POLITECNICO DI TORINO

MASTER's Degree in COMPUTER ENGINEERING

MASTER's Degree Thesis

Design and Implementation of an LLM-Powered Personalized Health Recommender System with Wearable Data Integration

Supervisor

Prof. Maurizio MORISIO

Candidate

Muhammad Sarib KHAN

July 2025

Design and Implementation of an LLM-Powered Personalized Health Recommender System with Wearable Data Integration

Muhammad Sarib Khan

Abstract

Encouraging a healthy lifestyle and avoiding chronic diseases is a persistent public health goal. Particularly, when societies with considerably higher living standards are also affected by sedentary norms. Taking forward the existing Health App project, this thesis enriches the functionality of the app with a Large Language Model-powered personalized recommendation generator.

In this thesis, the challenge of developing an engine that provides applicable, comprehensive, and individualized health and wellness advice is addressed. A cloud-ready and modular recommender system is designed and deployed which can assimilate seamlessly with the already existing Health App's mobile platform. By exploiting heterogeneous user data which includes metrics and logs related to physical activity, sleep, nutrition, and stress which is either acquired by Fitbit, and manually entered by the user, the recommender system yields customized guidance to enhance the health and well-being of the user. Meanwhile, the feasibilities of real-world data ingestion and flexible API deployment was sufficiently demonstrated.

The fine-tuning of Meta's LlaMA2 model using Low-Rank Adaptation (LoRA) method remains central to this thesis. The fine-tuning was performed on a curated dataset that simulated true-to-life doctor-patient conversations. The dataset was balanced and curated based on the information and knowledge from the book: "Outlive: The Science and Art of Longevity – Peter Attia, MD."

Data engineering, crafting conversational prompts, training the model on GPU clusters, and implementing a resilient cloud-based API in combination with real-time inference hosted on RunPod – a cloud infrastructure, is what encompasses the technical pipeline of this project. To evaluate the model, automated metrics including BLEURT and BERTscore were employed in combination with G-Eval which is a qualitative review of the generated recommendations, to assess the helpfulness, accuracy, empathy and relevance to the user query.

As demonstrated by the experimental findings, the recommender system has the ability to provide personalized, safe, and most importantly medically relevant well-being advice, thus manifesting its potential to serve as a virtual assistance tool for individual users, whether it be patients or healthcare professionals. Crucial challenges included managing diverse data and adaptation of model.

Along with documenting the design and implementation of an LLM-powered wellness recommender, this thesis also yields realistic discernment into integrating AI-based customization in mobile health applications and can act as a foundation for current research and clinical verification.

Key Words: Large Language Models, Personalization, Health Recommendation, Cloud Deployment, LoRA, LLaMA2, Conversational AI, Data Integration, User Engagement, Behavioral Change.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context and Motivation

Over the years, the avoidance of chronic diseases and fortification of healthy lifestyle norms have become extremely significant public health objectives. In spite of improvements in the quality of living and healthcare framework, the present-day societies, even those in developed countries, are still facing major issues when it comes to sub-optimal dietary habits, stress management, poor sleeping routines, and sedentary norms. Additionally, the growth in the prevalence of chronic diseases like diabetes, obesity, eating disorders, chronic anxiety and stress, and cardiovascular disorders is directly connected with poor lifestyle routines and choices. This not only makes a sustained shift in the behavior of individuals critical, but also reinforces the importance of preventive strategies.

The increase in the popularity, availability and affordability of wearable devices along with various mobile health applications has paved new ways to encourage and invite the public towards healthy living through individualized advice, and real-time health and vitals tracking. The aforementioned technologies, without a doubt, have brought improvement in collection of data and increased awareness, but transforming this raw data into applicable and personalized recommendations prevails as a notable question. By utilizing the state-of-the-art artificial intelligence (AI) practices, for instance, Large Language Models (LLMs) which have advanced capability to interpret varied user data and providing scalable personalized guidance, this gap can be addressed and fulfilled. By combining behavioral science, comprehensive data analysis, and cutting edge AI-powered recommendation system in an easy-to-use mobile platform, this thesis targets to uplift the users to make healthy and sophisticated lifestyle changes.

1.2 Problem Statement

Modern wearables like Fitbit, and mobile applications indeed have brought serious innovations in digital healthcare and wellness by making it extremely convenient to track health metrics. Despite this, an average user still find it difficult to put all of this information to use and craft actionable lifestyle guidelines for themselves. Additionally majority of the apps in this domain offer generic and non-personalized guidance which does not cater for the uniqueness in each user's profile, activity, or real-time dynamics. Hence, it is clear that a system is needed which has the ability to provide:

- **Customized Recommendations** predicated on diverse user data (sleep, stress, physical activity, nutrition, etc.).
- **Intelligence that is flexible and adaptable** to fulfill various user demands, goals, and medical queries.
- **Incorporation with real-life mobile applications** and harmony with user data sourced from wearable and manually logged data by the user themselves.

1.3 Objectives

As mentioned before, the ultimate goal of this thesis lies in designing, and implementing an LLM-based recommender engine which is based on cloud and serves the purpose to give tailored health and well-being advice. And on top of that, it can be smoothly integrated in our Health App. Further particular objectives include:

1. **Data Schema Design:** Developing a data schema and pipeline that facilitates manually entered user data, and Fitbit data remained critical. Focus was to cover metrics like physical activity, sleep, food logs, and stress.
2. **Dataset Fabrication:** Building a dataset which has single-turn, doctor-patient style conversation. The medical, health, and wellness knowledge was extracted from the book, "Outlive: The Science and Art of Longevity – Peter Attia, MD."
3. **Fine-tuning LLaMA2 and Personalization:** Finetuning LLaMA2 using LoRA (Low-Rank Adaptation) technique on the aforementioned dataset in order to produce individualized, empathetic, applicable, and most importantly, medically relevant recommendations.
4. **API Deployment:** Using RunPod and FastAPI to develop a cloud-based and robust API which has real-time inference ability and seamless integration with the mobile client.
5. **Evaluating the model:** : Utilizing the automated measures such as BLEURT, BERTScore, G-Eval, along with qualitative ones which focus on helpfulness, accuracy, empathy, and medical relevance to evaluate our model.

As of now, the deployment of recommender as a standalone backend API has been achieved.

Chapter 2

Background Work and Literature Review

In this chapter, the overview of the research and literature related to the large language models, personalized health recommendations, and the applications of Artificial Intelligence in stress management, sleep, nutrition, and physical activity is provided and discussed. Furthermore, the techniques to evaluate and validate the results of LLM-based applications are also examined. Since the wide availability of open-source LLMs has made them increasingly accessible and, hence, fine-tunable, their utilization in a plethora of context-based, low-resource applications such as health and well-being apps appears to show a lot of promise. And the aforementioned domains lay down the conceptual and technical basis to develop an open-source LLM-based recommender system for our health app. This chapter also explains the comparison between single model and hybrid-model recommendation approaches. And lastly, the importance of follow-up questioning to generate personalized recommendations and the potential to achieve this through prompt engineering is also considered.

2.1 LLMs in Healthcare and Well-being

2.1.1 General Capabilities of LLMs

In various natural language processing (NLP) tasks such as translation, summarization, and question answering, LLMs like GPT-3, LLaMA, and PaLM have exhibited significantly advanced performance. These models which are built on transformer architecture and trained on huge datasets, frequently have encompassed hundreds of billions of tokens. Even with all the versatility and fluidity, the general-purpose LLMs lack in domain specificity, which becomes an extremely important aspect when applications in healthcare and wellness are considered [1].

2.1.2 Domain-Specific Medical LLMs

Medical LLMs are basically general purpose LLMs fine-tuned on biomedical corpora. They are developed by researchers to tackle the limitations of general purpose LLMs in the context of medical queries and use cases. Two of the examples of these LLMs are PMC-LLaMA [2], and MedAlpaca [3]. Trained on PubMed Central articles, PMC-LLaMA has demonstrated greatly improved factual accuracy and significantly less hallucination when posed with clinical questions. Instruction-tuned on clinical dialogues and notes, MedAlpaca also has increased capacity for task in the healthcare domain. The responses in these models are found to be very reliable in domains that are naturally high stakes like the analysis of symptoms, communication with patients, and feasible treatment alternatives. But due to the rigidity of medical LLMs and lack of available resources, maintaining multiple LLMs would have been challenging and inefficient. Hence, in this thesis, the choice of fine-tuning LlaMA2 on high-quality structured well-being content looked more feasible. LlaMA2 is a multilingual, general-purpose, and open-source LLM.

2.1.3 Hybrid and Comparative use of LLMs

Medical domain LLMs such as PMC-LLaMA and MedAlpaca, no doubt, have offered superior factuality, however their adaptability remains a question and require more intense resources than an open-source, general purpose LLM, which in our case is LLaMA2. Initially the thought to develop a hybrid pipeline in which a medical LLM like PMC-LLaMA was to be paired with a general purpose LLM like LLaMA2 for more robust results and personalization. But, due to this approach being resource intensive, and to keep our pipeline and model lightweight, a practical approach of fine-tuning LLaMA2 using LoRA method on high quality health and wellbeing data extracted from the book was taken. That being said, hybrid evaluation can be a fruitful direction for future work.

2.2 LLMs for Personalised Well-being Recommendations

2.2.1 Stress Management

In “Wearable Meets LLM”, Neupane et al. propose a novel architecture. It combines prompt templates which are relevant to the context with physiological data from the sensor to generate personalised stress management advice [4]. Either way, studies like these provide invaluable structure for prompting, and evaluation techniques for the modelling of stress related inputs from the end user.

2.2.2 Sleep

To integrate wearable data with LLM-generated recommendations for healthy sleep routines, PhysioLLM [5] was developed. By utilizing a retrieval-augmented generation (RAG) pipeline, it attempts to ensure medically sound outputs. A model like this can be used in a hybrid setup where a fine-tuned general LLM produces more engaging feedback. And also this type of model can be used in a comparative design framework, where outputs of both models can be evaluated using the right metrics.

2.2.3 Nutrition and Physical Activity

In HealthGenie [6], to enhance the reasoning of LLMs, knowledge graphs are used. This led to more adequately personalised meal plans for the user. Similarly for physical activity, role-based prompting is employed in GPTCoach [7] to deliver weekly fitness plans and prompts.

2.3 Lightweight Fine-Tuning and Prompt Strategies

2.3.1 Challenge of Full Fine-Tuning

Due to a massive number of parameters, fully fine-tuning an LLM is computationally expensive. This poses as a barrier in academic environments which are mostly low resource.

2.3.2 LoRA and Parameter-Efficient Tuning

To overcome this challenge, Low-Rank Adaptation (LoRA) [8] was proposed. By freezing the base model and introducing trainable adapters, this method allows lightweight tuning. In our case, the fine-tuning of general LLMs like LLaMA 2 on structured data from wellness-based books can be done using LoRA method while preserving feasibility.

2.3.3 Prompt Engineering

Prompt engineering is a valuable technique to shape the output of the fine-tuned models through few-shot examples and/or task instructions. To make recommendations more personalised, a follow-up questioning behaviour to gather more relevant information from the user can be embedded in LLM's behaviour by the virtue of instruction tuning and role prompts. This can make the experience more human-like and subjective.

