Politecnico di Torino

Computer Engineering - Artificial Intelligence and Data Analytics
A.a. 2024/2025
Graduation Session October 2025

Large Language Model for Personalized Well-being Recommendations

Supervisor:
Maurizio Morisio

Candidate:
Anjali Narendra Vaghjiani

Abstract

Although large language models (LLMs) show promise for medical healthcare applications, their utility for personalized health monitoring using wearable device data remains underexplored. Here we introduce the Personal Wellbeing Coach Large Language Model, designed for providing well-being recommendations in food, fitness, sleep, and social-emotional health. The growing availability of wearable fitness tracking devices is increasing, and along with it, demand and new opportunities for personalized health monitoring.

However, the application of LLMs for providing customized personalized well-being insights from multimodal wearable data remains unexplored. So this thesis presents the development of Personal Well-Being Coach LLM, a domain- adapted model designed to generate personalized recommendations in tasks like nutrition, fitness, sleep, and mental health. Our LLM is built on top of Meta-Llama-3, which was initially pre-trained on medical datasets and diagnostic corpora to enhance domain knowledge. To adapt the model for personalized well-being, we have further finetuned on a custom dataset derived from wearable device data (Fitbit) along with domain-specific knowledge sources, including longevity-focused books, clinical guidelines, and health-related websites. Textual data is then reformatted into question-answer pairs using embedding-based retrieval to ensure relevance.

Both benchmark corpora and direct user inputs were employed in evaluation, testing of semantic similarity, accuracy, coherence, and personalization. The fine-tuned Well-Being Coach demonstrated significant improvements over the baseline model, as semantic similarity (Sentence-BERT cosine) went from 0.0834 to 0.7678, while BERTScore F1, BLEU, and ROUGE measures, which reflect increased similarity, structure, and overlap between generated text and reference texts, all improved. These results imply that the fine-tuned model not only acquires the medical knowledge of the base model but also generates more coherent, relevant, and personalized recommendations. More importantly, it is found to combine multimodal signals from wearable sensors and generate actionable, context-aware advice for users as per individual user profiles.

This work also shows and contributes a) a novel fine-tuned LLM for personalized well-being monitoring, b) a handcrafted multimodal dataset that combines the wearable data with health knowledge, and c) a set of evaluation rubrics and benchmarks for assessing personalized in healthcare-focused LLMs. The results clearly highlighted the potential of domain-adaptation LLM for health management delivers safe, personalized, and actionable responses.

Table of Contents

List of Figures III

1 Introduction 1
1.1 Context and Motivation . . . 1
1.2 Objective . . . 2

2 Literature Review 3
2.1 Overview of Medical Large Language Model . . . 3
2.2 Why Choose Medical LLMs Over Generic Models? . . . 4
2.2.1 Key Medical LLMs in Research . . . 4
2.3 Limitations of Medical LLMs for Wellbeing Coach . . . 7
2.3.1 Lack of Personalization . . . 7
2.3.2 Limited Focus on Wellness . . . 8
2.3.3 Risk of Hallucinations . . . 8
2.4 Fine Tuning . . . 8
2.4.1 Domain Adaptation Techniques . . . 8
2.4.2 Data-Driven Training . . . 10
2.5 Evaluation . . . 11
2.5.1 Evaluation Approches . . . 11

3 Base model Selection 13
3.1 Candidate Models . . . 13
3.1.1 Selection Criteria for Base Model . . . 15
3.2 Evaluation Framework . . . 16
3.2.1 Evaluation Results of Candidate Models . . . 17
3.3 Prompt Sample for Evaluation . . . 18
3.4 Model Selection Decision . . . 19

4 Data Collection and Dataset Preparation 21
4.1 Data Collection . . . 21
4.1.1 Data Sources . . . 22

4.2 Dataset Preprocessing 22
4.3 Example of Input-Output Data Pairs 23
    4.3.1 Books and Research Documents for Training 25

5 Fine Tuning 27
5.1 Model architecture and training setup 28
    5.1.1 Base Model Selection 28
    5.1.2 Model Architecture 28
5.2 Experimental Setup and System Requirements 32
    5.2.1 Hardware Environment 32
    5.2.2 Software Environment 32
    5.2.3 Tools and frameworks used 33
    5.2.4 Justification on resources used 33

6 Evaluation 34
6.1 Model Comparison (Base Model vs Fine Tuned Model) 34
    6.1.1 Quantitative Comparison 35
    6.1.2 Prompt based evaluation 40
6.2 LLM as a Judge (G-EVAL) 45
    6.2.1 Results Visualizations 46

7 Summary 48
7.1 Motivation and Problem Statement 48
7.2 Approach Summary 48
7.3 Evaluation and Key Findings 49
7.4 Contribution 49
7.5 Limitations 50
7.6 Future Work and Recommendations 51

Appendix 53

Bibliography 66

List of Figures

4.1 Flow of data Argumentation from Kaggle 24
4.2 Flow of data Argumentation from Books and Websites 25

5.1 Model architecture and fine-tuning pipeline 28

6.1 Quantitative Evaluation Comparison of both models 37
6.2 Semantic Similarity comparison 38
6.3 BERTScore comparison 39
6.4 G-EVAL results of the well-being coach 47

