Master Thesis Project 1

SmokeCtrl: An AI-Powered Solution for Tobacco Cessation

Submitted By -

Hardik Soni
20CS30023

Supervised By -

Prof. Jayanta Mukhopadhyay
Department of Computer Science and Engineering

Introduction

Application of AI in HealthCare

- Recent AI advances in NLP and LLMs enable tailored healthcare assistance.

- AI-driven apps now offer personalized advice, answer complex queries, and support behavior change.

- These technologies enhance accessible, effective healthcare support.

The Need for AI-Driven Tools

- Tobacco cessation is essential for reducing smoking-related diseases.

- Traditional programs lack real-time, personalized support.

- AI can provide tailored, immediate assistance for cravings, withdrawal, and quitting commitment.

Support with AI in Healthcare

- AI in healthcare, with RAG and quantized adaptation, outperforms traditional cessation aids.

- These technologies enable efficient, context-aware user interactions.

- Mobile AI boosts accessibility and engagement in tobacco cessation.

Motivation

Aim: Leverage AI-driven tools to address the public health challenge of *tobacco cessation*.

- Tobacco addiction is a leading cause of preventable illness and mortality.
- Current cessation programs lack personalization and accessibility for effective support.
- Healthcare apps provide general support, but advanced AI enables real-time, context-specific responses.
- Large language models (LLMs) offer tailored assistance for users' unique needs and circumstances.
    - **Goal 1:** Deliver scalable, personalized support on mobile platforms through efficient model adaptations.
    - **Goal 2:** Create an empathetic, interactive tool that reinforces users' motivation and commitment to quit smoking.
- Bridge AI technology with real-world healthcare needs.
- Focus on empowering users throughout their smoke-free journey.

Objective

GOAL : To develop and evaluate an AI-driven mobile application that provides personalized support for tobacco cessation, enhancing user engagement and success rates.

Application Development: Develop **SmokeCtrl** app with Flutter frontend and Spring backend for seamless cross-platform functionality.

Back-End Security and Data Management: Implement secure, efficient back-end with Spring for user authentication and data storage.

AI Integration: Integrate Llama 3.2 LLM for intelligent, context-aware user query responses.

Enhanced Response Accuracy: Use RAG with a persistent vector database to improve accuracy and relevance of AI-generated support.

Model Optimization for Mobile: Optimize Llama 3.2 for mobile by quantizing with Quantized Low-Rank Adaptation (qLoRA).

User Interface Design: Develop a user-friendly interface for easy navigation and accessibility in tobacco cessation support.

User Engagement Analysis: Evaluate user engagement and app effectiveness through interaction data and feedback analysis.

Public Health Impact Assessment: Assess SmokeCtrl's impact on public health by measuring its effectiveness in reducing tobacco use and healthcare costs.

Literature Review

The thesis broadly comprises of 3 widely researched fields, namely - Large Language Models(LLM's) and NLP in Healthcare, Retrieval-Augmented Generation(RAG), Quantized Low-Rank Adaptation (qLoRA). Here's a deeper dive into the state-of-the-art techniques used in these respective fields

LLMs in HealthCare

Below is a list of prevalent LLMs in healthcare:

❖ **Text Generation in LLMs:** GPT-3 and T5 generate coherent, context-aware responses across domains.

❖ **Conversational LLMs:** ChatGPT and LaMDA enable personalized healthcare applications, including addiction support.

❖ **Healthcare-Specific LLMs:** *BioBERT* and *ClinicalBERT* provide precise responses on symptoms, treatment, and behavioral health, supporting addiction and cessation efforts.

RAG for LLM's: A Survey

❖ *Gao (2023):* RAG improves LLMs by reducing hallucination and outdated knowledge, integrating external databases for accurate, context-rich, and traceable responses in knowledge-intensive tasks.

❖ *Lewis (2020):* Combining information retrieval with language generation, using both parametric (e.g., BART) and non-parametric memory (e.g., Wikipedia index), may provide more accurate, diverse, and factual responses.

Solution Methodology

Mobile Application Development

- **Frontend:** *Flutter* for cross-platform UI (Android & iOS) with responsive, accessible design.
- **Backend:** *Spring* framework for secure authentication.
- **Database:** *PostgreSQL* for reliable data management.

LlaMA Model Integration

The core AI engine uses the **Llama 3.2** model, configured to generate real-time, contextually relevant responses to user prompts. **Python's** asyncio library handles asynchronous processing, enabling fast and responsive interactions.

Retrieval Augmented Generation (RAG)

Implemented with **LangChain** and **Chroma**, RAG retrieves relevant information to enrich responses with contextually accurate data.