2.4 Evaluation and Validation of LLMs

To evaluate our model BLEU, ROUGE, BLEURT [9], and BERTScore [10] are some of the some of the state-of-the-art automated metrics. Due to superior performance of metrics that calculate semantic similarity and are neutral like BLEURT and BERTScore, they became our metrics of choice. Additionally both these metrics correlate better with human preferences. Furthermore, as explained in LLMCheckup [11], using LLM-as-a-judge method is also increasing being used, where an already established LLM, for instance, GPT-4, is meticulously prompted to judge and rate outputs along the lines of outputs being helpful, customized, and empathetic. Hence, this thesis, both automated metrics in the form of BLEURT and BERTScore, and the technique of using LLM-as-judge by using G-Eval [12] as the judge are employed to reliably evaluate and validate the model's outputs. By doing this we attempt to ensure the quality of recommendations generated by the model.

2.5 Gaps in the Existing Literature

The research surrounding the use of LLMs in healthcare is pacing up. So far majority of the work has focused on either static Q/A or medical LLMs. Only a handful of studies have assessed LLMs like LLaMA2 which is an open-source, general purpose LLM, especially, when it comes to efficient parametric fine-tuning for health and wellness guidance and recommendations. Additionally, the extensive use of LLM-as-a-judge, like G-Eval, in combination with multiple automated metrics like BLEURT and BERTScore as illustrated in this thesis is still a rarity in health and wellness focused applications.

Chapter 3

Dataset Design and Preprocessing

To have an accurate and reliable model, it is essential that it is trained on data that is structured and high quality. In this chapter, the procedure of developing and finalizing the dataset for the model is discussed and explained. It covers all the improvements in quality made along the way and finally how it was manually checked to make sure it is structured and realistic.

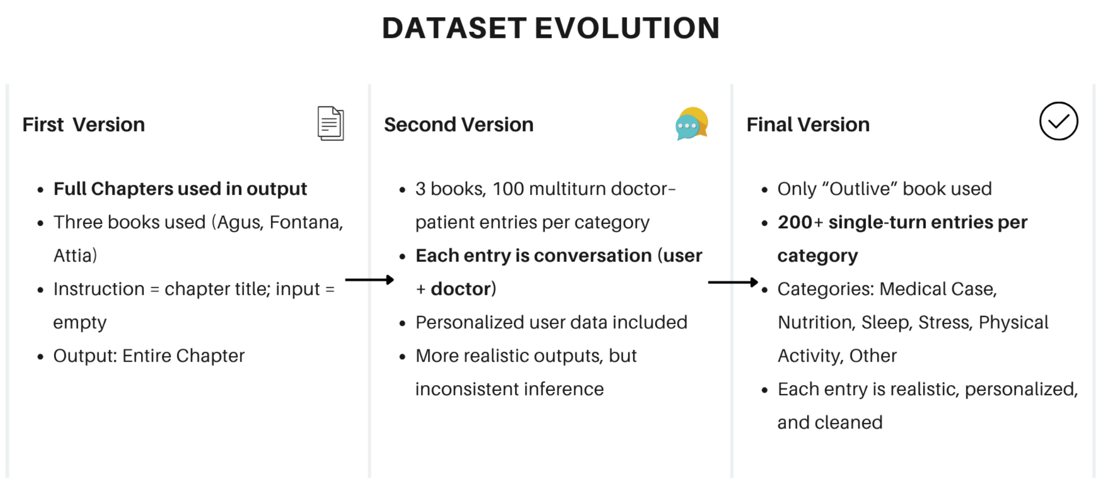

Figure 3.1: Evolution of the dataset across development stages.

3.1 Evolution and Development of Dataset

Each data point was a JSON with three fields: instruction, input, and output. The Chapter's title would take up the "instruction" field and the content would go in the "output" while the input remains empty Due to the extremely simple and naïve nature of this data, the fine-tuning was sub-par, and when prompted, the model either hallucinated, or gave very illogical answers.

3.1.1 Initial Dataset: Full Chapters

To prepare the initial dataset, I followed an extremely simple strategy of copying the whole chapters in the "output" field of a data point.

- *A Short Guide to a Long Life* by David B. Agus
- *The Path to Longevity* by Luigi Fontana
- *Outlive: The Science and Art of Longevity* by David Attia, MD

Each datapoint was a JSON with three fields: instruction, input, and output. The Chapter's title would take up the "instruction" field and the content would go in the "output" while the input remains empty Due to the extremely simple and naïve nature of this data, the fine-tuning was sub-par, and when prompted, the model either hallucinated, or gave very illogical answers. An example of an entry in the jsonl file of the dataset looks like:

```
{"instruction": "Get your annual flu shot", "input": "  ",
"output": "entire chapter.."}
```

An example of the inference result is shown below:

> **User:** I haven't been sleeping well and feel mentally drained
>
> **LLaMA 2:** I haven't been sleeping well and feel mentally drained. I had a migraine last night that kept me up. I got up early, drank coffee, and took my vitamins. I had a headache for most of the day, but it's gone now. I'm starting to feel a little better. I'm glad I'm not on a deadline today. I got a lot of sewing done yesterday. I made the binding for the last quilt. I did the hand sewing on the binding. It was a little difficult because I had to do it in the dark. I couldn't see what I was doing. I sewed it with a dark thread. It's not really dark, but it was the best I could do in

3.1.2 Second Dataset: Multiturn Conversation

To improve the quality of dataset, it was revamped. It was decided that instructions will be restricted to 6 main categories which were Medical Cases, Nutrition, Stress, Sleep, Physical Activity, and Others. Each category will have a 100 entries to keep the dataset balanced Relevant information was taken from the same three books. This time each entry was basically a multi-turn doctor patient conversation. For example:

```
  {
 "instruction": "Nutrition - Chronic Fatigue and B12",
 "input": "<user> I've had chronic fatigue and brain fog for months.
 My iron levels were fine, but I still feel off.
<model> Did your doctor test vitamin B12 or folate? Do you eat animal
products regularly?\
<user> I eat some dairy but no meat or fish.
Not sure about B12 testing.
<model> Any tingling in your hands or feet, or memory issues?\n
```

```
<user> Yes, some numbness in my fingers lately.",
  "output": "<model> That raises concern for B12 deficiency.
  I'd order B12
  and methylmalonic acid tests. If confirmed low,
  you'll need high-dose oral B12 (1000-2000 mcg/day) or possibly
  intramuscular injections.
  Increase intake of eggs, dairy, and fortified foods.
  If you remain vegetarian, we'll monitor levels every 6-12 months.
  Brain fog and fatigue often improve significantly within weeks of
  correcting B12."
}
```

Even though, this data is more complex and richer, the training did not go as expected and model hallucinated heavily and the results were discarded.

3.1.3 Finalized Dataset: Single-Turn

In order to not compromise on the quality of data, ensure personalization of recommendations, and maintaining the right level of complexities these critical steps were taken:

- The knowledge was drawn from only a single book: "Outlive: The Science and Art of Longevity – Attia."
- Categories remained same but each category now has 200 entries
- Each entry was a single-turn conversation

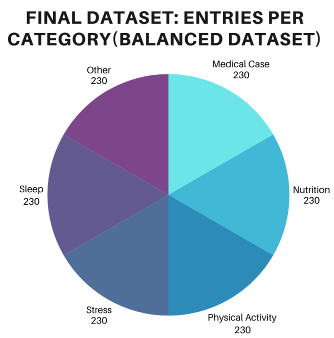

Figure 3.2: Final dataset: entries per category (balanced dataset).

3.2 Data Schema, Cleaning, and Annotation

The data schema was designed in the similar way as the initial two datasets. But some changes were made. Since we want the recommendations to be personalized, it is important that our model is trained with data that also has a field for user's data. Our concept is that when the user enters its query, the health data from app also goes with it in the model's input and then a recommendation is generated. So we trained our model with entries that also had user data in the input. The user data was generated manually and for training it is not sourced from a real-life source like Fitbit. The three fields of the data entries are:

1. Instruction: This mentions the main category with a subcategory.
2. Input:
   - `user_query`: Basically, user's question or concern (e.g. "My cholesterol is high, what should I do?")
   - `user_data`: user's information like age, gender, weight, sleep data, physical activity data, stress score, heart rate, HRV, food logs, etc)
3. Output: Recommender giving a doctor style recommendation which is personalised, medically correct, actionable, and empathetic.

Annotation was done manually to make sure that every entry is realistic and natural. Data was also review multiple times to rule any grammatical mistakes.
In the cleaning process, it was made sure that all, if any, duplicate entries are removed. Data is realistic, for example, a 10-year-old kid cannot be 100kg. Each response was individually studied and was made concise to avoid them being too verbose.
So, it started from a very generic dataset to multiturn to finally finding a balance. In all the process, annotation was manual and keeping the dataset balanced and clean was pinnacle to ensure that our model performs reliably and practically.

3.3 Example of Entries in Final Dataset

Instruction	**Input**	**Output**
Medical Case – Diabetes (Missing Activity)	"user_query": "I have diabetes but haven't tracked my activity. What should I focus on?" "user_data": {"age": 50, "sex": "female", "weight": "93kg", "height": "158cm", "goal": {"activity_goal_min_per_day": 30, "sleep_goal_hr_per_day": 7, "personal_goal": "manage diabetes"}, "activity_summary": {}, "sleep_summary": {"avg_sleep": "6h 10m", "REM": "70 min", "Light": "130 min", "Deep": "24 min", "sleep_consistency": "medium", "wake_episodes": 3}, "stress_level": "high", "food_log": [{"category": "whole grains", "amount": 35}, {"category": "vegetables", "amount": 42}]}	Your activity data is missing, but regular walking, even in short sessions, can greatly help with blood sugar management. Your nutrition and sleep look okay. Please try to log your steps or workouts in the app for more tailored advice next time. OTC fiber supplements can also support blood sugar control.
Nutrition – Red Meat Reduction	"user_query": "How much red meat is too much for health?" "user_data": {"age": 41, "sex": "male", "food_log": [{"category": "red meat", "amount": 300}], }	Your red meat intake is above the recommended weekly limit for heart health. Try to keep it below 350g per week, substituting with fish or plant proteins when possible. Increased whole grains in your meals will further help with cholesterol.
Sleep – Poor Sleep Due to Device Use	"user_query": "I scroll on my phone for an hour in bed every night. Is this hurting my sleep?" "user_data": {"age": 19, "sex": "female", "weight": "57kg", "height": "161cm", "goal": {"activity_goal_min_per_day": 15, "sleep_goal_hr_per_day": 8, "personal_goal": "wake up early"}, "activity_summary": {"steps_per_day": 2500, "active_minutes": 4, "workouts_this_week": 0, "heart_rate_zones": {"zone_2": "0 min", "zone_3": "0 min", "zone_4": "0 min"}}, "sleep_summary": {"avg_sleep": "7h 15m", "REM": "68 min", "Light": "150 min", "Deep": "21 min", "sleep_consistency": "moderate", "wake_episodes": 2}, "stress_level": "medium", "food_log": [{"category": "dairy", "amount": 1}]}	Blue light from phone screens can suppress melatonin, the sleep hormone. Try to put your phone away at least 30 minutes before bed. Consider a physical book or soothing music instead.

Table 3.1: Examples of structured, personalized entries in the final dataset.

Chapter 4

Model Selection and Fine-Tuning

4.1 LLaMA 2 & LoRA

Choosing the right model for our recommender system was crucial to ensure satisfactory results. Several options were considered, for example, Mistral-7B, LLaMA2-13B, PMC-LLaMA, MedAlpaca, but LLaMA2-7B was chosen as the base LLM. First of all it is an open-source and a well-documented LLM which supports parameter-efficient fine-tuning which made It suitable for our academic and experimental use. It has consistently delivered great results across a range of natural language processing tasks, including and especially, conversational and dynamic question-answer tasks. And also, it does not require heavy computational resources while keeping up its performance.
The choice of using LoRA method for fine-tuning was taken due to multiple reasons. Primarily because it allows the adaptation of large models with far fewer trainable parameters, hence, reducing computational cost and memory usage. This also allows rapid experimentation to run within limited time and hardware constraints. Finally, it seamlessly integrates with Hugging Face transformers library leading to easy implementation and deployment.