Chapter 1

Introduction

1.1 Context and Motivation

Large language models (LLMs) have grown in recent years, showing revolutionary developments in various fields, including text generation, content creation, and natural language comprehension. Such models have expanding capabilities, and they are increasingly being integrated into various domains and a wide spectrum of applications. Also, researchers have begun to explore their potential in the medical and healthcare field, from aiding clinical decision-making to enhancing patient education and engagement. Many popular foundational models like GPT, LlaMa are versatile but require further fine-tuning for downstream tasks. While many foundational models can perform well in a generic context, they lack the domain-specific knowledge about most industrial or business applications, and especially in the medical and health domains. Recently, the reached and development towards Medical LLM is boosting, and it is doing quite well in the medical-specific domain to give answers and solve questions related to medicine. But there are no models so far in the market that can provide suggestions and recommendations to the user by analyzing their data and their progress in the well-being world. General models are capable enough to provide recommendations when asked with correct prompting, but their responses are sometimes inaccurate, or the model hallucinates due to missing domain-related training and tuning of the model. When general-purpose large language models provide health improvement suggestions, their responses are often repetitive and too generic, or unengaging, which makes it difficult for users to stay motivated and follow the advice. Effective well-being guidance should be not only accurate but also engaging and actionable, encouraging users to maintain healthy habits consistently. Currently, there are no LLMs capable of collecting data from users' fitness trackers, such as Fitbit, or health applications, converting that data into meaningful prompts, and generating personalized recommendations for diet, exercise, sleep, and social well-being. Additionally, this domain is highly sensitive and requires careful attention towards both user privacy and the accuracy of the recommendations. So to solve such challenges, this work is to develop a personalized well-being coach. The proposed model can intelligently analyze personalized user data and is trained to provide personalized, practical, and safe recommendations that could help in improvements in ones overall health while respecting user privacy and delivering guidance in an engaging, motivating way.

1.2 Objective

The primary objective of this thesis is to develop a domain-specific large language model capable of acting as a personalized well-being coach. The model must consider accurate, context-aware, and engaging recommendations personalized for individual user's lifestyles, dietary habits, fitness levels, sleep patterns, and health conditions.

1. **Selection and evaluation of model:** The first step towards the development of this model begins by selecting an appropriate base model. We began by considering the medical domain pretuned models and evaluated them using fitness-related prompts, providing user data as input to each candidate model. The model that produced the most accurate and relevant outputs for our domain was then selected as the base foundation for further development.

2. **Data for Fine Tuning:**Once the base model was selected, the next step was to identify the gap between the model's existing capabilities and the specific requirements for a personalized wellbeing coach. Based on this analysis, relevant data were collected to address these gaps, including books, website articles, blogs, and nutrition and fitness guidelines tailored to different diet types, user profiles, and regional contexts.

3. **Fine Tuning:** Once we have appropriate data that aligns with our goal, the model is ready for fine-tuning. This approach allowed the model to specialize in personalized well-being tasks while preserving its general knowledge.

4. **Evaluation:** Finally, after fine-tuning, the model has to be evaluated to ensure that it can generate accurate, relevant, and engaging personalized well-being suggestions, comparing general-purpose LLMs in both quality and usefulness.

Chapter 2

Literature Review

This chapter reviews the literature on developing domain-specific, personalized large language models (LLMs) for well-being. It explores the evolution of LLMs, the need for domain-specific models, and their applications in medical and wellness fields. The chapter also shows methods for personalization techniques and the importance of ethical considerations. It concludes by identifying gaps in the existing research, providing a clear rationale for the present study.

2.1 Overview of Medical Large Language Model

In recent years, large language models (LLMs) have catalyzed transformative advancements across a wide range of domains, from natural language understanding to content generation. Their growing capabilities have enabled their integration into diverse applications, including the medical field, where they are increasingly being used to support clinical decision-making and improve patient education and participation.

LLM plays a central role in healthcare, facilitating communication between patients and care providers. Progress in LLMs has made it possible to develop AI systems that understand and generate medical language, enabling richer interactions and collaboration between humans and AI. These models have shown impressive performance on multiple-choice medical benchmarks and possess extensive knowledge of medical conditions and regulations. However, despite their strengths, current medical LLMs are still less effective in providing personalized recommendations for fitness, diet, sleep and social well-being, considering unique user inputs. They often deliver generic advice, mainly because they lack knowledge and personalized guidance.

2.2 Why Choose Medical LLMs Over Generic Models?

When it comes to building a well-being coach, which is a personalized recommendation system, choosing a medical LLM over a generic one is not just a preference; it is necessary to consider the principles of safety, accuracy, and trust. While generic LLMs like ChatGPT are incredible generalists, capable of answering questions on a wide range of topics, they are trained on vast, unfiltered internet data, which can contain a lot of misinformative data. This generalist approach is a significant risk in the healthcare domain. In the context of health and wellness, this poses a serious risk, as a generic model might produce plausible-sounding but incorrect medical advice, an error that could have serious consequences for a person's health.