Quantized Low-Rank Adaptation (qLoRA)

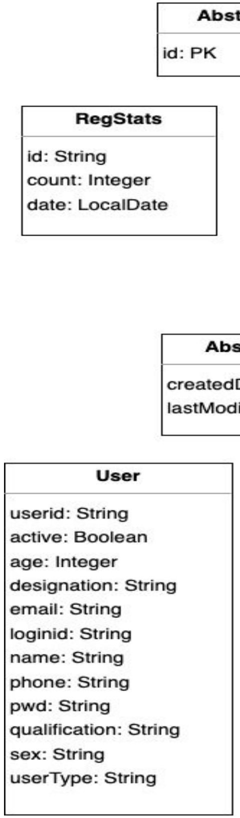
qLoRA fine-tunes large models like Llama 3.2 efficiently, reducing memory and computational demands for real-time, accurate responses on resource-constrained devices.

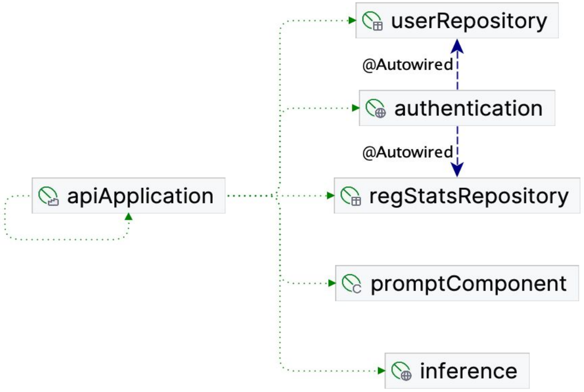
Evaluation

Performance is evaluated using *ROGUE, MMLU, ARC, SQuAD*, and *Hellaswag* metrics, measuring coherence, reasoning, accuracy, and relevance for healthcare support.

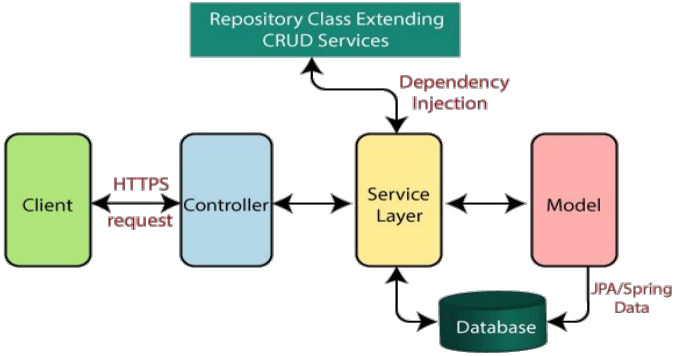
Mobile Application Development

We have developed a dynamic and user-friendly mobile application with a simple, intuitive UI using Flutter for the frontend, Spring for the backend, and PostgreSQL for the database, ensuring efficient storage of user data and registration analytics.