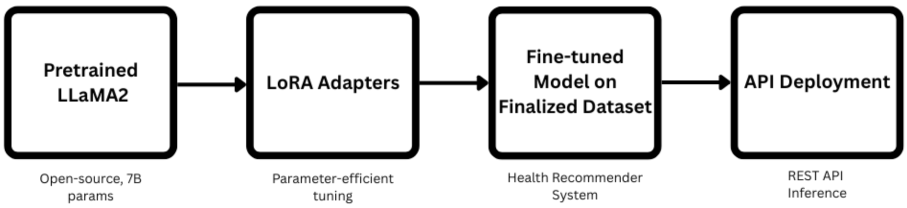

Figure 4.1: Model architecture and fine-tuning pipeline: Pretrained LLaMA 2, LoRA adaptation, domain-specific fine-tuning, and API deployment.

4.2 Training Pipeline and Infrastructure

Using modular and scalable training pipeline, the model's fine-tuning was supervised.

- **Data Preparation**: The finalized dataset was formatted in JSONL file with the three fields: instruction, input, and output
- **Model Initialization**: Model (LLaMA2) was loaded with Hugging Face Transformers and, of course, LoRA adapters configured for parameter-efficient fine-tuning.

Hyperparameter	Value
Model	LLaMA2-7B
Fine-tuning Method	LoRA
Learning Rate	1e-5
Batch Size	1
Number of Epochs	10
Max Tokens per Sample	1024
LoRA Rank	8
Validation Split	10%
GPU	NVIDIA A100 SXM

Table 4.1: Main hyperparameters and training setup for fine-tuning LLaMA2.

By leveraging cloud GPU resources, specifically A100 SXM, training was performed on RunPod. Python scripts were prepared for configuration, model, dataset, and trainer. It was made sure that the logging of training metrics and checkpoints is automated.

4.3 Model Validation& Monitoring

To make sure the training and evaluation go by smoothly, it was made sure that multiple tools and only best practices are engaged. Weights & Biases (WandB) platform was used to monitor real-time changes in loss curves, validation metrics, and system resources during training. This also made comparison of different training runs and hyperparameter configurations fairly simple. And last provided the visualizations to spot any overfitting.
10% of the dataset was used for periodic validation checks during training. Training checkpoints were saved at regular intervals, and code files were version-controlled using GitLab for full reproducibility.
The use of LLaMA2 and LoRA, propped up by stable and robust infrastructure and observations, lead to efficient and effective fine-tuning of the model. Hyperparameters were carefully selected, and monitoring was done systematically. Finally keeping the code modular made sure that the trained model shows high-performance and is read for deployment as the backbone of health recommender system.

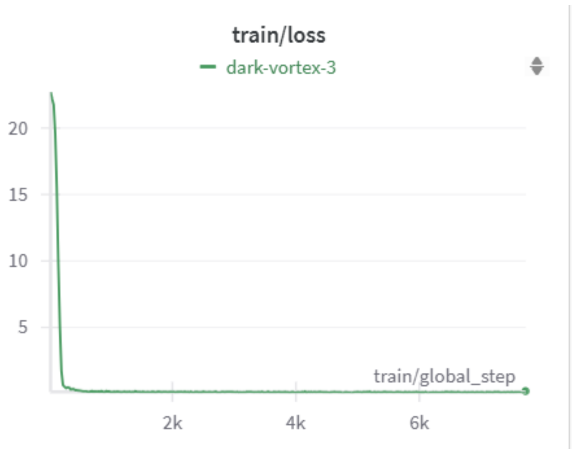

Figure 4.2: Training loss curve over steps.

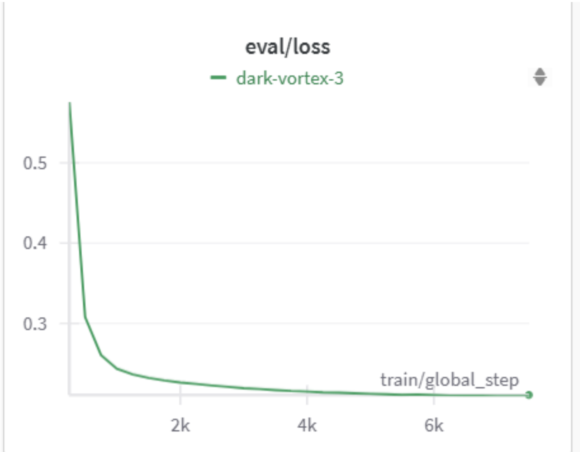

Figure 4.3: Evaluation loss curve over steps.

Chapter 5

System Implementation

The end-to-end implementation of the health recommender system from the backend to cloud deployment is discussed below. The focus was to create a scalable, modular, and maintainable system.

5.1 System Architecture

Modular approach was taken to design the system, hence, machine learning inference, REST API, and user interface components are managed separately. (at the time of writing this report, connection with mobile client hasn't been made). Keeping this structure enables independent and efficient development of each module.

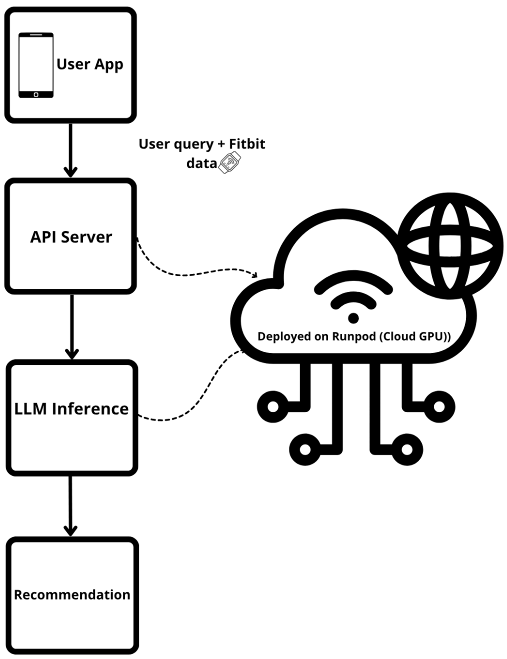

Figure 5.1: System architecture of the health recommender, showing the flow of data from the user app and Fitbit into the cloud-hosted API server and LLM inference engine, deployed on RunPod.

5.2 API and Backend Design

To build the backend, FastAPI was used. It is a state-of-the-art Python web framework which is optimised for building robust, high-performing APIs. API exposes two main APIS. First being /recommend. It takes a POST request with a JSON payload that contains user query and health data and returns a personalized recommendation from LLM. Second being /health which It simply returns a status message and is used for monitoring the availability of the service in deployment environments.

To validate the input a Pydantic model (RecommendRequest) is used to ensure that all necessary data is provided in the right format, hence reducing the risk of runtime errors and increasing reliability.

As server sets up, a dummy inference is run to "warm-up" the model to reduce latency for first real user request. It is a common practice during the deployment of LLMs in production environments. When a request is sent to /recommend by the client, the backend extracts the relevant field, put it in correct format, and triggers the run_inference function. The model finally generates a recommendation and returns a JSON object in the response.

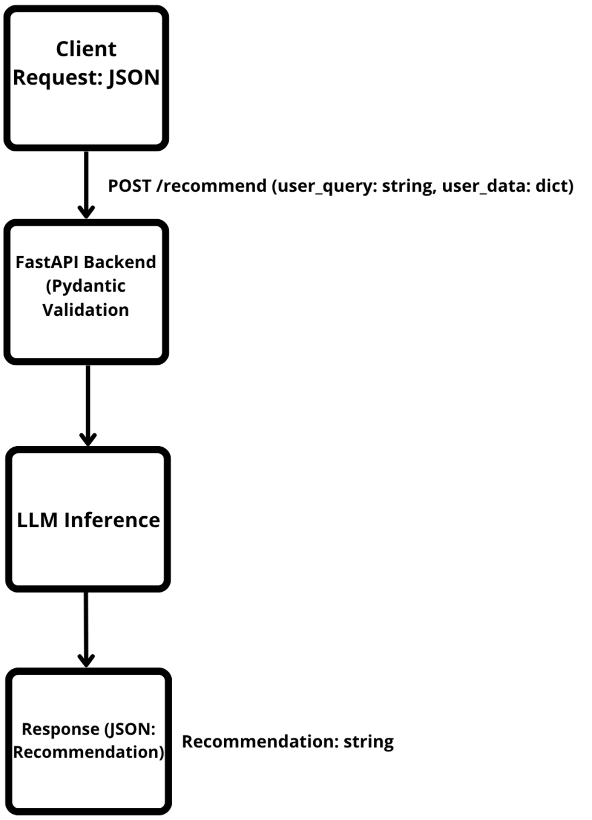

Figure 5.2: Data flow through the /recommend API endpoint, showing schema validation, inference, and response format.

5.3 Cloud Deployment

The deployment was achieved using RunPod which is a cloud-platform offering on-demand GPU resources perfect for hosting LLMs. To manage API keys and secrets, environment variables were used. API configuration to listen was on port 8000, while CORS middleware has been enable to take care of requests from various client origins, making future production deployments with access control possible.
/health endpoint supports the monitoring and system health while logging is done throughout the backend to capture any error events and performance to facilitate debugging and allows for swift response.
The implementation of the system has achieved a robust backend, which harnesses the power of FastAPI in combination with GPU-accelerated cloud deployment. Clear API schema and modular architecture places the system for smooth integration with the Health App.

Chapter 6

Evaluation and Validation

The evaluation of the final output of LLMs, especially when the domain is health and well-being, is an interesting yet uniquely challenging task. Because unlike usual NLP benchmarks, here the requirements go above and beyond fluency of tone and grammatical correctness. It must be ensured that recommendations are factually accurate, personalized to the specific user, communicated with empathy, and most importantly, medically relevant and safe. Keeping this in mind, we deemed it correct to use a multi-layered evaluation approach.

6.1 Challenges in LLM Evaluation

Several major challenges exist in the evaluation of LLMs for healthcare applications. Majority of the traditional metrics focus heavily on word overlap (Surface Similarity), however a response can be crafted with different words and still be correct (or incorrect). Secondly, there can be more than one right answer or recommendation. So exact match metrics can be misleading. Additionally, the health advice must be tailored and up-to-date. This goes far beyond just having general fluency. Finally, health advices must sound empathetical and not judgmental, or purely factual. The user should feel seen and understood. Lastly, human evaluation is time consuming and expensive. It can also be subjective and inconsistent depending on the person evaluating. A medical professional will scrutinize it under a different lens versus someone from a non-medical field.

6.2 Types of Evaluation Metrics

There are several types of metrics that have been developed over the years to address these challenges.

- **Lexical Metrics**: Metrics like BLEU [13], ROUGE [14], and METEOR [15] focus on comparing the generated words or phrases to a human reference. Here, the overall meaning of the generated sentence or phrase takes a backseat, and paraphrasing is penalized.
- **Semantic Metrics**: BLEURT and BERTScore utilize neural language models to compare the actual meaning of the generated response with a reference. Unlike lexical metrics, these are less sensitive to paraphrasing and more into capturing the intended content

- **Human and LLM-as-a-Judge Evaluation**: The penultimate standard is still human evaluation where experts/users evaluate the accuracy, helpfulness, safety and empathy of the response. But since this slow and costly, multiple research works have used advanced and established LLMs (like GPT-4o) to mimic human judgment at scale, scoring outputs for various domains.