In contrast, a medical LLM is a specialist. It is specifically fine-tuned on a selected, high-quality dataset of medical knowledge, including peer-reviewed research, clinical guidelines, and verified patient data. This specialized training allows models to understand complex medical terminology and nuance, providing recommendations that are not only accurate but also clinically sound. Ultimately, a well-being tool must be built on a foundation of reliability and safety. A medical LLM offers a solid foundation, ensuring that the advice it gives is not just a guess from the internet, but an informed recommendation backed by a wealth of specialized knowledge, which is critical for a tool that impacts a person's health and wellness.

2.2.1 Key Medical LLMs in Research

There are numerous open-source, fine-tuned medical large language models (LLMs) available today, providing a strong and valuable foundation for developing domain-specific applications. Many of these models can be accessed and utilized through platforms such as Hugging Face, which hosts a wide range of pre-trained and fine-tuned models suitable for further customization. For this thesis, several commonly used medical LLMs were considered as potential base models. Previous studies have also listed and analyzed many different healthcare-domain LLMs, providing valuable insights about their capability and use cases [1]. A selection of these models, along with their key features, is summarized in the table 2.1.

Model	**Features**	**Base model**	**Developer**	**No. of Parameters**	**open**	**local**	**Usage cost**
BioBERT	Pre-trained on biomedical literature (PubMed, PMC) for biomedical text mining, question answering, and named entity recognition	BERT	Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang	110M	✓	✓	16 GB GPU
DeepSeek Quen 8B Medical Reasoning	This model was fine-tuned on deepseek-ai/Deep SeekR1-0528-Qwen3-8B using a medical reasoning dataset (mamachang/medicalreasoning) with 4-bit quantisation for memory-efficient training.	Deep Seek-R1-0528-Qwen3-8B	Abid Ali Awan	8B	✓	✓	12 GB GPU
MedPalm	Med-PaLM is a large language model (LLM) designed to provide high-quality answers to medical questions.	PaLM2	Google Research	340B	✗	✗	paid via google API

BioMistral	BioMistral is a suite of Mistral-based further pre-trained open source models suited for the medical domain and pretrained using textual data from PubMed Central Open Access.	Mistral-7B-Instruct-v0.1	Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gouraud, Mickaël Rouvier, Richard Dufour,	7B	✓	✓	16GB VRAM
Contact Doctor	This model is a fine-tuned multimodal version of Meta-Llama-3-8B-Instruct on custom "BioMed-Data" text and image datasets.	Llama-3-8B-Instruct	Contact Doctor Healthcare Private Ltd	8.03B	✓	✓	16 GB VRAM
MedGemma	MedGemma demonstrates advanced medical understanding and reasoning on images and text.	Gemma3	Google Research	4B and 27B	✓	✓	12-32GB GPU

MedAlpaca 7B	An advanced LLM, fine-tuned specifically for medical question-answering. Built upon the foundations of Alpaca, it utilizes a diverse array of medical texts.	LLaMA	Tianyu Han, Lisa C.Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figueroa, Alexander Löser, Daniel Truhn, Keno K. Bressem,	7B	✓	✓	12GB RAM

Table 2.1: Key Medical LLMs and Their Features and Properties

2.3 Limitations of Medical LLMs for Wellbeing Coach

2.3.1 Lack of Personalization

One major limitation of existing medical LLMs is their lack of personalization. While these models are highly knowledgeable in clinical and medical domains, they are primarily trained on general datasets and provide broad, one-size-fits-all recommendations. They usually do not take into account individual user demographic data such as age, gender, health conditions, dietary preferences, activity levels, or lifestyle habits. In the context of well-being applications, this limitation is particularly important to ensure that the personalized recommendations are safe, relevant, and actionable. Without personalization, advice on nutrition, exercise, or mental health may be ineffective, unmotivating, or even potentially harmful. Addressing this gap is a key motivation for the present work, which incorporates user-specific data collected from a fitness belt or fitness app used to generate tailored, context-aware suggestions that can improve engagement and overall outcomes in well-being coaching.

2.3.2 Limited Focus on Wellness

Existing medical LLMs have a key limitation: they're designed for clinical problems, not for everyday well-being. While these models are great at understanding diseases, treatments, and scientific research, their focus is narrow. They lack the training to provide meaningful, realistic guidance on things like nutrition, fitness, sleep, or mental and social health. This means they can't offer the kind of practical, personalized advice people need to improve their daily habits and overall health. To fix this, we're fine-tuning a medical LLM with a new kind of data. By training it specifically on information about nutrition, physical activity, sleep, and mental wellness, we're giving it the knowledge it needs to generate truly useful and personalized recommendations for a healthier lifestyle.

2.3.3 Risk of Hallucinations

A big problem with medical LLMs is that they can "hallucinate," meaning they make up information that sounds right but actually is wrong. Even though these models know a lot about medicine, their knowledge of things like diet, fitness, sleep, and mental health can be incomplete. This knowledge gap can cause them to give different answers to the same question or create recommendations that don't make sense. If a user asks for advice and the model just speculates to fill in what it doesn't know, it can provide misleading or even unsafe information.