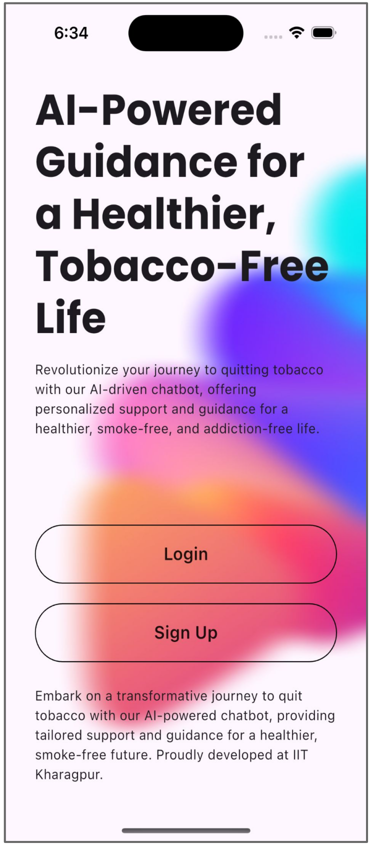


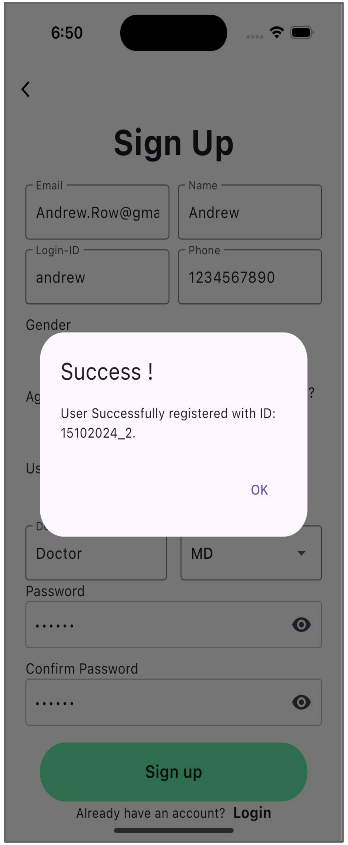
Spring Boot flow architecture

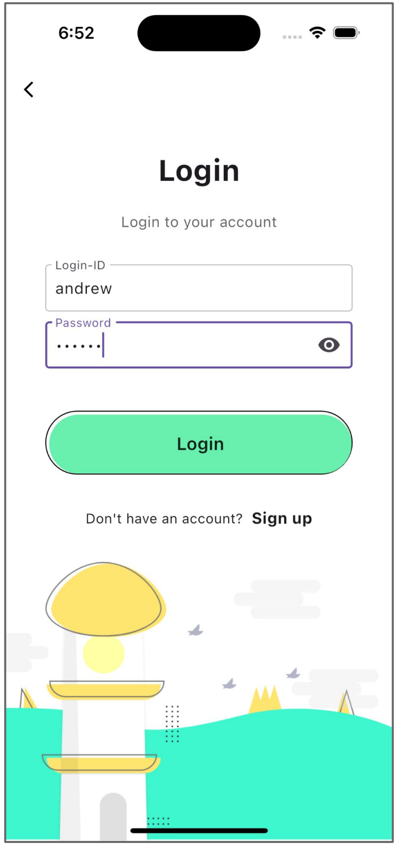


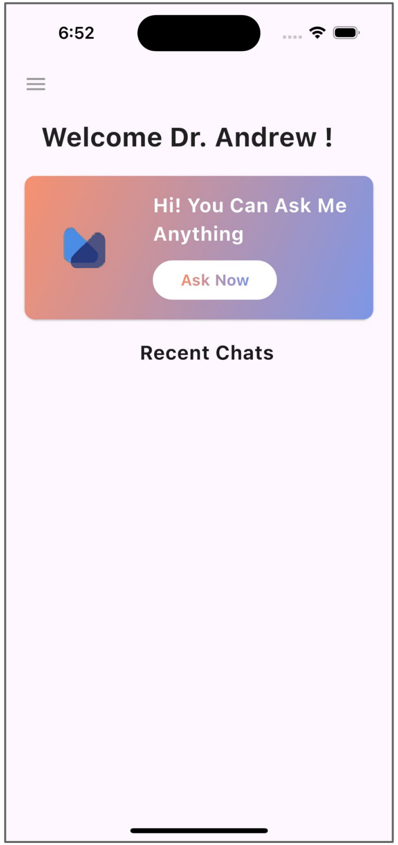
Mobile Application's Interface

Below are all the different components from the application, starting with Home Page, Sign-up, Login, Dashboard Page and Chat Screen.









LlaMA Model Integration

We integrated the **Llama 3.2** model for real-time response generation using the *Llama.cpp* library. To optimize storage space, we utilized the efficient quantized model compression format ".*gguf*".

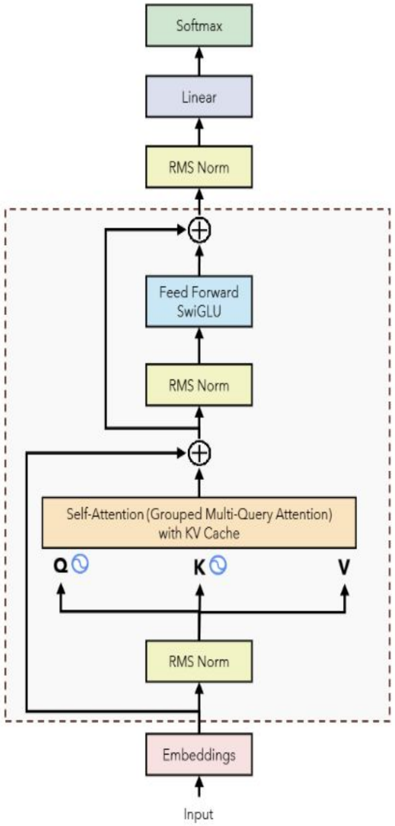
	Training Data	Params	Input modalities	Output modalities	Context Length	GQA	Shared Embeddings	Token count	Knowledge cutoff
Llama 3.2 (text only)	A new mix of publicly available online data.	1B (1.23B)	Multilingual Text	Multilingual Text and code	128k	Yes	Yes	Up to 9T tokens	December 2023
		3B (3.21B)	Multilingual Text	Multilingual Text and code					
Llama 3.2 Quantized (text only)	A new mix of publicly available online data.	1B (1.23B)	Multilingual Text	Multilingual Text and code	8k	Yes	Yes	Up to 9T tokens	December 2023
		3B (3.21B)	Multilingual Text	Multilingual Text and code					

The quantized **Llama 3.2** model maintains the same architecture and parameters as the base version, with the only change being a reduction in context length from **128k** to **8k** tokens. This adjustment results in:

a) improved inference times, offering better efficiency in storage and computation.

b)While the shorter context length *doesn't significantly impact task performance*, it leads to **faster processing** and **reduced resource usage.**

This makes it ideal for real-time applications, especially in resource-constrained environments like one in case of mobile devices.