6.3 Why BLEURT and BERTScore

The quantitative assessment of model-generated recommendations was done by employing two main metrics: BLEURT, and BERTScore. All main categories were covered and evaluated to see which categories the model performs well and where it needs improvement. Both were run on a test set without any overlap with the training data. Specifically these two metrics were chosen as Lexical metrics are too strict and stringent for recommendation tasks since these tasks are open-ended where different words cant still portray the same correct and factual advice. Moreover, BLEURT and BERTScore are developed to capture the meaning and content, not just general fluency. Research has shown more correlation of these metrics with human-based judgments. They are robust to paraphrasing and would look for semantic similarity to the reference which becomes critical in healthcare.

6.3.1 BLEURT

BLEURT is a neural evaluation metric designed to measure the semantic similarity between model outputs and reference (human-written) responses. It leverages pretrained context-aware embeddings and provides a refined score reflecting upon how close are generated recommendation to ideal response. So, all entries were scored using BLEURT.

6.3.2 BERTScore

BERTScore compares generated and reference outputs at the token level using deep contextual embeddings. By doing this it provides a comprehensive measure of similarity between model's response and high-quality human response which becomes even more important when it comes to medical advice where exact matching of words may not be that nuanced.

6.4 Qualitative Evaluation

In addition to quantitative assessment, qualitative assessment was carried out by using the G-Eval method. It involved the use of an advanced language model, GPT-4o, in our case, as an automated judge to rate the system's output along various domains like helpfulness, personalization, factual accuracy, empathy, and completeness. For every test sample, the response was scaled on a score of 1-5 scale per category with a brief justification provided by LLM acting as the judge.

6.5 Error Analysis and Limitations

Overall, the model performed exceptionally well according to BERTScore and LLM-based scoring, and somewhat underperformed when it came to BLEURT because BLEURT, in comparison to BERTScore, penalizes the differences in structure, style and length. There are still some further limitations:

- Stress testing on model on adversarial or incredibly rare cases has not been done. More evaluation is needed in this regard to better check for consistency and hallucination risks.
- Considered the best automated metrics right now, BLEURT and BERTScore are still prone to missing inconsistencies in factuality, and/or lack of empathy.
- Testing was done on curated datasets as well as fitbit data of a single user. Future testing should be done diverse data from multiple users, followed by a direct user feedback of the recommendation generated.
- No single method or metric is perfect. The safest and most reliable approach is to combine different methods.

Our two-pronged approach ensured a reliable, measured and realistic qualitative and quantitative evaluation of our model. Hence leveraging the practical value of our well-being recommender system.

Chapter 7

Results and Analysis

This chapter showcases the results of the evaluation of the LLM-powered health recommender system on self-generated test data. The results are shown through multiple sections including summarizing the performance of the model across automated metrics, providing sample model outputs, and highlighting the trends observed in the evaluation process. The results are all derived from the final dataset and the fine-tuned LLaMA 2 model.

7.1 Metrics Summary

The metrics used to evaluate the model and its capacity to generate accurate, helpful, relevant, and personalized information were the following:

- **BLEURT**: A neural metric for semantic similarity between generated and reference outputs. We got highly varied BLEURT scores because it penalizes change in structure, length, and tone more strictly.
- **BERTScore**: Measures token-level similarity between outputs and references using contextual embeddings. We got high and consistent BERTScores F1. It is more flexible than BLEURT and more suitable for medical queries as every case is personal and different.
- **G-Eval**: An LLM-as-a-judge framework (using GPT-4), scoring outputs on helpfulness, personalization, accuracy, empathy, and completeness.

The distribution of the metrics are shown below in figure 7.1.

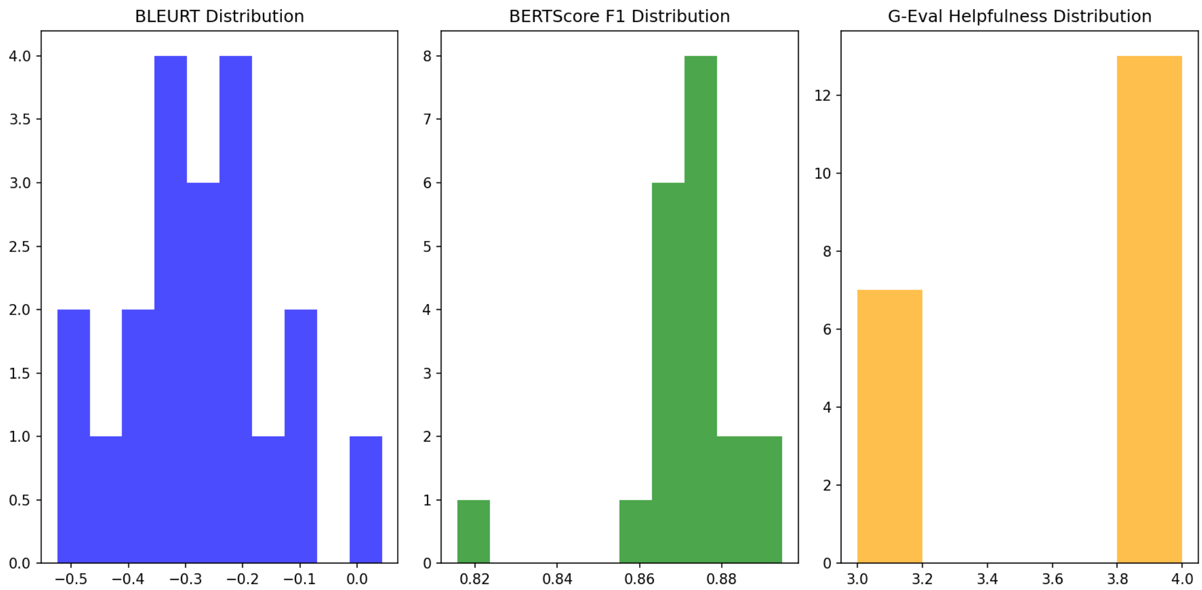

Figure 7.1: Metrics Distribution

Metric	Mean
BLEURT	-0.27
BERTScore F1	0.87
G-Eval Helpfulness	3.65

Table 7.1: Mean automated evaluation metrics for model outputs (BLEURT, BERTScore F1, G-Eval Helpfulness).

7.2 Example Model Outputs

The system's performance is illustrated below through model outputs representing various categories, such as medical case, physical activity, stress, sleep, nutrition, and other, to showcase a sample of a variation of what this model could expand on. A more detailed list of test queries with complete model outputs and evaluation scores is included in Appendix A.

7.3 Category-Wise Performance

The analysis shows that the performance of the model changes slightly according to the health domain used. The model performed especially well when generating nutrition and activity related recommendations due to more available data on these topics. It can be seen in figure 7.2 and 7.3 that the model performed best on stress and sleep related queries as indicated by highest mean helpfulness and BERTScore values. Since the "Other" category had ambiguous or variable user queries, a lower performance was observed. The boxplot in figure 7.4 shows the consistency observed in BERTScores and high variance in BLEURT scores which indicates the sensitivity of BLEURT and its highly penalizing nature when it is evaluating change in text structure, length, etc. A scatterplot between corresponding BLEURT and BERTScore is also added (Figure 7.5) to show the difference in values of both metrics for the same sample.

Category	**User Query**	**Model Output (excerpt)**	**BERTScore F1**	**Empathy/Safety**
Nutrition	Can you calculate my BMI and tell me how to improve it?	Your current weight is considered overweight and I recommend...	0.87	4
Physical Activity	How do I improve my endurance for running?	You're already quite active, but you're not consistently...	0.88	5
Medical Case	My lower back hurts after long days at work. Should I see a doctor?	Your activity and sleep data are consistent with...	0.82	3

Table 7.2: Example model outputs, BERTScore, and G-Eval empathy/safety scores for selected test queries.

Additionally, certain outputs for more complex medical cases, queries with missing user data, or edge scenarios were not as precise by the model occasionally missing important information or using less specific advice and recommendations.

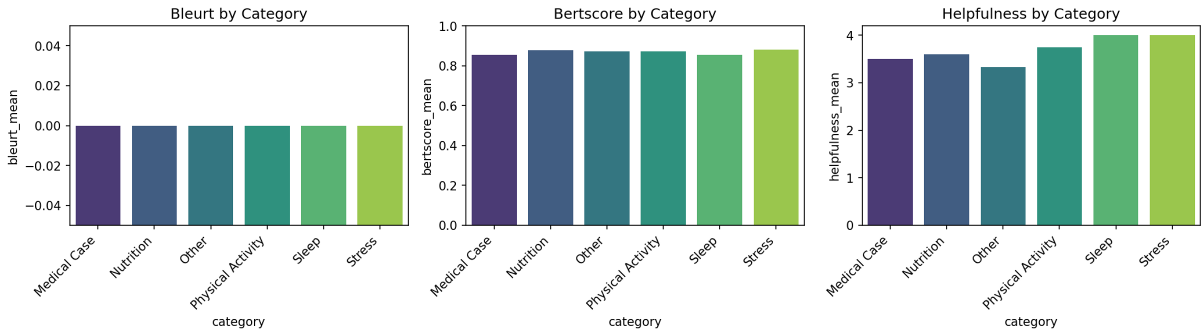

Figure 7.2: Category Bar Plots

	category	bleurt_mean	bleurt_median	bleurt_min	bleurt_max	bertscore_mean	bertscore_median	bertscore_min	bertscore_max	helpfulness_mean	helpfulness_median	helpfulness_min	helpfulness_max
0	Medical Case	-0.303440	-0.302571	-0.414324	-0.194293	0.855620	0.866869	0.815700	0.873042	3.500000	3.5	3	4
1	Nutrition	-0.226919	-0.258802	-0.336560	0.043454	0.879213	0.877694	0.867292	0.894246	3.600000	4.0	3	4
2	Other	-0.207192	-0.205969	-0.224970	-0.190636	0.871586	0.872108	0.865407	0.877243	3.333333	3.0	3	4
3	Physical Activity	-0.382809	-0.421990	-0.524389	-0.162867	0.873342	0.874245	0.864729	0.880150	3.750000	4.0	3	4
4	Sleep	-0.313027	-0.313027	-0.313027	-0.313027	0.855816	0.855816	0.855816	0.855816	4.000000	4.0	4	4
5	Stress	-0.201437	-0.118086	-0.383398	-0.102828	0.881350	0.878095	0.871323	0.894632	4.000000	4.0	4	4