2.4 Fine Tuning

Fine-tuning large language models(LLMs) is important for accommodating these advanced algorithms to specific tasks, disciplines. This process enhances the model's performance on technical tasks and significantly broadens its connection across colorful fields. This means we can take advantage of the model's being knowledgeable while further training it to perform our specific tasks, perfecting delicacy, and making it more applicable for the particular use case.

What are the possible ways to do it? How can we further train the model? There are a few ways to adapt and train a model. Each approach has its own benefits and drawbacks.

2.4.1 Domain Adaptation Techniques

Full Fine-Tuning

Full Fine-tuning is the most straightforward system of conforming a pre-trained model to a new task or dataset. In this process, all the parameters of the model, including the weights in the layers, are streamlined during training. The model starts with a pre-trained set of weights and also undergoes fresh training to acclimate its parameters to perform better on the new data or task. The primary advantage of full fine-tuning is its capability to completely optimize a model for a specific task. Since all parameters are modified during the fine-tuning process, this method allows the model to fully adapt to the target domain, achieving high delicacy and task-specific performance. Still, this system has some downsides. First, full fine-tuning requires a significant quantum of computational resources, especially for large models. The need for large memory capacity and extended training times can be a hurdle for numerous inventors. also, fine-tuning all parameters might lead to overfitting, particularly when the size of the dataset is small or lacks diversity. To break this problem, an indispensable approach has been under consideration, similar to LoRA.

LoRA: Low-Rank Adaptation

Low-Rank adaption(LoRA) was introduced as a more effective solution to full fine-tuning. Rather than streamlining all the parameters of the model, LoRA focuses on modifying a lower subset of the model's weights by introducing low-rank matrices into the layers of the model. Specifically, LoRA decomposes the weight updates into two lower matrices of lower rank, which are much more computationally effective to train.

LoRA operates on the principle that numerous deep literacy models, especially large models, don't require full updates to their parameters to achieve task-specific performance advancements. rather, a low-rank approximation of the weight updates is sufficient. By indurating the original model parameters and only training the low-rank matrices, LoRA reduces the number of parameters that need to be streamlined, leading to a dramatic reduction in memory operation and computational cost. One trade-off is that LoRA may not always achieve the same position of performance as full fine-tuning for further complex tasks, as the low-rank approximation may not prisoner all the necessary nuances of the data.

qLoRA: Quantized Low-Rank Adaptation

Quantized Low-Rank adaptation (qLoRA) is an extension of LoRA that incorporates quantization into the low-rank adaptation process. Quantization refers to the fashion of reducing the precision of the model's parameters, similar to converting floating-point values to lower-precision formats(e.g., 8-bit integers rather of 32- bit floats). This process reduces the memory footprint and computational conditions indeed further, making it possible to fine-tune large models on indeed more constrained hardware. QLoRA combines the benefits of LoRA's low-rank adaptation with quantization's memory effectiveness. By quantizing both the pre-trained model weights and the low-rank adaptation matrices, qLoRA reduces the overall computational burden of the fine-tuning process. This system is particularly useful in surroundings where offers such as memory, storehouse, and computational power are limited, such as in edge bias or in situations requiring rapid-fire model deployment.

The crucial advantage of qLoRA is its capability to maintain the effectiveness of LoRA while also reducing the size of the model, allowing for indeed brisk training and conclusion. The trade-off is that quantization can lead to a loss in model perfection, which may affect performance. In practice, qLoRA's performance is frequently similar to LoRA but with further reduced memory and computational outflow.

2.4.2 Data-Driven Training

These approaches focus on how the model learns from data.

Supervised Learning

Supervised fine-tuning involves conforming a pre-trained Large Language Model(LLM) to a specific downstream task using labeled data. In supervised fine-tuning, the fine-tuning data is collected from a set of responses validated beforehand. That's the main difference from the unsupervised ways, where data isn't validated beforehand. While LLM training is(generally) unsupervised, Finetuning is(generally)supervised.

The supervised fine-tuning process allows the model to learn task-specific strokes and nuances present in the labeled data. By conforming its parameters according to the specific data distribution and task conditions, the model becomes specialized for performing the target task.

For our well-being trainer, many books, exploration papers, and other high-quality resources on fitness, nutrition, internal health, and life guidance were precisely collected. This information was also converted into structured input-output pairs suitable for supervised learning.

Few-shot learning

There are some cases where collecting a large labeled dataset is impractical. Few-shot learning is one of the ways to make a model understand a task by providing a few examples, or "shots," at the beginning of the prompt. This gives the model the necessary context without the need for extensive fine-tuning.

2.5 Evaluation

As LLMs have recently gained significant attention due to their remarkable capabilities in the performance of various domains. However, it is very important to evaluate these models across diverse tasks, checking their performance accuracy, correctness, and reliability.

There are many ways and techniques for the evaluation, but to begin with, one of the very first things required is the benchmark. The benchmarks can be categorized into general capability benchmarks or specialized and diverse benchmarks. As from their name, their task is also to evaluate generic performance and domain-specific task performance, respectively.