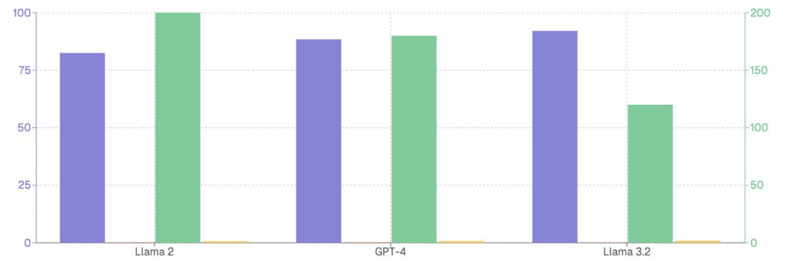
LLaMA

Why LLaMA?

Category	Decode (tokens/sec)	Time-to-first-token (sec)	Prefill (tokens/sec)	Model size (PTE file size in MB)	Memory size (RSS in MB)
1B BF16 (baseline)	19.2	1.0	60.3	2358	3,185
1B SpinQuant	50.2 (2.6x)	0.3 (-76.9%)	260.5 (4.3x)	1083 (-54.1%)	1,921 (-39.7%)
1B QLoRA	45.8 (2.4x)	0.3 (-76.0%)	252.0 (4.2x)	1127 (-52.2%)	2,255 (-29.2%)
3B BF16 (baseline)	7.6	3.0	21.2	6129	7,419
3B SpinQuant	19.7 (2.6x)	0.7 (-76.4%)	89.7 (4.2x)	2435 (-60.3%)	3,726 (-49.8%)
3B QLoRA	18.5 (2.4x)	0.7 (-76.1%)	88.8 (4.2x)	2529 (-58.7%)	4,060 (-45.3%)

Performance and Cost Comparison of Language Models

This chart compares Llama 2, GPT-4, and Llama 3.2 across various performance metrics and cost. Llama 3.2 shows improvements in accuracy, latency, and cost efficiency compared to its predecessors and competitors.



Model	Tasks Supported	Accuracy (%)	Latency (ms)	Throughput (req/sec)	Cost Per Inference ($)
Llama 2	Language Only	82.5	200	1.2	0.10
GPT-4	Multimodal	88.4	180	1.4	0.12
Llama 3.2	Multimodal	92.1	120	1.7	0.09

For tobacco cessation, using an open-source model like **Llama 3.2** locally ensures privacy by processing sensitive data directly on the user's device, avoiding use of external servers. Llama 3.2's efficiency makes it suitable for:

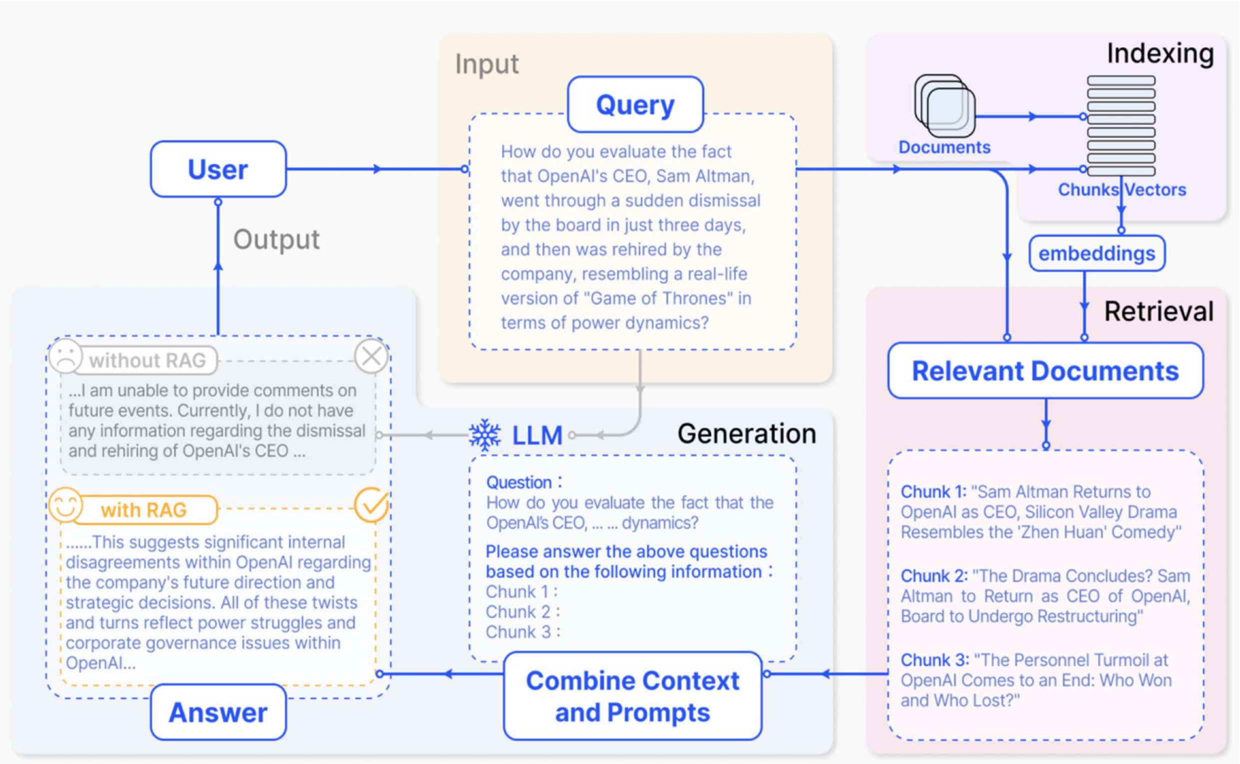
a) *resource-limited* hardware

b) allowing *real-time, private support* without needing powerful cloud infrastructure.

Its multimodal capabilities also enable personalized, interactive support—such as reminders and progress tracking—at a low cost, making it an ideal choice for tobacco cessation assistance.

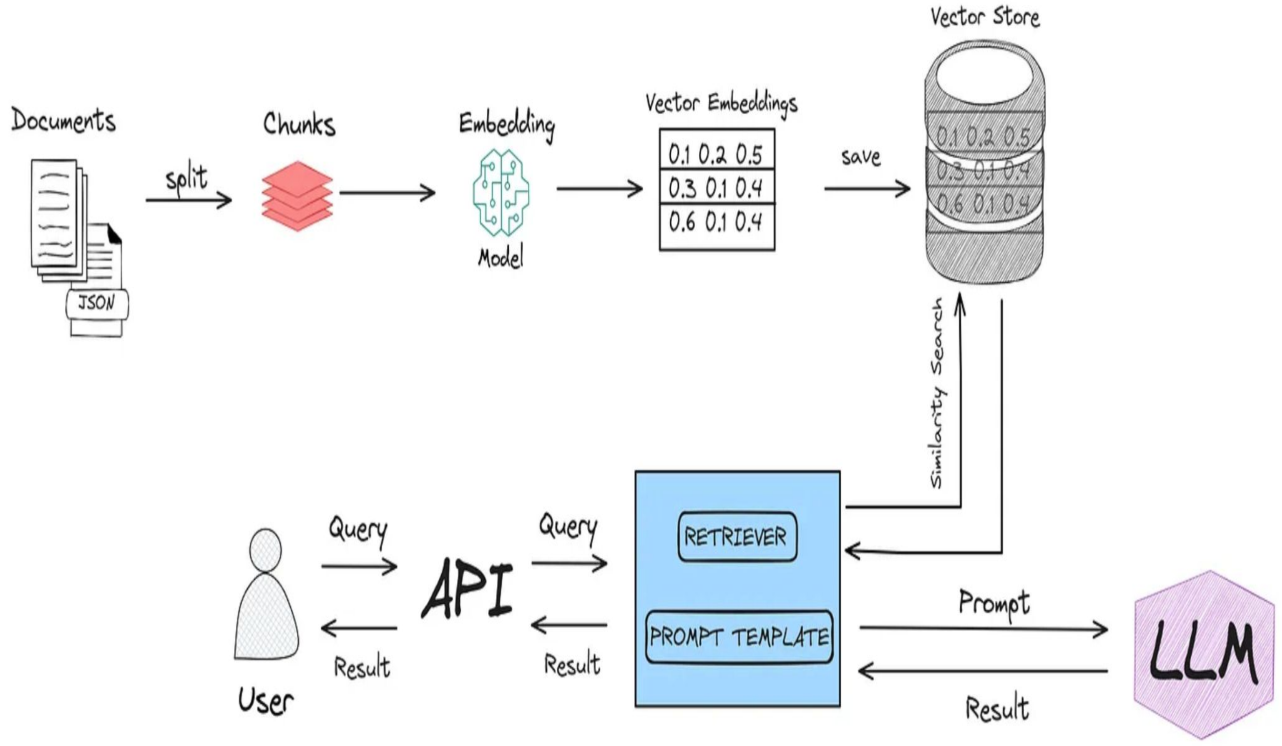
Retrieval Augmented Generation (RAG)