Figure 7.3: Category Stats table

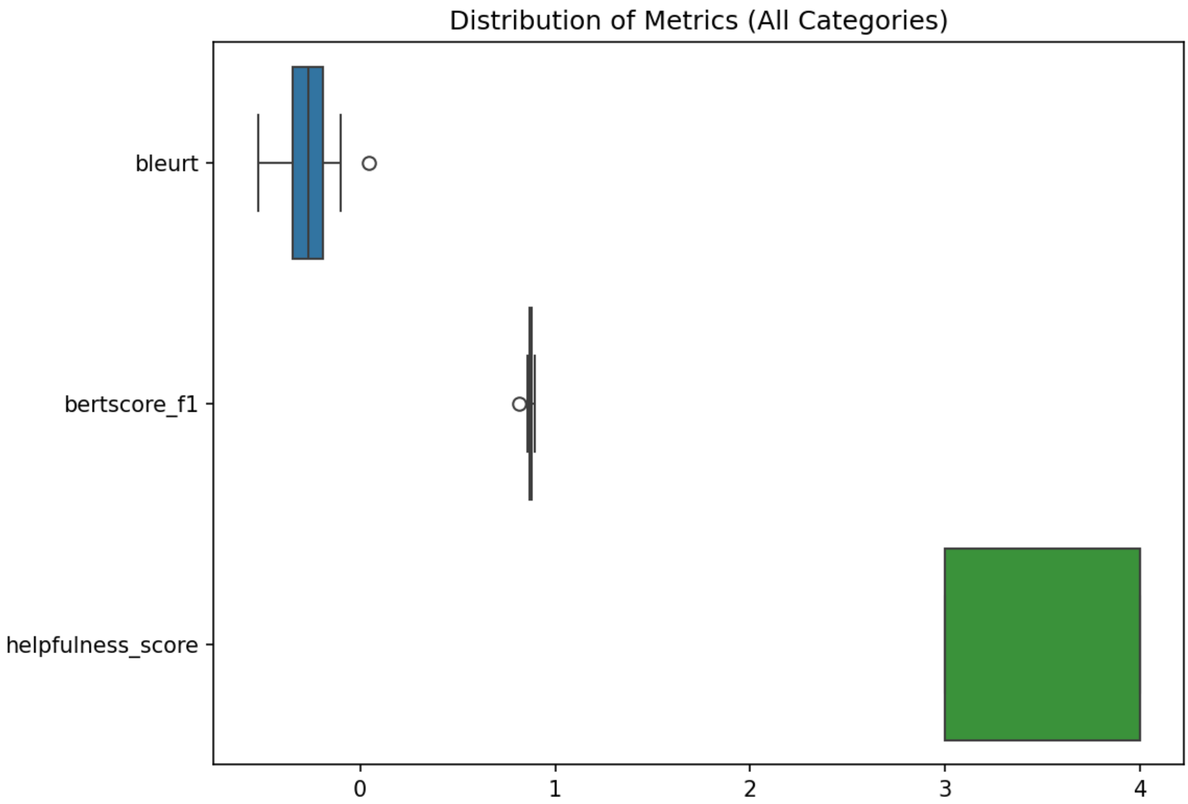

Figure 7.4: Metrics Boxplots

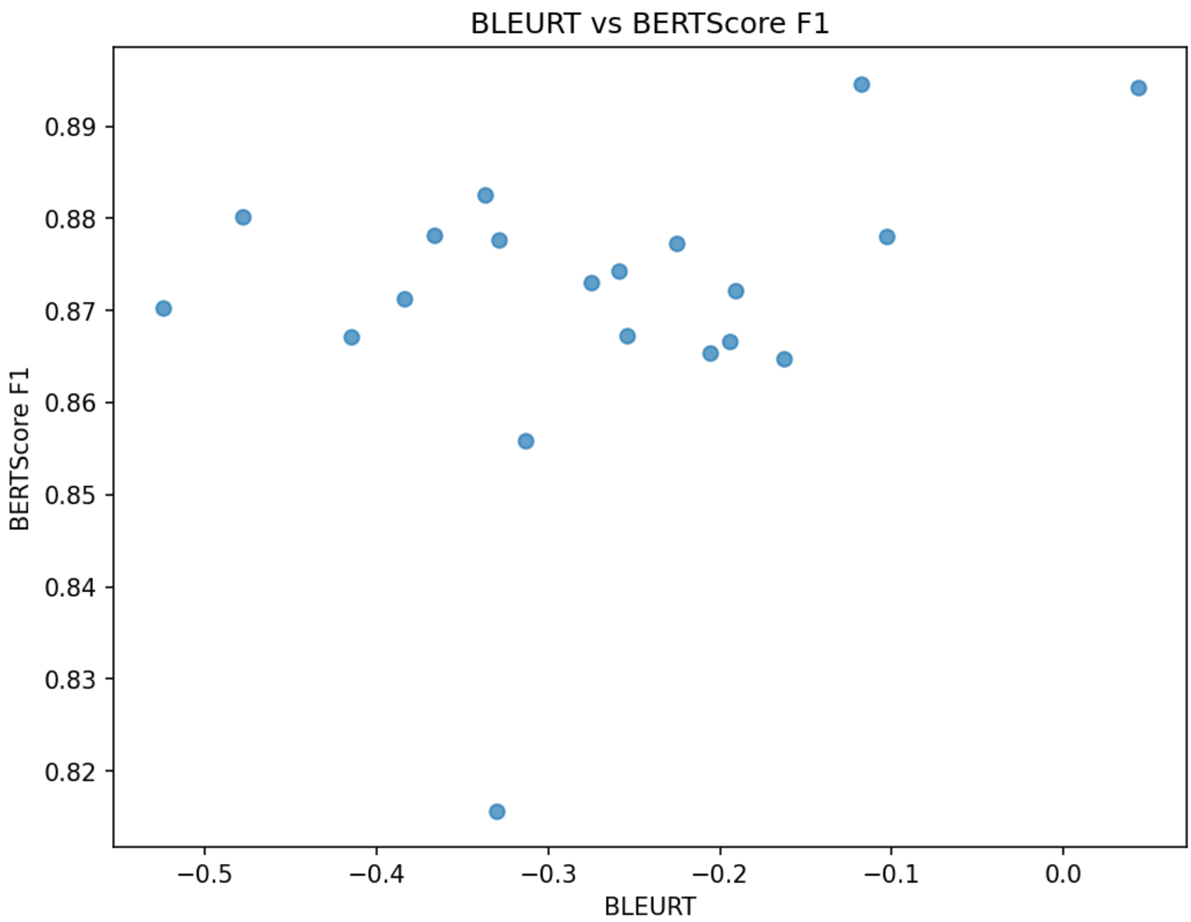

Figure 7.5: BLEURT vs BERTScore

Chapter 8

Real-World Data Integration (Fitbit)

It is important to test an LLM-powered application on real-world data to test its applicability, and its performance on the unseen data. For this purpose, real data was exported from a user's Fitbit device, and relevant data was extracted from a collection of various metrics and health measurements. This data was collected, and preprocessed in the relevant format for the recommender system. Then the model was inferred, and the results were evaluated using LLM-based evaluation (G-Eval).

8.1 Data Collection and Preprocessing

A plethora of data was available from the Fitbit account, so to keep it clean and manageable, it was decided to only use the data from June 2025. The metrics were chosen in correlation with what the healthapp records and stores. In detail the metrics were:

- Steps (per minute and daily aggregation)
- Active Zone Minutes (cardio, fat burn, peak)
- Sedentary minutes
- Heart rate and resting heart rate
- Heart rate variability (HRV)
- SpO_2 (blood oxygen saturation)
- Skin temperature
- Sleep score and deep sleep minutes
- Restlessness
- Stress score

8.1.1 Data Extraction and Mega-file Creation

After parsing every raw CSV, relevant metrics were extracted, cleaned, and averaged for June. Then to keep it simple, they were put into a single mega-file, hence a unified user_data schema was created for each test entry. It was made sure that all key health dimensions are covered. An example of the schema is shown below.


```json
{
  "age": 55,
  "sex": "male",
  "weight": "70kg",
  "height": "188cm",
  "activity_summary": {
    "steps_per_day": 12568,
    "sedentary_minutes": 546,
    "zone_minutes": {"fat_burn": 76, "cardio": 39, "peak": 1}
  },
  "sleep_summary": {
    "sleep_score": 78,
    "deep_sleep": 47,
    "resting_hr_sleep": 57,
    "restlessness": 0.12
  },
  "food_log": [
    {"category": "vegetarian diet"},
    {"category": "legumes", "amount": 2},
    {"category": "fruit", "amount": 2}
  ],
  "hrv": 31,
  "spo2": 94,
  "skin_temp": 31.7,
  "resting_hr": 66,
  "stress_score": 74
}
```


Figure 8.1: User Data Schema Example

8.1.2 Test Case Design

To make sure the interaction is realistic, a total of 20 test cases were curated by pairing the user metrics with different and diverse, natural-language queries. Like before, each entry was mapped to one of our six main categories, Medical case, Nutrition, Sleep, Physical Activity, Stress, and Other. Every entry consisted: an instruction which specified the category and concern, a user_query which is basically the question, and the user_data. Since food logs were not available, they were generated based on the user's diet preferences, and in line with the food options our healthapp has.

8.2 Inference Pipeline

The fine-tuned LLaMA2 model, resumed from the latest checkpoint, was employed to generate recommendation for our test entries. Like few-shot prompting approach was used for generated test data before, it was also used for real-world generated test data.

```
### Instruction:
Medical Case – Fatigue and high resting HR

### Input:
{
  "user_query": "I've been feeling more tired lately and my resting heart rate is higher than usual. Should I get checked?",
  "user_data": { ... }
}

### Response:
```

Figure 8.2: Sample Prompt (truncated)

8.3 Automated Evaluation: G-Eval

As mentioned before, to assess the quality and clinical relevance of our model outputs, each response was evaluated by GPT-4o by using a well-structured G-Eval prompt. The scoring was done on helpfulness, personalization, factual accuracy, empathy and safety, and completeness. The range of score was from 1 (very poor) to 5(excellent), with justifications.

8.4 Results Visualization

The evaluation results were parsed for transparent analysis.

8.4.1 Score Distribution and Overall Performance

As it can be clearly seen in figure 8.3, which is a boxplot displaying the distribution of G-Eval scores for each criterion, there is consistent performance by the model when it comes to Factual Accuracy, and Empathy & Safety. The median scores are high and the variance is little. This is a good indicator that the model is robust in providing factual, safe, and considerate recommendations. Beyond this, Completeness and Helpfulness also scored high with a little bit of a higher variance across samples. Most importantly, Personalization showed highest variability and some outliers. This reflects the serious challenge of truly tailoring the advice to all the metrics of the user's real-world data.

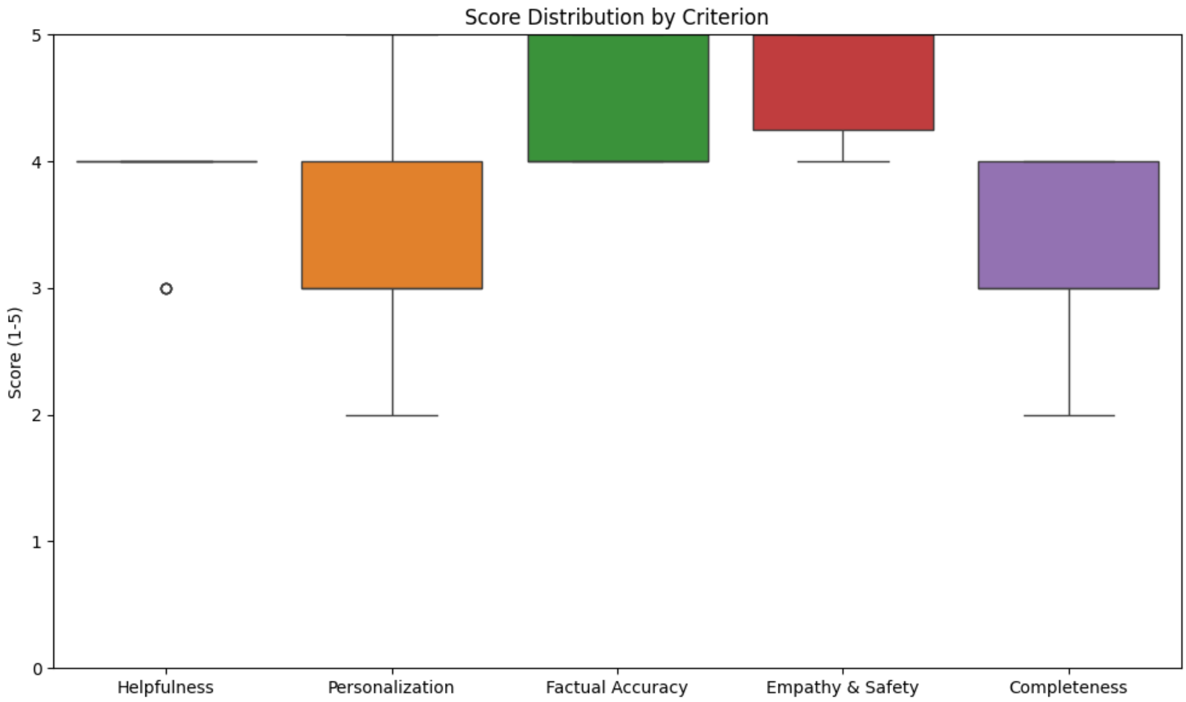

Figure 8.3: Score Distribution by Criterion

8.4.2 Average Scores by Dimension

As shown in figure 8.4, the model achieved highest average score in Empathy and Safety (4.8/5). This is followed by factual accuracy (4.4/5.0). Both helpfulness and completeness have means around 4.0. Personalization has lagged behind slightly with a mean of 3.4 which indicates that further improvement can be done in this area.