Evaluation is not only essential for evaluating the accuracy of a model after training, but it also serves as a key strategy for selecting the most suitable base model for domain-specific fine-tuning. Once we have an appropriate benchmark and the model for the training or inference, the next steps towards the evaluation are to design the prompt and set the decoding parameters for the response generation.

2.5.1 Evaluation Approches

Evaluation approaches can be divided into three main orders as automatic evaluation, mortal evaluation, or LLMs as a judge. Numerous automatic evaluation styles attempt to measure the same rates of a summary that mortal observers would consider. Those rates include ignorance, consonance, applicability, factual thickness, and fairness. Similarity in content or style to a reference textbook can also be an important quality of a generated textbook.

So there are two approaches in automatic evaluation, which would be reference-grounded and reference-free. **Reference grounded** This system evaluation standard is grounded on comparing the generated textbook to the reference-grounded mortal ground verity textbook, numerous of which are developed for traditional NLP tasks. Many implementations are N-gram grounded criteria. Metrics BLEU(Bilingual Evaluation Understudy), ROUGE(Recall- acquainted Understudy for Gisting Evaluation), and JS divergence JS2 are imbrication grounded criteria that measure the similarity of the affair textbook and the reference textbook using n-grams.

Reference-free(environment-grounded) criteria produce a score for the generated textbook and don't calculate on ground verity. Evaluation is substantially grounded in the environment or source documents; numerous of these criteria were originally designed to address the challenge of creating ground truth data. Quality-grounded criteria similar to SUPERT and BLANC are substantially concentrated on assessing the quality of the content generated. While another metric, ROUGE-C, is a revision of ROUGE without the need for the reference, and uses the source textbook as the environment for the comparison. Reference-free criteria have shown better correlations to the mortality evaluations compared to reference-grounded criteria, but there are limitations to using them as the single measure of process on a task.

LLM- grounded Evaluation or LLM as a Judge, it can be either prompt-grounded, where an evaluation is grounded on a prompt to judge some textbook, or embedding-grounded criteria.

Chapter 3

Base model Selection

In this chapter, we describe the process of selecting a suitable base model for developing the personalized well-being coach. Since the performance of the final system strongly relies on the choice of the base model, many models were considered, and a careful evaluation was carried out. Several open-source large language models (LLMs), including some fine-tuned on medical and health-related domains, were explored. The goal was to identify the model that not only performed best on well-being-related tasks but also demonstrated robustness, reliability, and adaptability for further fine-tuning.

The core limitation of general-purpose LLMs such as ChatGPT, Gemini, and others raises our privacy concerns. Beyond accuracy issues, these models often suffer from hallucinations, provide recommendations that overlook established guidelines, and fail to account for individual preferences. Since they are trained on broad, general-purpose datasets, they lack the depth of domain-specific knowledge necessary for well-being applications. In particular, general models struggle to fully understand medical terminology, psychological principles, and physiological processes required to deliver safe and meaningful health guidance. For this reason, we chose to work with an open-source, medically fine-tuned LLM as the base model. Unlike general-purpose models, these are specifically trained on health-related data, making them more reliable and relevant for building a personalized well-being coach.

3.1 Candidate Models

To begin our project, our first task was to look at the field of available AI models. For this, we started exploring platforms like Hugging Face, which acts as a vast library for open-source LLMs. Our search was specifically focused on finding models that had already been trained or "fine-tuned" for medical and healthcare purposes.

We chose this approach because a model with a medical background starts with a significant advantage. Unlike generic LLMs, which might mix up medical facts with general internet chatter, these specialized models already have a deep and reliable understanding of health-related concepts. This stronger foundation means they are far more likely to provide accurate and safe well-being recommendations, saving us from having to teach them the fundamentals of health from scratch.

A few examples of the considered model are shown in the Table 3.1.

Model	**Features**	**Link**
DeepSeek	This model was fine-tuned on deepseek-ai/DeepSeek-R1-0528-Qwen3-8B using a medical reasoning dataset (mamachang/medical-reasoning) with 4-bit quantization for memory-efficient training.	[2]
BioMistral	BioMistral is a suite of Mistral-based further pre-trained open source models suited for the medical domain and pre-trained using textual data from PubMed Central Open Access.	[3]
ContactDoctor	This model is a fine-tuned multimodal version of Meta-Llama-3-8B-Instruct on custom "BioMedData" text and image datasets.	[4]
MedAlpaca	MedAlpaca-7B a large language models specifically fine-tuned for medical domain tasks. It is based on LLaMA (Large Language Model Meta AI) and contains 7 billion parameters.	[5]

Table 3.1: Candidate Medical LLMs with References

3.1.1 Selection Criteria for Base Model

The base model was selected based on a few criteria such as:

Domain Knowledge

When selecting the base model, we try to select the model that has a solid understanding of both medicine and general wellness. While a strong background in medical knowledge was crucial for safety and accuracy, a model that can understand broader topics like nutrition, fitness, sleep, and mental health was necessary. This combination of clinical expertise and general well-being knowledge was a key factor towards the decision, as it allowed us to provide truly comprehensive and personalized guidance to our users.

