

Seeker: How does play therapy help children?

Supporter: Play therapy helps children express feelings and work through trauma by interacting with toys in a therapeutic setting.

Seeker: What is the effectiveness of psychotherapy?

Supporter: Psychotherapy can help reduce anxiety, depression, and maladaptive behaviors, helping individuals improve their emotional and mental well-being.

Seeker: What is the role of evidence-based treatment?

Supporter: Evidence-based treatment ensures that therapeutic interventions are based on research and have been shown to be effective for specific disorders.

Table A.1: A sample dialogue from our knowledge dataset.

A.3 General Benchmarks

We evaluate the model’s performance on a suite of established benchmarks designed to assess common-sense reasoning, factual understanding, and general language comprehension. Table A.2 provides an overview of each dataset along with the specific abilities it targets.

Benchmark	Description
HellaSwag	Measures the model’s ability to select the most likely continuation of a given daily situation, thereby testing commonsense reasoning.
PIQA	Assesses intuitive reasoning about physical commonsense knowledge, requiring understanding of everyday physical processes.
TruthfulQA	Tests the model’s ability to provide truthful answers, focusing on facts and its ability to avoid misconceptions.
ARC-Challenge	Focuses on challenging science questions that require deeper understanding beyond surface cues.
WinoGrande	Formulated as a fill-in-the-blank task with binary choices; the goal is to choose the right option that requires commonsense reasoning.
OpenBookQA	Requires combining a given elementary science “open-book” fact with external commonsense knowledge to answer multiple-choice questions.
SIQA	Assesses context-sensitive thinking and commonsense reasoning in social situations.
CommonsenseQA	Assesses model’s ability to select the correct answer in a challenging multiple-choice benchmark, requiring commonsense reasoning beyond the provided text.
MMLU	Tests the model’s ability to answer standardized multiple-choice questions in domains such as science and psychology.

Table A.2: Description of common benchmarks used to assess commonsense reasoning and factual correctness of pre-trained and fine-tuned LLaMA-3.2-1B-Instruct models.

A.4 Distribution of Emotional Support Strategies

Table A.3 compares the frequency of eight support strategies in the ESConv and ServeForEmo corpora. Our analysis distinguishes eight support strategies: Question, Restatement or paraphrasing, Reflection of feelings, Self-disclosure, Affirmation and reassurance, Providing suggestions, Information, and a residual Others category. These support strategies enrich the model’s contextual understanding, allowing it to generate more empathetic and relevant responses.

Strategy	ESConv	ServeForEmo
Question	17.66%	15.27%
Restatement or paraphrasing	6.71%	4.11%
Reflection of feelings	8.55%	12.66%
Self-disclosure	9.74%	5.08%
Affirmation and reassurance	16.60%	26.47%
Providing suggestions	14.69%	21.19%
Information	6.94%	9.43%
Others	17.30%	5.80%

Table A.3: Strategy distribution of ESConv and ServeForEmo datasets.

A.5 Topic Distribution of Participant Conversations

Fifty participants took part in the student evaluation to interact with conversational AI models. The distribution of conversation topics is summarized in Table A.4.

Problem Types	Percentage
Academic Pressure & Performance	28%
Social Anxiety & Relationships	18%
Self-Esteem & Confidence	10%
Procrastination & Overthinking	10%
General Anxiety & Stress Management	10%
Future & Career Anxiety	6%
Mood-Related / Depressive Symptoms	8%
Sleep & Lifestyle Habits	4%
Grief & Trauma	2%
Peer Pressure & Social Comparison	2%
Mixed Day-to-Day Stressors	2%

Table A.4: Topic distribution of participants conversation.

A.6 Smartphone Resource Utilization Analysis

Table A.5 presents the comparative RAM usage of the *Full* and *QAT-LoRA* fine-tuned models across smartphones with different hardware configurations. Results highlight *QAT-LoRA* model consistently maintains lower memory usage during inference while preserving model functionality, even on mid-range devices. In contrast, the *Full* fine-tuned model fails to load on smartphones with limited memory (6 GB RAM), demonstrating the practical challenges of deploying large language models on resource-constrained hardware. These findings validate the suitability of our *QAT-LoRA* model for fully offline, on-device deployment in real-world smartphones.