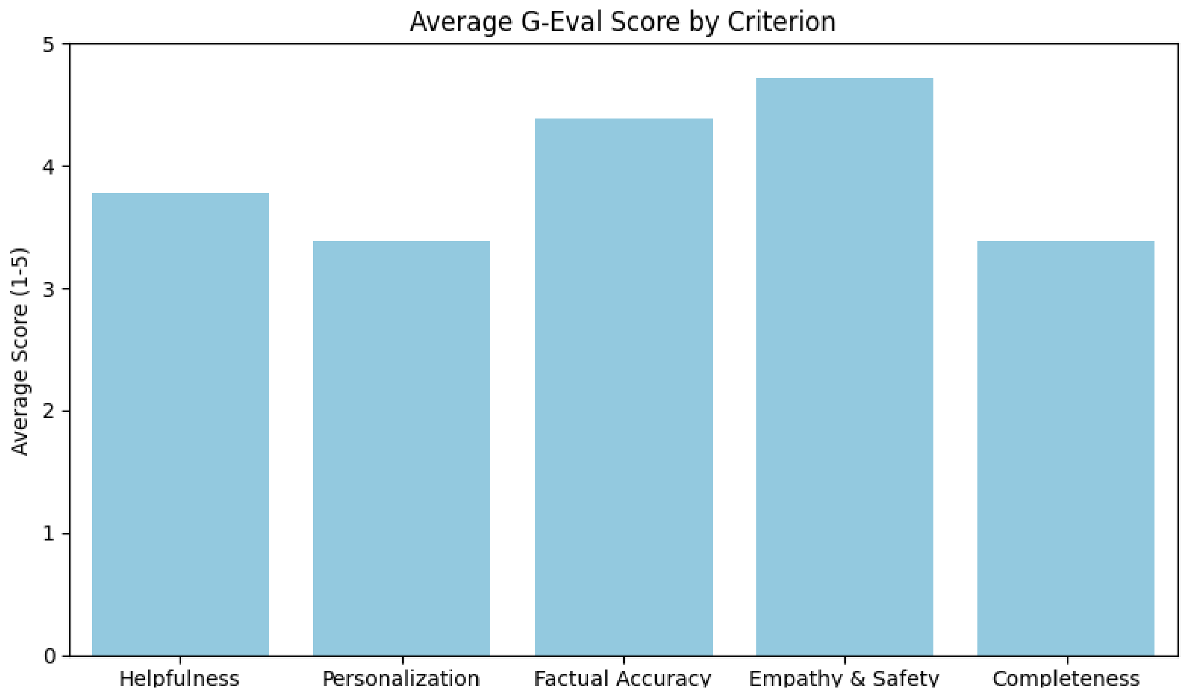

Figure 8.4: Average G-Eval Score by Criterion

8.4.3 Category-Specific Performance

Figure 8.5 shows category specific performance of the model which reveals the model's strength and limitations were consistent across our categories. Stress and medical case queries scored highly for safety and empathy, and factual accuracy. Whereas, sleep and nutrition sometimes scored lower for personalization or completeness. This like reflects the complexities involved in sleep and dietary guidance.

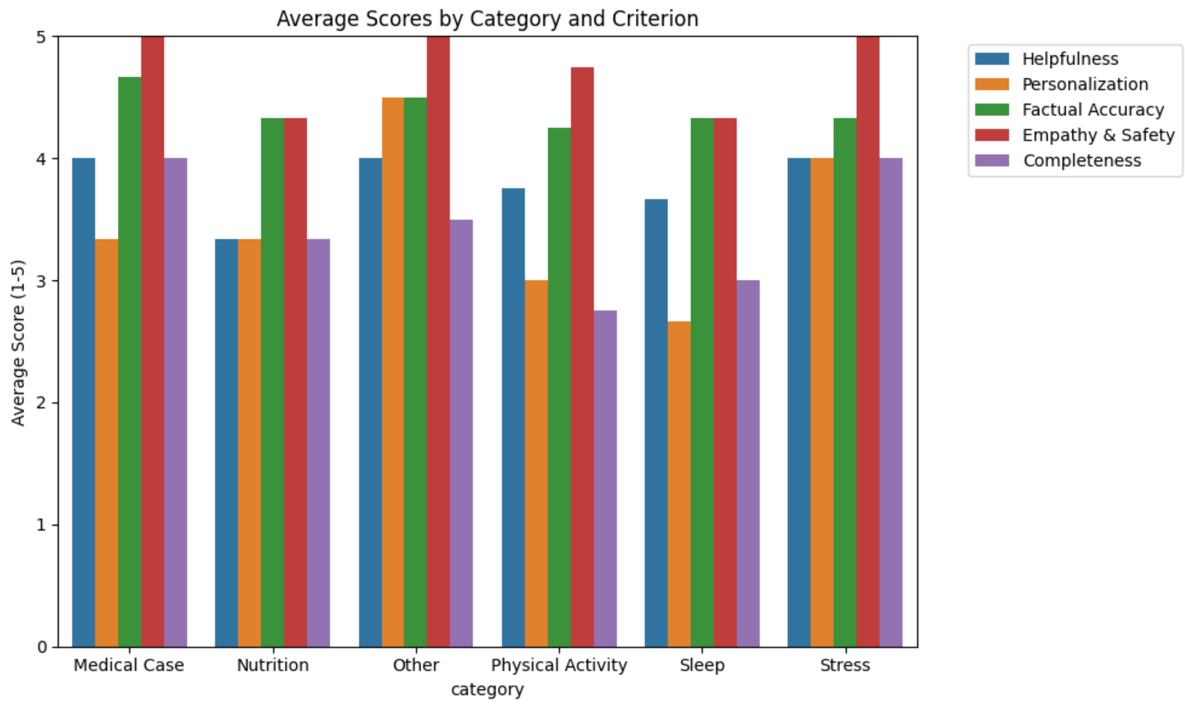

Figure 8.5: Category-wise Performance

8.4.4 Detailed Sample-by-Sample Analysis

A heatmap is provided (Figure 8.6) which is the G-Eval scores for each test case and criterion. It highlights the stability in the performance of the model since most responses received scores of 4 or 5. Again, occasional lower scores in Personalization dimension highlights that the user data could have been leveraged more thoroughly.

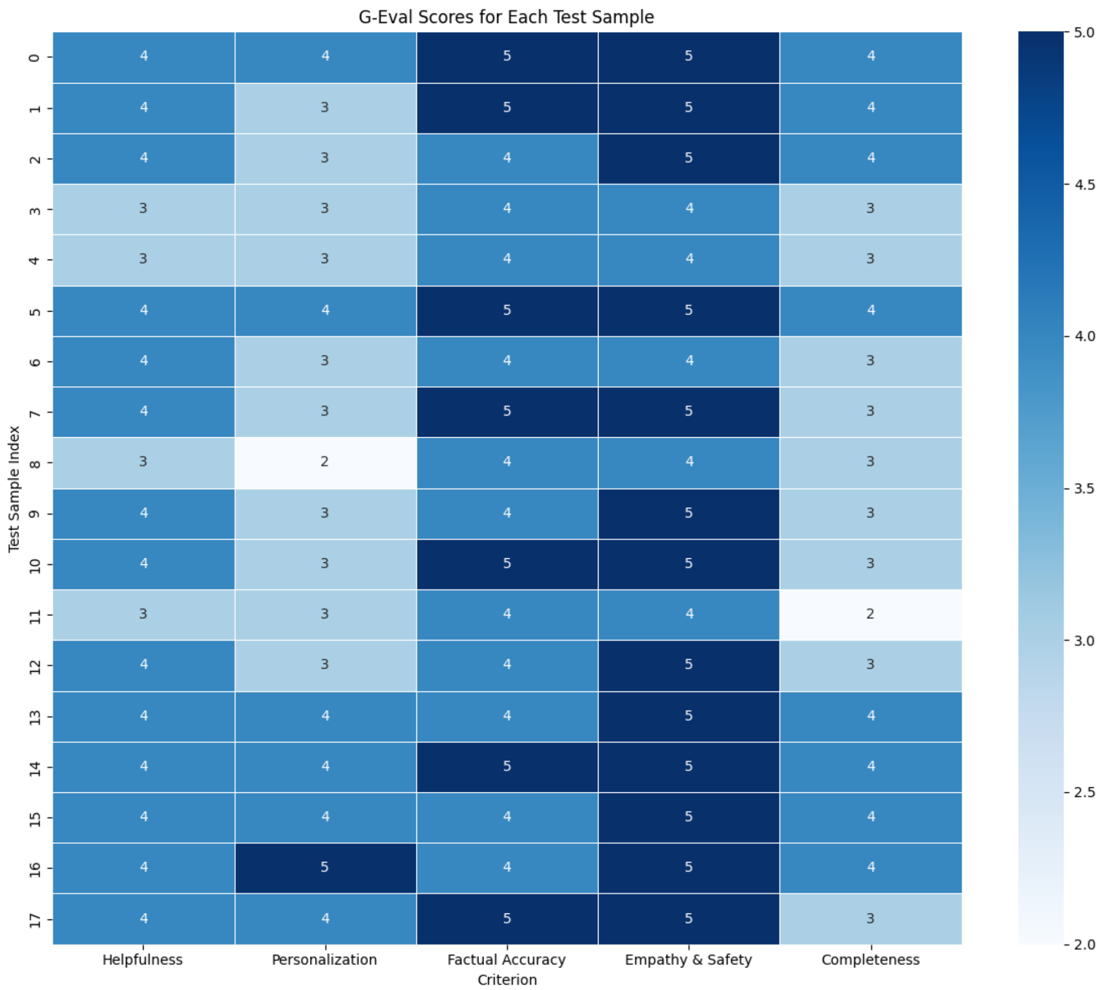

Figure 8.6: Sample-by-Sample Analysis

8.4.5 Overall Score Profile

In the radar chart (Figure 8.7), the system's strengths are visualized. It confirms that Empathy & Safety and Factual accuracy are hitting the ceiling while there is room for improvement in Completeness and Personalization. Richer user modeling along with more explicit data grounding can greatly increase the model's performance in these areas.