In this project, we implemented Retrieval-Augmented Generation (RAG) using LangChain with a Chroma database to enhance the contextual relevance of responses, leveraging Llama 3.2 for personalized, real-time support in tobacco cessation



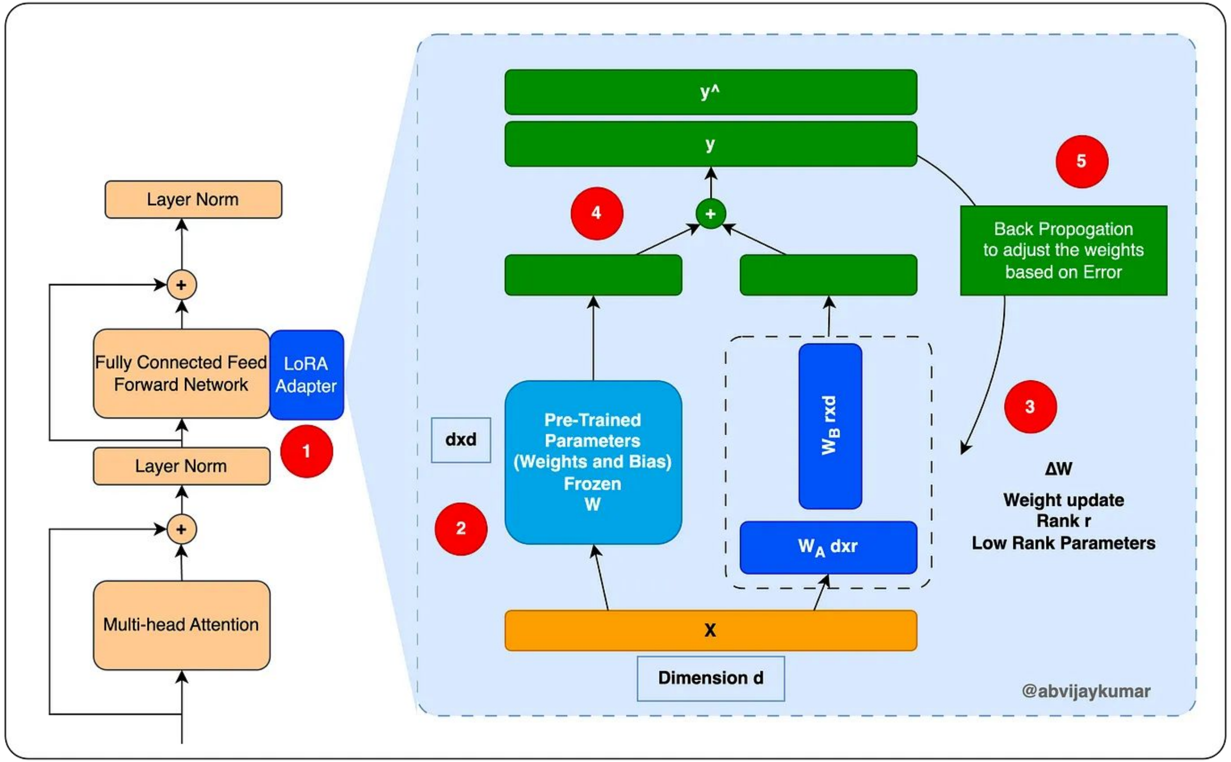
Retrieval Augmented Generation (RAG)

The RAG system uses the **sentence-transformers/all-mpnet-base-v2** model for generating vector embeddings, which are stored in Chroma DB. With LangChain's similarity search, it retrieves relevant, contextually accurate information, optimizing response quality for knowledge-intensive tasks.



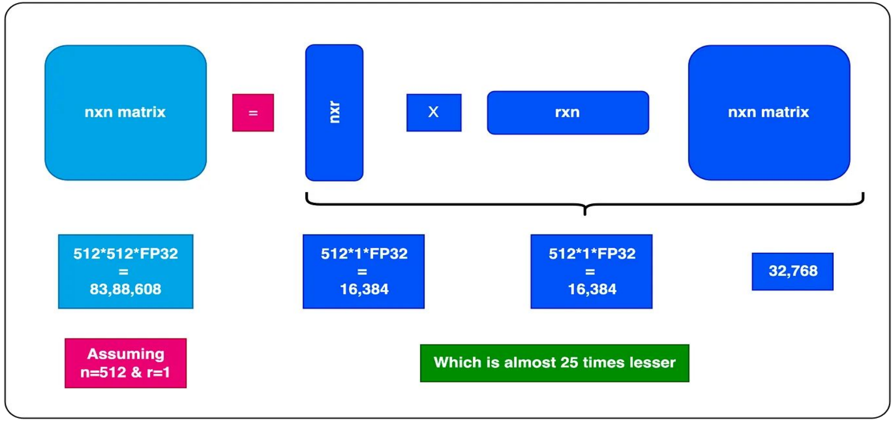
Quantized Low-Rank Adaptation (qLoRA)