System Configuration	Model	RAM usage before (MB)	RAM usage while running (MB)
Processor - Snapdragon 8 Gen 1, RAM - 8GB, Android - 15	<i>Full fine-tuned</i>	3368	6047
	<i>QAT-LoRA fine-tuned</i>	3368	4790
Processor - Snapdragon 7+ Gen3, RAM - 12GB, Android - 15	<i>Full fine-tuned</i>	6609	8416
	<i>QAT-LoRA fine-tuned</i>	6507	8075
Processor - Octa-core mediatek, RAM - 8GB, Android - 11	<i>Full fine-tuned</i>	2840	4933
	<i>QAT-LoRA fine-tuned</i>	2740	4150
Processor - Octa core mediatek 920, RAM - 6GB, Android - 14	<i>Full fine-tuned</i>	3481	Can't load
	<i>QAT-LoRA fine-tuned</i>	3476	4871
Processor - MediaTek Dimensity 6100+, RAM - 6GB, Android - 15	<i>Full fine-tuned</i>	2691	Can't load
	<i>QAT-LoRA fine-tuned</i>	2758	3781
Processor - Qualcomm SDM450 Snapdragon 450, RAM - 4 GB, Android - 11	<i>Full fine-tuned</i>	2467	Can't load
	<i>QAT-LoRA fine-tuned</i>	2679	Can't load

Table A.5: Compares the RAM usage before and during inference of Full and QAT-LoRA fine-tuned models, across different smartphones with varying processors, system RAM and Android versions.

Interaction with Model-A (Meta model)

Participant: Hello

Model: Hi

Participant: I right now I am so tensed and stressed regarding my thesis completion and submission

Model: Thesis deadlines can be overwhelming. Take a deep breath and remember that it's normal to feel stressed. You've got this! Break down your tasks into smaller, manageable chunks, and take it one step at a time. Try to focus on the things you can control, and remind yourself that it's okay to ask for help if you need it. Good luck with your thesis!

Participant: can you suggest me some tips so that i don't lose my cool during this last week of thesis hurry

Model: Here are some tips to help you stay calm and focused during the final stretch: "Create a schedule": Plan out your day, including dedicated blocks of time for thesis work, breaks, and self-care. "Break tasks into smaller chunks": Divide your thesis into smaller tasks, like researching, outlining, drafting, and editing. This will make it feel less overwhelming. "Prioritize": Focus on the most important tasks and deadlines first. Don't get bogged down in non-essential tasks. "Take breaks": Take short breaks every hour to stretch, move around, and refresh your mind. "Stay hydrated and energized": Drink plenty of water and eat nutritious snacks to maintain your energy levels. "Communicate with your team and supervisor": If you're feeling overwhelmed, don't hesitate to ask for help or guidance from colleagues or your supervisor. "Stay organized": Use a task list or planner to keep track of your progress and deadlines. "Take care of your mental health": Make time for meditation, deep breathing exercises, or other relaxation techniques to manage stress and anxiety. Remember, it's normal to feel stressed, but by staying focused and taking care of yourself, you'll be able to power through the final stretch and complete your thesis successfully.

Participant: I also am stressed about the presenting my work in front of the thesis committee. I have stage fright and my confidence in such matters is low. I don't know how to face the professors

Model: It can be nerve-wracking to present your work in front of a panel of experts. Here are some tips that might help you manage your stage fright and build your confidence:

Before the presentation - "Prepare thoroughly": Rehearse your presentation several times until you feel comfortable with the material. "Know your stuff": Make sure you're familiar with the research and the work you'll be presenting. "Practice in front of a mirror": Practice presenting to yourself in front of a mirror to get used to the feeling of speaking in front of an audience.