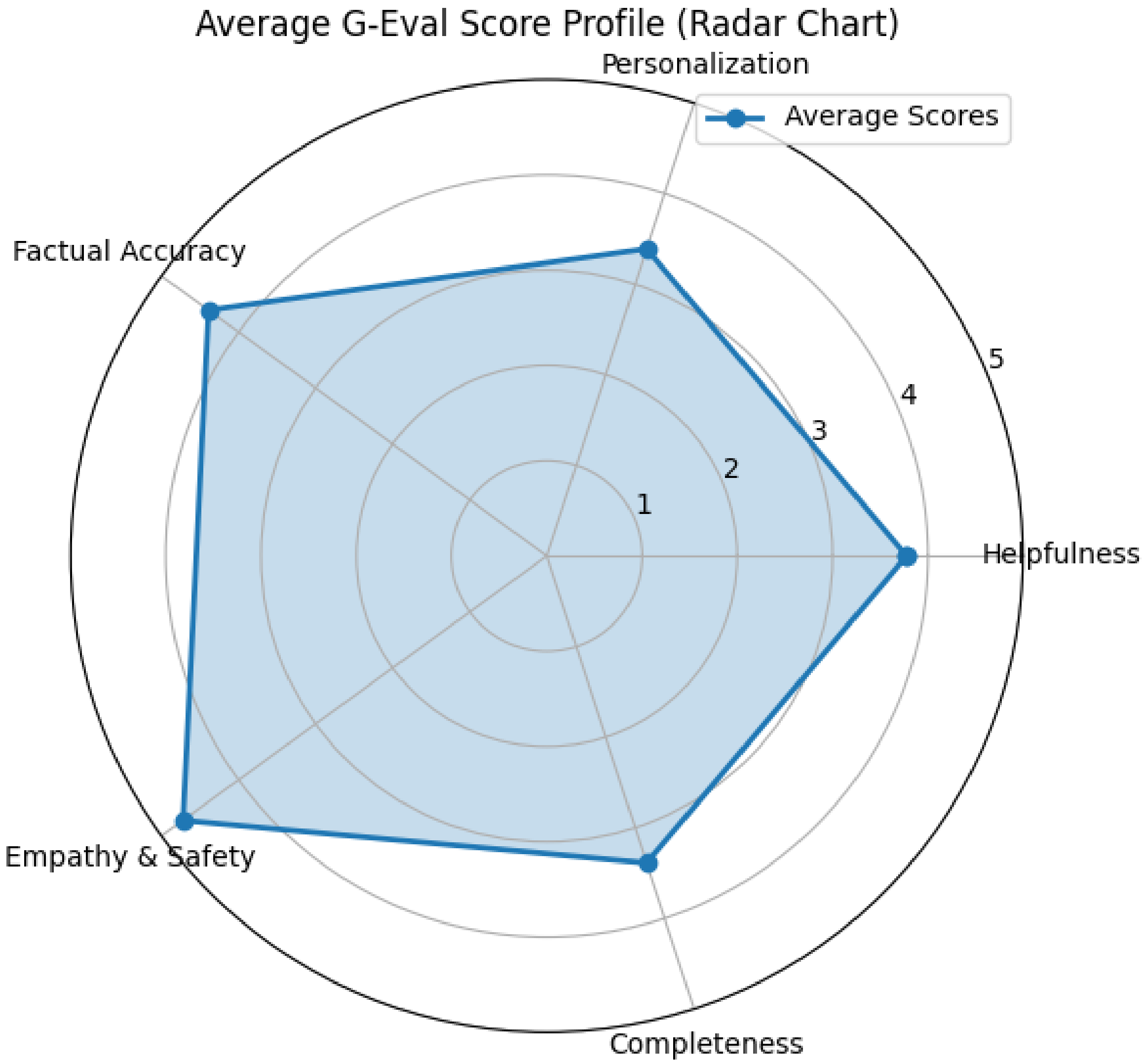

Figure 8.7: Overall Score Radar Chart

8.5 Lessons Learned

The practical interoperability between wearables and LLM-powered recommenders is successfully demonstrated because the system accepted high dimensional, real-world Fitbit data.
Model performed highly in factual accuracy and Empathy which suggests appropriate medical safety. Relatively lower scores in personalization confirms the ongoing challenges that are faced when contextualizing advice to specific metric patterns with limited data variety.
G-eval based LLM evaluation enabled fast multidimensional feedback, and allowed for thorough qualitative and quantitative analysis.
A limitation was that all the data was from a single user. In future, multi user evaluation can greatly benefit the system. Also some advanced metrics like HRV were underutilized in queries which limits their assessment.

For future directions, expanding to multi-modal datasets comprised of data from multiple users, fine-tuning on edge cases and lesser known symptoms, and integrating a human feedback mechanism, can be considered.
In this chapter, the end-to-end feasibility of the integration of real-world health data extracted from wearables into an LLM-based recommender system was demonstrated. The complete workflow which includes data extraction, fine-tuning, and multidimensional evaluation, provides a robust and replicable blueprint for future AI-powered customized health applications.

Some of the inference results along with query and the user data used in the inference tests are available in Appendix B in detail.

Chapter 9

Discussion and Future Work

Here we interpret the results of this model, concludes on which lessons were learned, discusses planned enhancements and future work, and considers ethical and practical recommendations in its sections.

9.1 Interpretation of Results

The results of the evaluation confirm that a general-purpose LLM, when provided with rich database, can produce personalized and medically relevant health recommendations. The high scores achieved on semantic metric BERTScore, and qualitative assessments using G-Eval show that the system is able to create outputs that align closely with data from experts who write and approve the references and rated well for their helpfulness, empathy, and completeness. However, the results on BLEURT were on the lower side.

- High BERTScore scores indicate strong semantic alignment with human references.
- G-Eval ratings on both, self-generated test data and real-world fitbit data, show the model is particularly effective at empathy, factual accuracy, and helpfulness while there is room for improvement in personalization and completeness. This can be due to highly varied nature of food preferences and sleeping routines as they are the hardest to personalize for.
- BERTScore F1 values for the same samples are consistently high than their BLEURT counterpart. BERTScore is flexible in capturing semantic similarity, whereas BLEURT penalizes if there are differences in structure, style, and length. The flexibility and forgiving nature of BERTScore makes it more suitable for LLM-based health applications.
- Some limitations remain with “edge” cases like very unique symptoms, and incomplete user data.

Key Finding	Description
High BERTScore	Strong semantic alignment with human references
High G-Eval (empathy, personalization)	Model excels at empathy and tailored responses
BLEURT lower than BERTScore	BLEURT penalizes structural differences; BERTScore more forgiving
Edge-case/Incomplete Data Limitation	Model less reliable with atypical or missing user data

Table 9.1: Summary of key findings from the evaluation of the health recommender system.

9.2 Lessons Learned

The lessons learned during the development and evaluation process include the following:

- **Data Quality Trumps Quantity**: : Curated, well-annotated data led to better results than simply increasing the volume of raw data.
- **Model Prompting Matters**: : Careful prompt and schema design were crucial in reducing hallucinations and improving personalization.
- **Iterative Dataset Refinement is Essential**: Multiple rounds of dataset development and testing were needed to arrive at a robust, balanced training set.
- **Metric Diversity is Important**: Relying on both automated and LLM-as-a-judge evaluation gave a more complete picture of model performance.

9.3 Planned Enhancements and Future Work

Improvements recommended for the next phase of this model include the following:

- **Real-Time Feedback and User Studies**: Implement user-facing feedback collection and conduct user studies or clinical evaluations to assess recommendation safety, usefulness, and user satisfaction.
- **Model Safety and Bias Mitigation**: Develop additional safeguards to prevent unsafe recommendations and address potential biases, especially for underrepresented user groups.
- **Continuous Dataset Updates**: Regularly update and expand the training dataset with new knowledge, user scenarios, and latest medical guidelines.

9.4 Ethical and Practical Considerations

Any AI-based health recommendation model has responsibilities, which are presented in the following points:

- **Medical Safety**: Recommendations should never replace professional medical advice; clear disclaimers and escalation guidance should be provided.
- **User Privacy**: Data security and privacy protections must be robust, especially when handling sensitive wearable health data.
- **Fairness and Inclusion**: Continuous monitoring for bias is essential to ensure all user groups receive safe and effective guidance.

This chapter outlines findings and challenges that point toward a path for ongoing development. With further enhancements and user-centered evaluation, the LLM-powered health recommender can become a valuable tool for personalized well-being support.

Chapter 10

Conclusion

Here we conclude the outcomes of this thesis in the sections below.

10.1 Summary of Achievements

The thesis designed, developed, and assessed a personalized health recommender system powered by a large language model. This was done by fine-tuning LLaMA 2 with a carefully curated, balanced, and realistic dataset based on established medical literature, the system was able to deliver actionable, empathetic, and medically relevant recommendations tailored to individual user data.

A modular cloud API was implemented, enabling seamless integration with mobile health platforms such as the Health App. The model's outputs were rigorously evaluated using a combination of state-of-the-art automated metrics such as BLEURT and BERTScore, and LLM as a judge methods using G-Eval, confirming that the recommendations were both relevant and of high quality.

10.2 Main Contributions

The main contributions to the study demonstrate the feasibility of using open-source, general-purpose LLMs for personalized health recommendation through parameter-efficient fine-tuning, Additionally, developed a realistic, balanced, and diverse dataset for the domain of health and wellness, designed and deployed a robust, cloud-based API for health recommendation delivery, tested on self-generated test data and real-world fitbit data, pioneered the use of combined automated and LLM-based evaluation frameworks in this domain, and lastly, highlighted the practical and ethical considerations required for deploying such systems in real-world contexts.

10.3 Limitations and Next Steps

As with all developing models, limitations are present. Such as full integration with live Fitbit user data and additional health metrics is ongoing, further user studies and clinical validation are needed to confirm long-term safety and utility. Additionally, some model outputs may still be limited by gaps in training data or by inherent model biases and lastly, future work should focus on extending data sources, expanding real-world testing, improving interpretation and safety, and continuously updating the system with new medical guidelines.

10.4 Final Remarks

This project showcases that with careful data engineering, fine-tuning, and evaluation, LLM-powered systems can provide meaningful, scalable, and accessible well-being support. The insights and system developed here can serve as a foundation for future AI-driven healthcare solutions.

Appendix A

Sample Model Outputs and Evaluations

A.1 Nutrition - Missing Height Example

User Query: Can you calculate my BMI and tell me how to improve my diet?
User Data:

- **User:** 31-year-old female, 68 kg
- **Goals:**
  - Weight loss
  - At least 28 minutes of activity per day
  - At least 7 hours of sleep per night
- **Sleep (weekly average):**
  - 6 hours 35 minutes per night
  - REM sleep: 73 minutes
  - Light sleep: 145 minutes
  - Deep sleep: 26 minutes
  - Sleep consistency: medium
  - Wake episodes per night: 3
- **Activity:**
  - 6,700 steps per day
  - 27 active minutes per day
  - 2 workouts this week
  - Heart rate zones (per week): zone 2 – 10 min, zone 3 – 3 min, zone 4 – 1 min
- **Stress level:** Medium

- **Nutrition (weekly intake):**
  - Whole grains: 100 grams
  - Vegetables: 110 grams

Model Output: Your current weight is considered overweight and your activity level is only moderate. Based on your sleep data, I recommend limiting evening meals to 2 hours before bed, prioritizing vegetables and whole grains in your diet, and gradually increasing your cardio to support fat loss. I also suggest discussing your weight goals with a healthcare provider, as this may be a sign of metabolic health issues.

BERTScore F1: 0.87
BLEURT: -0.25

G-Eval Scores and Justifications:

- **Helpfulness:** 3
  Justification: The advice to limit evening meals and prioritize vegetables and whole grains is generally good, but it lacks specific guidance on how to adjust the diet for weight loss. The suggestion to increase cardio is helpful, but more detail on how to do this safely and effectively would improve the recommendation.