Multiple prompts covering similar tasks were tested on the candidate models, and the model that consistently produced the most accurate outputs was selected as the base model. The reliability of the outputs was further evaluated using prompt engineering and repeated measurements, where the same prompt was provided multiple times to verify the consistency of the model and also to ensure the model did not hallucinate. Additionally, evaluations were also conducted using other LLMs as judges, comparing responses against one another, and by measuring the outputs against verified healthcare datasets containing supervised input-output pairs. This comprehensive evaluation ensured that the selected model could provide accurate, consistent, and trustworthy well-being recommendations.

Model Size and Computational Feasibility

Another aspect we must consider while selecting the base model is the size of the model, because it can directly affect the computational resources required for inference and fine-tuning. Models that were too large could be impractical to train or deploy efficiently, while smaller models might lack sufficient knowledge or reasoning capabilities. A balance between performance and resource requirements was therefore necessary. For this aspect, we considered using and testing a small model of around 8B parameters.

Transparency and Ownership

One of the other important factors considered during the base model selection was its open-source property, which can allow us to run it locally. Using the open-source model, we gain the freedom to pre-train weights, which allows us to fine-tune it for the specific needs of our well-being domain. Running the model locally also means that user data and model outputs stayed in our control, ensuring to protect privacy and have proper ownership of the data. This criterion worked hand-in-hand with performance and resource considerations, giving us both the flexibility to experiment and the confidence that personal health information would be handled responsibly.

3.2 Evaluation Framework

To select the best model among the candidate models, an evaluation was carried out. A prompt-based testing framework was developed, where a diverse set of well-being-related prompts was designed to cover:

- Nutrition and diet
- Fitness and physical activity
- Sleep hours
- Mental and social wellness
- Overall Suggestions

All candidate models were tested using the same set of prompts to ensure a fair comparison. The outputs were evaluated based on several criteria listed below:

- **Accuracy and Relevance:** Whether the recommendations were factually correct and applicable to the user scenario.
- **Personalization Potential:** The model's ability to adapt advice based on user-specific information.
- **Clarity and Motivation:** Whether the responses were easy to understand, engaging, and motivating for the user.
- **Consistency:** The ability to provide stable and coherent answers across repeated prompts, avoiding contradictions or hallucinations.

In addition to that, the qualitative domain-specific evaluation was performed to make sure that the model can handle queries and has proper domain knowledge. For the same metrics, like Semantic Search, BERTScore, BLUE, ROUGE-1, ROUGE-2, and ROUGE-L were employed. The output of such an evaluation is shown in table 3.2.

Table 3.2: Evaluation results of candidate models across four domains (Food, Fitness, Mental Health, Fitbit Data).

Metric	**Contact Doctor**				**DeepSeek Qwen3 8B**			
	Food	Fitness	Mental	Fitbit	Food	Fitness	Mental	Fitbit
Semantic Search	0.6435	0.5020	0.6805	0.0843	0.5282	0.6794	0.7036	0.0748
BERTScore	0.8058	0.8249	0.8677	0.8709	0.8136	0.8382	0.8715	0.8448
BLEU	0.0000	0.0114	0.0296	0.0000	0.0000	0.0246	0.0385	0.1213
ROUGE-1	0.1736	0.1638	0.2505	0.0028	0.0637	0.2671	0.2850	0.4827
ROUGE-2	0.0058	0.0248	0.0604	0.0000	0.0013	0.0573	0.0772	0.1647
ROUGE-L	0.1712	0.1130	0.1567	0.0028	0.0571	0.1542	0.1847	0.2238

Metric	**MedAlpaca**				**BioMistral**			
	Food	Fitness	Mental	Fitbit	Food	Fitness	Mental	Fitbit
Semantic Search	0.1490	0.3861	0.6408	0.0527	-0.0356	0.4631	0.6134	0.0513
BERTScore	0.2959	0.7994	0.8712	0.7684	0.0000	0.8258	0.8788	0.7547
BLEU	0.0000	0.0082	0.0244	0.0000	0.0000	0.0117	0.0417	0.0000
ROUGE-1	0.0241	0.1249	0.2345	0.0086	0.0000	0.1303	0.2679	0.0025
ROUGE-2	0.0012	0.0211	0.0502	0.0000	0.0000	0.0307	0.0607	0.0000
ROUGE-L	0.0194	0.0895	0.1599	0.0057	0.0000	0.0946	0.1778	0.0025

3.2.1 Evaluation Results of Candidate Models

Table 3.2 presents the quantitative evaluation of the candidate LLMs across four domains: Food, Fitness, Mental Health, and Fitbit Data, using multiple metrics including Semantic Search, BERTScore, BLEU, ROUGE-1, ROUGE-2, and ROUGE-L. Specifically, for each model, the metric score for a domain was calculated as the average score of all prompts from that domain, providing a summary measure of performance. The results shows why ContactDoctor was selected as a base model for our well-being coach.

Metrics Description

- **Semantic Search** computes embedding-based similarity between model outputs and human reference responses. For instance, Contact Doctor achieved the highest Semantic Search in Food (0.6435) and Mental Health (0.6805), indicating good semantic alignment with human references. We can also see that DeepSeek Qwen3 model shows best performances. But other model simply shows it could not perform well in food and fitness with (0.149 and -0.0356).