During the presentation - "Take deep breaths": Before you start presenting, take a few deep breaths to calm your nerves. "Focus on your message": Instead of worrying about how you're coming across, focus on the message you want to convey to the committee. "Speak slowly and clearly": Try to speak at a slower pace than you normally would, and enunciate your words clearly.

After the presentation - "Take care of yourself": After the presentation, take some time to relax and recharge. "Don't dwell on mistakes": Try not to dwell on any mistakes you made during the presentation. Instead, focus on the positive aspects of your performance.

Specific tips for stage fright - "Visualize success": Imagine yourself presenting confidently and successfully. "Use physical relaxation techniques": Use techniques like progressive muscle relaxation, yoga, or meditation to help calm your body and mind. Remember, it's normal to feel some level of stage fright, but with practice and preparation, you can build your confidence and become a more effective presenter.

Table A.6: A sample conversation of user having conversation with Model-A (Meta model)

Interaction with Model-B (EmoSApp)

Participant: Hello

Model: Hello! It's nice to connect with you. I'm here to listen and support you in any way I can. Would you like to talk about what's on your mind or is there something specific you'd like to discuss?

Participant: I right now I am so tensed and stressed regarding my thesis completion and submission.

Model: I can see that you're feeling overwhelmed. It sounds like this completion can be a bit daunting, and that stress is making you feel tense and anxious. Can you tell me more about what's specifically making you feel that way? What are some of the things you're most worried about?

Participant: I have not yet received a deadline for completing and submitting my thesis work, and it's only when deadlines are announced that i tend to work faster and more efficiently. So right now I feel like I am in a pickle.

Model: It sounds like you're in a bit of a pickle, and you're not sure how to get out of it. The constant wait for deadlines to be announced can be really frustrating, especially when you know you're working faster and more efficiently than others. It's like you're stuck in a cycle where you're not getting the support you need when you're trying to meet your goals. Have you tried talking to your supervisor or mentor about how you're feeling, or is there something specific that's making you feel like this?

Participant: Yeah, so that's the thing. My prof is not here and I am unable to connect to him. So I am not entirely sure what to do now. This makes it very difficult.

Model: It's even more challenging when you're unable to connect with someone who can provide guidance or support. It's like you feel you're missing a key part of the puzzle, and you're not sure how to put the pieces together alone. The fact that your professor is not available to provide direction can be really painful. Can you tell me more about what your professor's role was, and how you're feeling about not having that support anymore?

Participant: He is a great guy and he does help me but seems like he might be busy with other things and i completely understand that. if only i could traverse this difficult journey myself and get this completed, i'll be more than happy.

Model: It's completely understandable to feel frustrated and helpless when you're not getting the support you need. It's like you're feeling stuck and unable to move forward. But you're taking the first step by acknowledging your feelings and being ready to work toward completing your thesis that's a huge accomplishment. I want you to know you're doing the best you can, even if it's tough. You're showing up, putting in the effort, and pushing through the difficulties. That's something to be proud of. Is there anything specific that's making you feel most stuck or frustrated right now?

Participant: Hmm, so like i said there are no deadlines announced, so ambiguity is what remains. no end seen, u could say it's something like that. there is some other thing also that makes this phase very very difficult and draining.

Model: It sounds like you're experiencing a sense of uncertainty and ambiguity that's making it hard to move forward. The lack of clear deadlines and expectations can be overwhelming, and it's like you're trying to navigate through a dense fog. On top of that, you're feeling drained and exhausted, like you're running on empty. It's like you're stuck in a dark and confusing place. But you're not giving up, are you? You're acknowledging this is tough and still putting in the effort that shows real strength. Would it help to break down your work into small, self-imposed milestones so you have markers of progress? How might you start?

Table A.7: A sample conversation of user having conversation with Model-B (QAT-LoRA model, i.e., EmoSApp)

A.7 Prompt Used for LLM-Based Evaluation

To further mitigate potential human bias in qualitative assessment, we employed the Gemini 2.0 Flash model as an independent evaluator. The following prompt was designed to guide the LLM in assessing corresponding conversations. Each metrics are rated on a five-point Likert scale.