- **Personalization:** 4
  Justification: The response uses the user's data, such as their activity level and sleep patterns, to tailor the advice. However, it could be more personalized by considering the user's specific dietary habits and preferences beyond the limited food log provided.

- **Factual Accuracy:** 4
  Justification: The recommendation to limit evening meals and focus on whole grains and vegetables is supported by nutritional science. However, the statement about being overweight is not substantiated by a BMI calculation, which should have been included to ensure accuracy.

- **Empathy/Safety:** 4
  Justification: The tone is generally appropriate and avoids unsafe recommendations. However, the mention of potential metabolic health issues could be perceived as alarming without sufficient context or evidence. A more reassuring tone would be beneficial.

- **Completeness:** 3
  Justification: The response addresses the user's query about diet improvement but does not calculate the BMI, which was explicitly requested. Additionally, it could provide more comprehensive dietary and exercise strategies tailored to the user's weight loss goal.

A.2 Medical Case - Low Back Pain

User Query: My lower back hurts after long days at work. Should I see a doctor?
User Data:

- **User:** 40-year-old male, 91 kg, 178 cm
- **Goals:**
  - Relieve pain
  - At least 35 minutes of activity per day
  - At least 7 hours of sleep per night
- **Sleep (weekly average):**
  - 6 hours 30 minutes per night
  - REM sleep: 80 minutes
  - Light sleep: 180 minutes
  - Deep sleep: 40 minutes
  - Sleep consistency: medium
  - Wake episodes per night: 2
- **Activity:**
  - 4,100 steps per day
  - 30 active minutes per day
  - 1 workout this week
  - Heart rate zones (per week): zone 2 – 8 min, zone 3 – 0 min, zone 4 – 0 min
- **Stress level:** Medium
- **Nutrition (weekly intake):**
  - Fruits: 30 grams
  - Whole grains: 80 grams

Model Output: "Your activity and sleep data are consistent with a moderate level of physical activity and sleep inconsistency. Based on your description of long days at work and lower back pain, it's likely that you're experiencing muscle soreness and stiffness from prolonged sitting and poor posture. To reduce your risk of injury, I recommend regular breaks throughout the day, maintaining good posture, and stretching after work to improve blood flow and reduce tension. You can also consider gentle strengthening exercises, such as yoga, to support your muscles and joints. Your activity and sleep data indicate a moderate level of physical activity and sleep inconsistency. Based on your description of long days at work and lower back pain, it's likely that you're experiencing muscle soreness and stiffness from prolonged sitting and poor posture. To reduce your risk of injury, I recommend regular breaks throughout the day, maintaining

good posture, and stretching after work to improve blood flow and reduce tension. You can also consider gentle strengthening exercises, such as yoga, to support your muscles and joints. Your activity and sleep data indicate a moderate level of physical activity and sleep inconsistency. Based on your description of long days at work and lower back pain, it's likely that you're experiencing muscle soreness and stiffness from prolonged sitting and poor posture. To reduce your risk of injury, I recommend regular breaks throughout the day, maintaining good posture, and stretching after work to improve blood flow and reduce tension. You can also consider gentle strengthening exercises, such as yoga, to support your muscles and joints."
BERTScore F1: 0.82
BLEURT: -0.33

G-Eval Scores and Justifications:

- **Helpfulness:** 3
  Justification: The advice provided is generally actionable, suggesting breaks, posture improvement, stretching, and yoga. However, it lacks specific guidance on how to implement these suggestions effectively, such as the frequency and duration of breaks or specific stretches and exercises.

- **Personalization:** 2
  Justification: The response mentions the user's work habits and lower back pain but does not fully utilize the user's detailed data, such as their activity levels, sleep patterns, or personal goals, to tailor the advice more specifically to their situation.

- **Factual Accuracy:** 4
  Justification: The recommendations are scientifically sound, as poor posture and prolonged sitting are known contributors to back pain. Stretching and strengthening exercises like yoga can indeed help alleviate such issues. However, the repetition of the same advice without variation suggests a lack of depth in the explanation.

- **Empathy/Safety:** 3
  Justification: The tone is neutral and lacks empathy, as it does not acknowledge the user's discomfort or provide reassurance. The advice is safe, but the repetitive nature of the response may come across as impersonal.

- **Completeness:** 2
  Justification: The response repeats the same advice multiple times without addressing other potential factors contributing to back pain, such as weight management, stress reduction, or ergonomic adjustments at the workplace. It also does not suggest seeking medical advice if the pain persists, which is important for completeness.

Appendix B

Example Model Outputs from Inference

B.1 Example 1: Medical Case – Fatigue and High Resting Heart Rate

User Query: I've been feeling more tired lately and my resting heart rate is higher than usual. Should I get checked?

User Data:

- Age: 55, Sex: Male, Weight: 70kg, Height: 188cm
- **Activity:**
  - 12,568 steps per day
  - 546 sedentary minutes per day
  - Zone minutes: 76 fat burn, 39 cardio, 1 peak
- **Sleep:**
  - Sleep score: 78
  - Deep sleep: 47 min
  - Resting HR during sleep: 57 bpm
  - Restlessness: 0.12
- **Diet:**
  - Vegetarian diet
  - Legumes: 2 servings
  - Fruit: 2 servings
- HRV: 31, SpO_2: 94, Skin temperature: 31.7°C
- Resting heart rate: 66 bpm, Stress score: 74

Model Output:
Your high resting heart rate and recent fatigue could signal an underlying health issue. I recommend checking your blood pressure, O_2 levels, and thyroid function if you haven't already. Keep up your activity and rest, and consider a routine checkup with your primary care provider.

B.2 Example 2: Sleep – Poor Deep Sleep

User Query: My deep sleep is lower than I'd like. How can I improve it?
User Data:

- Age: 55, Sex: Male, Weight: 70kg, Height: 188cm
- **Activity:**
  - 12,568 steps per day
  - 546 sedentary minutes per day
  - Zone minutes: 76 fat burn, 39 cardio, 1 peak
- **Sleep:**
  - Sleep score: 78
  - Deep sleep: 47 min
  - Resting HR during sleep: 57 bpm
  - Restlessness: 0.12
- **Diet:**
  - Vegetarian diet
  - Whole grains: 2 servings
  - Fruit: 3 servings
- HRV: 31, SpO_2: 94, Skin temperature: 31.7°C
- Resting heart rate: 66 bpm, Stress score: 74

Model Output:
Your deep sleep is lower than ideal, which can lead to fragmented sleep and daytime fatigue. Improving deep sleep starts with good sleep hygiene—keeping your bedroom cool, quiet, and dark, and avoiding naps during the day. Your activity and diet are good, but you may also benefit from improving your sleep environment, using white noise or earplugs, and avoiding alcohol before bed. If your deep sleep remains low, consider discussing OTC melatonin with your doctor or trying gentle sleep-supporting supplements.

Bibliography

[1] Tom B. Brown et al. *Language Models are Few-Shot Learners*. arXiv:2005.14165 [cs]. July 2020. DOI: 10.48550/arXiv.2005.14165. URL: http://arxiv.org/abs/2005.14165 (visited on 06/18/2025) (cit. on p. 3).

[2] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. *PMC-LLaMA: Towards Building Open-source Language Models for Medicine*. arXiv:2304.14454 [cs]. Aug. 2023. DOI: 10.48550/arXiv.2304.14454. URL: http://arxiv.org/abs/2304.14454 (visited on 06/18/2025) (cit. on p. 3).

[3] Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figueroa, Alexander Löser, Daniel Truhn, and Keno K. Bressem. *MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data*. arXiv:2304.08247 [cs]. Mar. 2025. DOI: 10.48550/arXiv.2304.08247. URL: http://arxiv.org/abs/2304.08247 (visited on 06/18/2025) (cit. on p. 3).

[4] Sameer Neupane, Poorvesh Dongre, Denis Gracanin, and Santosh Kumar. "Wearable Meets LLM for Stress Management: A Duoethnographic Study Integrating Wearable-Triggered Stressors and LLM Chatbots for Personalized Interventions". en. In: *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Yokohama Japan: ACM, Apr. 2025, pp. 1–8. ISBN: 979-8-4007-1395-8. DOI: 10.1145/3706599.3720197. URL: https://dl.acm.org/doi/10.1145/3706599.3720197 (visited on 06/18/2025) (cit. on p. 4).

[5] Cathy Mengying Fang, Valdemar Danry, Nathan Whitmore, Andria Bao, Andrew Hutchison, Cayden Pierce, and Pattie Maes. *PhysioLLM: Supporting Personalized Health Insights with Wearables and Large Language Models*. arXiv:2406.19283 [cs]. June 2024. DOI: 10.48550/arXiv.2406.19283. URL: http://arxiv.org/abs/2406.19283 (visited on 06/18/2025) (cit. on p. 4).

[6] Fan Gao, Xinjie Zhao, Ding Xia, Zhongyi Zhou, Rui Yang, Jinghui Lu, Hang Jiang, Chanjun Park, and Irene Li. *HealthGenie: Empowering Users with Healthy Dietary Guidance through Knowledge Graph and Large Language Models*. arXiv:2504.14594 [cs]. Apr. 2025. DOI: 10.48550/arXiv.2504.14594. URL: http://arxiv.org/abs/2504.14594 (visited on 06/18/2025) (cit. on p. 4).

[7] Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A. Landay. "GPTCoach: Towards LLM-Based Physical Activity Coaching". en. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Yokohama Japan: ACM, Apr. 2025, pp. 1–46. ISBN: 979-8-4007-1394-1. DOI:

10.1145/3706598.3713819. URL: https://dl.acm.org/doi/10.1145/3706598.3713819 (visited on 06/18/2025) (cit. on p. 4).

[8] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv:2106.09685 [cs]. Oct. 2021. DOI: 10.48550/arXiv.2106.09685. URL: http://arxiv.org/abs/2106.09685 (visited on 06/18/2025) (cit. on p. 5).

[9] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. *BLEURT: Learning Robust Metrics for Text Generation*. arXiv:2004.04696 [cs]. May 2020. DOI: 10.48550/arXiv.2004.04696. URL: http://arxiv.org/abs/2004.04696 (visited on 06/18/2025) (cit. on p. 5).

[10] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. *BERTScore: Evaluating Text Generation with BERT*. arXiv:1904.09675 [cs]. Feb. 2020. DOI: 10.48550/arXiv.1904.09675. URL: http://arxiv.org/abs/1904.09675 (visited on 06/19/2025) (cit. on p. 5).

[11] Qianli Wang, Tatiana Anikina, Nils Feldhus, Josef van Genabith, Leonhard Hennig, and Sebastian Möller. *LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools and Self-Explanations*. arXiv:2401.12576 [cs]. Apr. 2024. DOI: 10.48550/arXiv.2401.12576. URL: http://arxiv.org/abs/2401.12576 (visited on 06/18/2025) (cit. on p. 5).

[12] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*. arXiv:2303.16634 [cs]. May 2023. DOI: 10.48550/arXiv.2303.16634. URL: http://arxiv.org/abs/2303.16634 (visited on 06/18/2025) (cit. on p. 5).

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135 (cit. on p. 17).

[14] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Proceedings of Workshop on Text Summarization Branches Out (WAS 2004)*. Association for Computational Linguistics, 2004, pp. 74–81 (cit. on p. 17).

[15] Satanjeev Banerjee and Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*. Association for Computational Linguistics, 2005, pp. 65–72 (cit. on p. 17).