- **BertScore** evaluates contextual token-level similarity. Contact Doctor and DeepSeek achived great results with score between (0.8-0.9) reflecting accurate semantic and contextual comprehension.

- **BLEU** measures exact word overlap with reference answers, so it works best when responses should match closely. In our wellness tasks, scores were low because models often rephrased or personalized answers. This shows a limitation of BLEU, low scores don't always mean poor quality, but can reflect flexibility or personalization.

- **ROUGE-1/2/L** compute overlapping unigrams, bigrams, and longest common subsequences. Contact Doctor's ROUGE-1 in Food (0.1736) and ROUGE-2 in Mental Health (0.0604) indicate that it can capture significant content phrases and structural coverage of reference responses.

Accuracy is also most visible in BLEU, ROUGE, and BERTScore, which estimate how closely generated responses approximate reference responses. Contact Doctor's high ROUGE and BERTScore in Food and Mental Health indicate precise domain content.

Personalization is indirectly demonstrated by Semantic Search and BERTScore, since more semantic similarity suggests the model to comprehend prompt context and user-specific details. Doctor's Fitbit Semantic Search (0.0843) and BERTScore (0.7709) reflect its capability to provide contextually appropriate advice for activity tracking.

3.3 Prompt Sample for Evaluation

Approximately 45-50 prompts were designed thoughtfully from the four foundations of the system: fitness, diet, sleep, and mood. They were not only meant to test the models' background knowledge in these areas but also test how well an LLM could map user-provided information in a manner that would generate personalized suggestions. Second, the prompts were employed to try the simple accuracy and consistency of each selected model compared to stable reference sources. An example of these prompts and sample model responses is found in Appendix 7.6.

What do we observe from this prompts?

- **Prompt 1** (stress activity): Contact Doctor provided mechanisms like decreased cortisol and increased endorphins in patient-friendly language, whereas DeepSeek was also correct, and MedAlpaca/BioMistral were brief. This aligns with Contact Doctor's and Deep Seek higher ROUGE-2 in Mental Health, reflecting its ability to capture causal bi-grams.

- **Prompt 2** (aerobic exercise): Contact Doctor and DeepSeek both were accurate with rationale, but MedAlpaca was factually wrong, consistent with its weaker Fitness semantics and ROUGE scores.

- **Prompt 3** (pre/post-workout nutrition): Contact Doctor offered actionable menus, timing, and hydration guidance, while DeepSeek was correct but generic, and other models produced brief or repetitive outputs. This mirrors the superior Food ROUGE scores of Contact Doctor.

- **Prompt 4** (personalized day plan): Contact Doctor used the user's profile (mood, yoga, steps, hydration) and included safety nuances, whereas BioMistral remained generic, showing the model's superior semantic alignment and personalization capability.

3.4 Model Selection Decision

The final selection of the base model for our domain well-being coach was evaluated across different wellness domains such as food, fitness, mental health, social wellness. Model was not selected solely based on generic knowledge but its ability to generate the personalized responses based on the user data such as gender, age, height, weight, daily_steps, food_consumptions and sleep pattern. These parameters were essential for well-being coach to generate the personalized recommendations.

While the model like MedGemma showed strong medical knowledge, due it its large size, it was computationally expensive to run, limiting its practical usability. Other models, such as BioMistral and MedAlpaca, could generate responses for the generic tasks but struggled when presented with personalized user data. As we could see in the evaluation results presented in Table 3.2. For example, BioMistral achieved a high BERTScore in the Mental domain (0.8788) but showed poor performance for Food and other domains(BERTScore 0, Semantic Search 0.0356), making it unreliable across categories. Similarly, MedAlpaca showed decent Fitness performance (BERTScore 0.7994) but very weak scores in Food and other domains(Semantic Search 0.1490, ROUGE-1 only 0.0241), indicating difficulty in handling nutrition-specific queries. Both models also failed to provide responses for Fitbit data.

Semantic Search and BERTScore highlight the model's ability to generate contextually relevant recommendations considering user queries. Contact Doctor LLM consistently achieved higher semantic similarity and BERTScore values across diet, fitness, mental health, and Fitbit-based user data benchmarks compared to other models.

BLEU and ROUGE scores evaluated the structural similarity and coherence of the generated responses. **Contact Doctor and DeepSeek** both the LLM demonstrated higher ROUGE-1, ROUGE-2, and ROUGE-L scores, indicating that its outputs were not only accurate but also well-structured and aligned with the intended response templates.

In contradictory, other models showed variability across different metrics, especially while generating the user's personal input data collected from their fitness watch.

We selected Contact Doctor as the base LLM because it returned the most balanced and uniform results at all times. While DeepSeek Qwen did obtain slightly higher on some metrics, but their answers weren't always more tailored and detailed. For example, DeepSeek did well in reference text overlap in fitness, but Contact Doctor did better on food-related tasks (Semantic Search 0.64 vs. 0.53; ROUGE-1 0.17 vs. 0.06). In our timely assessments, Contact Doctor stood out by using user-specific data like steps, sleep, and diet to generate short, actionable suggestions, whereas the remaining models were shallow or merely could not generate output and hallucinated with non-relevant outcomes. This combination of high scores, enhanced personalization, and consistent performance across domains made Contact Doctor the most practical and effective choice.