We applied **qLoRA** compression on Llama 3.2 to optimize memory and computational efficiency, focusing on essential model parameters for accurate, resource-efficient performance in mobile environments



Quantized Low-Rank Adaptation (qLoRA)

We applied qLoRA fine-tuning to Llama 3.2, optimizing key parameters for accurate, resource-efficient performance in mobile environments.

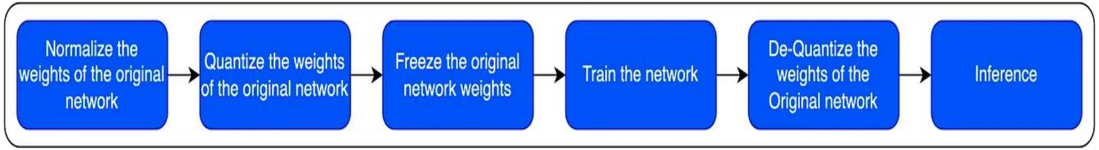


Let's say we have a FP32 weight, with a value of 0.2121. a 4-bit split between -1 to 1 will be the following number positions.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
-1	-0.8667	-0.7334	-0.6001	-0.4668	-0.3335	-0.2002	-0.0669	0.0664	0.1997	0.333	0.4663	0.5996	0.7329	0.8662	1

0.2121 is closest to 0.1997, which is the 10th position. Instead of saving the FP32 of 0.2121, we store 10.

Steps for PEFT-qLoRA



Performance Evaluation

The evaluation metrics used to measure performance include *ROGUE* for text summarization quality, *MMLU* for general language understanding, *ARC* for reasoning capabilities, *SQuAD* for question-answering accuracy, and *Hellaswag* for contextual inference and narrative understanding.

- **ROUGE:** Evaluates the quality of text summarization by comparing overlap between generated and reference summaries.
- **MMLU (Massive Multitask Language Understanding):** Measures general language understanding across a wide range of topics and tasks.
- **ARC (AI2 Reasoning Challenge):** Assesses reasoning abilities, particularly in answering standardized science questions.
- **SQuAD (Stanford Question Answering Dataset):** Tests question-answering accuracy by evaluating responses based on reading comprehension of passages.
- **Hellaswag:** Evaluates a model's ability to perform contextual inference and understand narrative completion by predicting the most likely continuation of a given context.

Capability	Benchmark	# Shots	Metric	1B bf16	1B QLoRA	3B bf16	3B PTQ
General	MMLU	5	macro_avg/acc	49.3	49.0	63.4	60.5
Re-writing	Open-rewrite eval	0	micro_avg/rougeL	41.6	41.2	40.1	40.3
Summarization	TLDR9+ (test)	1	rougeL	16.8	16.8	19.0	19.1
Instruction following	IFEval	0	Avg	59.5	55.6	77.4	73.9
Math	GSM8K (CoT)	8	em_maj1@1	44.4	46.5	77.7	72.9
	MATH (CoT)	0	final_em	30.6	31.0	48.0	44.2
Reasoning	ARC-C	0	acc	59.4	60.7	78.6	75.6
	GPQA	0	acc	27.2	25.9	32.8	32.8
	Hellaswag	0	acc	41.2	41.5	69.8	66.3
Tool Use	BFCL V2	0	acc	25.7	23.7	67.0	53.4
	Nexus	0	macro_avg/acc	13.5	12.5	34.3	32.4
Long Context	InfiniteBench/En.QA	0	longbook_qa/f1	20.3	N/A	19.8	N/A
	InfiniteBench/En.MC	0	longbook_choice/acc	38.0	N/A	63.3	N/A
	NIH/Multi-needle	0	recall	75.0	N/A	84.7	N/A
Multilingual	MGSM (CoT)	0	em	24.5	24.4	N/A	N/A

Category	Benchmark	# Shots	Metric	Llama 3.2 1B	Llama 3.2 3B
General	MMLU	5	macro_avg/acc_char	32.2	58
	AGIEval English	3-5	average/acc_char	23.3	39.2
	ARC-Challenge	25	acc_char	32.8	69.1
Reading comprehension	SQuAD	1	em	49.2	67.7
	QuAC (F1)	1	f1	37.9	42.9
	DROP (F1)	3	f1	28.0	45.2
Long Context	Needle in Haystack	0	em	96.8	1