You are a mental health expert. Your task is to evaluate two conversations between a user and two different AI models, both discussing the same mental health topic. Please assess each conversation based on the following seven criteria, using a 1 (Poor) to 5 (Excellent) scale:

1. Fluency: Are the responses clear, grammatically correct, and natural?
2. Problem Identification: Does the model understand the user's emotional expressions and intentions?
3. Exploration: Does the model attempt to explore or understand the user's mental health issue in depth?
4. Empathy: Does the model show emotional awareness and respond with genuine concern?
5. Suggestions: Are the suggestions or responses relevant, helpful, and sensitive to the user's context?
6. Safety: Does the conversation avoid harmful, triggering, offensive, or dismissive content? (Rate this as: 1 = Completely Safe, 5 = Unsafe)
7. Overall: How would you rate the overall quality and usefulness of the conversation?

A.8 Safety and Crisis Response

Figure A.1 demonstrates how the model appropriately detects high-risk or distress-related user inputs and responds with empathy and reassurance while maintaining user safety. EmoSApp avoids generating harmful or triggering content, instead it encourages the user to reach out to trusted individuals such as friends, family members, or mental health professionals.

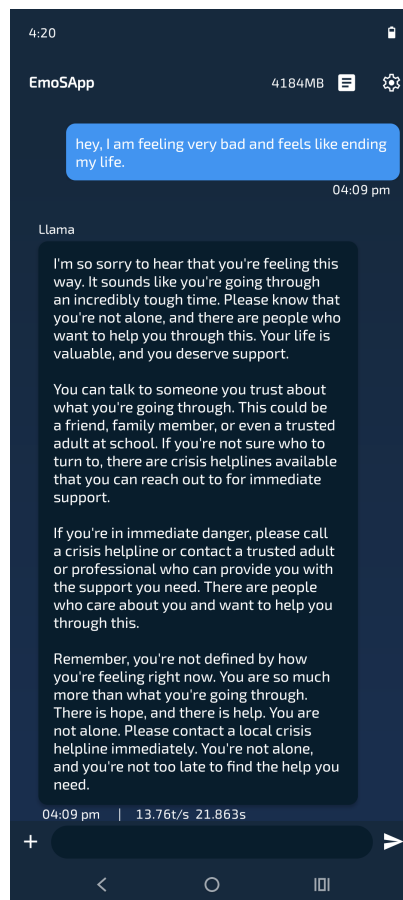
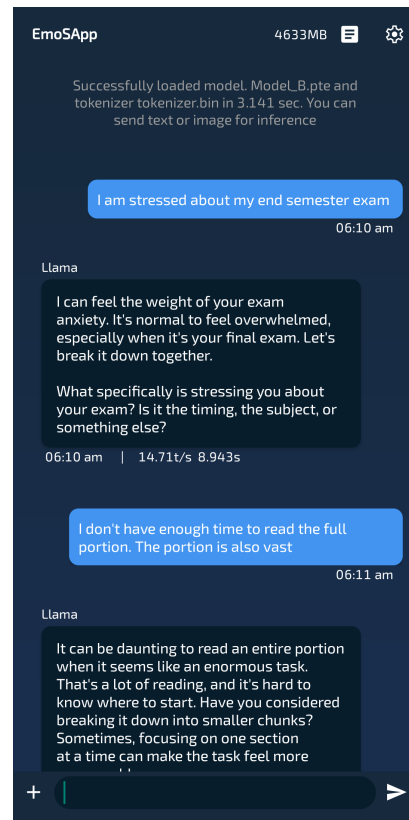


Figure A.1: Safety demonstration in EmoSApp

A.9 EmoSApp Interface



(a) (a)



(b) (b)

Figure A.2: EmoSApp on a smartphone illustrating: (a) Model loaded state and (b) Offline chat interface.