Curated Performance Scores

Category	MMLU	ARC	SQuAD	Hellaswag	Rogue-L	Rogue-2
Llama 3.2 (1B)	49.3	59.4	49.2	41.2	0.031356	0.0070013
Llama 3.2 (3B)	63.4	78.6	67.7	69.8	-	-
LLama 3.2 (1B) sce.v.1	47.5	55.2	46.7	39.1	0.041926	0.0098073
LLama 3.2 (1B) sce v.2	45.3	52.0	44.5	37.0	0.036292	0.0072468
LLama 3.2 (1B) sce v.c	46.4	53.1	45.5	38.0	0.038109	0.0084276

We fine-tuned the model with three conversation sets: sce.v.1, sce.v.2, and sce.v.c (combined): **sce.v.1 (scenario 1)**, **sce.v.2 (scenario 2)**, and **sce.v.c (their combined version)**. The initial model, fine-tuned on the HuggingFaceH4/ultrachat_200k dataset, was further tuned using curated conversation sets. The following inferences were drawn from the results:

- **General Performance Decline:** Metrics like MMLU, ARC, SQuAD, and Hellaswag dropped, reflecting a shift to specialized language patterns.
- **Improved ROUGE Scores:** ROUGE-L and ROUGE-2 scores increased, indicating better alignment with scenario-specific phrasing.
- **Limitations in Actions and Medication:** The model struggled with system actions and medication inputs due to limited dataset representation.\z
- **Combined Fine-Tuning:** Fine-tuning on sce.v.c improved generalization but reduced specificity for individual scenarios.

Future Considerations: Balancing domain-specific and general tasks, and adding metrics for system actions and medication, could improve performance.

Technology Stack Used

The SmokeCtrl project uses the following technology stack:

Front-End:
- ★ **Flutter 3.x:** Cross-platform mobile development for iOS and Android.

Back-End:
- ★ **Spring Boot 3.x:** Secure back-end support for data management, authentication, and API endpoints.

Language Model & Frameworks:
- ★ **Llama 3.2 (1B-3B):** Core model for personalized responses.
- ★ **LangChain:** Implements Retrieval-Augmented Generation (RAG).
- ★ **Llama.cpp:** Loads quantized GGUF format models.
- ★ **Hugging-Face Transformers:** For embedding generation and LLM integration.

Database & Storage:
- ★ **Chroma Database:** Stores vector embeddings for RAG.
- ★ **PostgreSQL 15.x:** Secure data storage and session tracking.

Fine-Tuning:
- ★ **qLoRA:** Optimizes memory and performance for mobile with 4-bit NormalFloat quantization.

Additional Tools:
- ★ **Python Asyncio 3.11:** Manages asynchronous request handling.
- ★ **Argparse:** Parses command-line arguments for model inference.

Future Work

In my ongoing research for Master’s Thesis Part 2 (MTP2), I'm committed to further exploration in these areas:

Personalized AI Response Enhancement	Enhance response personalization by training models on user feedback, delivering tailored, supportive advice for each individual's cessation journey.
Relapse Risk Prediction and Proactive Support	Integrate predictive analytics to detect relapse risk, enabling timely interventions that keep users committed.
Accessibility Through Multilingual Support	Expand language options to make SmokeCtrl accessible to non-English speakers, broadening its global reach and impact on tobacco cessation.
Community and Peer Support Features	Create community spaces with forums, group challenges, and progress tracking to foster peer support, motivation, and accountability.
Comprehensive Health Tracking and Clinical Partnerships	Gather detailed health data and collaborate with HCW’s to validate *SmokeCtrl's* effectiveness, creating a stronger support network and improving user outcomes.

References

- OpenAI. (2022). ChatGPT: Optimizing Language Models for Dialogue.
- Thoppilan, R., et al. (2022). LaMDA: Language Models for Dialogue Applications.
- Lee, J., et al. (2019). BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining.
- Huang, K., et al. (2020). ClinicalBERT: A Pre-trained Language Representation Model for Clinical Notes.
- Hu, E. J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models.
- Dettmers, T., et al. (2023). Quantized Low-Rank Adaptation for Efficient Model Fine-Tuning.
- Houlsby, N., et al. (2019). Parameter-Efficient Transfer Learning for NLP.
- Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
- Reimers, N., and Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation.
- Google. (2020). Flutter: Beautiful native apps in record time.
- Facebook. (2015). React Native: A Framework for Building Native Apps using React.
- Johnson, R., et al. (2003). The Spring Framework: Simplifying Java Development.

Thank You