Chapter 4

Data Collection and Dataset Preparation

This chapter describes the process of collecting and preparing domain-specific data to fine-tune the selected base model into a personalized well-being coach. Since the model's performance mostly depends on the quality and relevance of the training data, careful consideration was given to sourcing, curating, and structuring the dataset.

4.1 Data Collection

Since the quality of data is one of the important factors determining both the performance and accuracy of large language models. Several books and articles covering diverse aspects of wellbeing, such as fitness, nutrition, mental health, social health, and longevity, were collected and processed into a structured dataset. Given that the base model already contained broad wellness knowledge from its prior training on medical-related data, the focus was mainly placed on collecting highly domain-specific information that could be used to enhance personalization and context knowledge for the model.

In addition, an open-source dataset containing user fitness data (e.g., from Fitbit) was found on Kaggle [6] was used, since such data represents the actual inputs the LLM would receive in real-world usage. The data set was reformatted into supervised learning examples, where user data served as input, and depending on the user, the personalized recommendation was an output.

4.1.1 Data Sources

Data was gathered from multiple sources to ensure comprehensive coverage of well-being topics:

- **Books and Research Articles** - High-quality resources were collected on various domains such as nutrition, fitness, mental health, social well-being, and longevity. A few of the examples are listed below:
    - Dan Buettner. *The Blue Zones Solution: Eating and Living Like the World's Healthiest People* [7]
    - MICHAEL GREGER, M.D. *How not to Die* [8]
    - LUIGI FONTANA, *THE PATH TO LONGEVITY* [9]
    - PETER ATTIA, MD, *OUTLIVE THE SCIENCE and ART OF LONGEVITY* [10]
    - HÉCTOR GARCÍA AND FRANCESC MIRALLES, *IKIGAI The Japanese Secret to a Long and Happy Life* [11]
    - Other additional data from:
        - Government and Health Organization Data:- WHO, CDC, NHS, or other official health guidelines on nutrition, physical activity, and mental well-being.[12], [13], [14], [15], [16], [17], [18], [12],[19], [20].
- **Open-Source User Data** – Datasets from fitness trackers (e.g., Fitbit [6]) and other publicly available wellness datasets were utilized to simulate real-world user inputs. These datasets were then transformed into structured input-output pairs to create custom templates for supervised fine-tuning.
- **Websites and Blogs** – Expert-verified online content and many more guidelines and information related to lifestyle, sleep, and exercise routines were also collected and used for model learning.[21], [22], [23], [24]

4.2 Dataset Preprocessing

This is the most important and often the most time-consuming stage. It involves collecting, cleaning, formatting, and annotating the data that will be used for fine-tuning. The quality of this dataset directly impacts the performance of the fine-tuned model. This stage may include data augmentation, deduplication, and splitting into training, validation, and test sets.

This could include removing personally identifiable information (PII), irrelevant conversational filler, spam, and badly formatted entries. Standardize greetings and closings. Convert the raw data into a structured format suitable for instruction fine-tuning.

After the useful data were collected, including books, articles, research articles, and dietary guidelines considering different age groups and diet-specific meal plans, the next step is to preprocess the data to make it more suitable for supervised fine-tuning for a well-being coach.

The first step towards converting the textual knowledge involved converting data resource knowledge into question-answer pairs. This structured format allowed the model to learn rules and guidance from these sources, teaching the system to think before it provides personalized recommendations. For example, nutritional guidance for various diets or age-specific meal recommendations was transformed into inputs representing user queries and outputs representing the appropriate response.

Next, open-source datasets from Kaggle were used, such as Fitbit user data. These datasets contained user fitness data, including age, sex, height, weight, daily step counts, sleep duration, exercise sessions, and minutes of workouts. Since the goal of the well-being coach was to provide weekly personalized recommendations, daily data was converted from daily into weekly summaries using Python libraries. For the diet consumption of week dummy randomized weekly food intake values, including fruits, vegetables, legumes, dairy, grains, meat, and nuts, were added to simulate realistic dietary patterns.

The final preprocessed dataset included around 1,500–2,000 input-output example pairs. Because our base medical LLM already had a strong understanding of general well-being concepts, the main purpose of this dataset was not to teach basic wellness knowledge, but to train the model to generate responses according to a specific, structured template. This approach ensured that the fine-tuned model could produce consistent, personalized, and well-organized recommendations customized to each user scenario.

For that, output responses were carefully created for each unique user's data, following a predefined template designed to maintain structure and clarity. The template was organized into five key sections.

4.3 Example of Input-Output Data Pairs

This section shows a few examples of how the collected from Kaggle and converted into supervised training examples (input-output pairs). Each input is real user fitness data, while each output is a personalized response that the model is expected to generate, considering the user input and predefined template. The sole purpose of such data training was to teach the model to follow the strict template of response. As we evaluated previously, the base model already had good basic

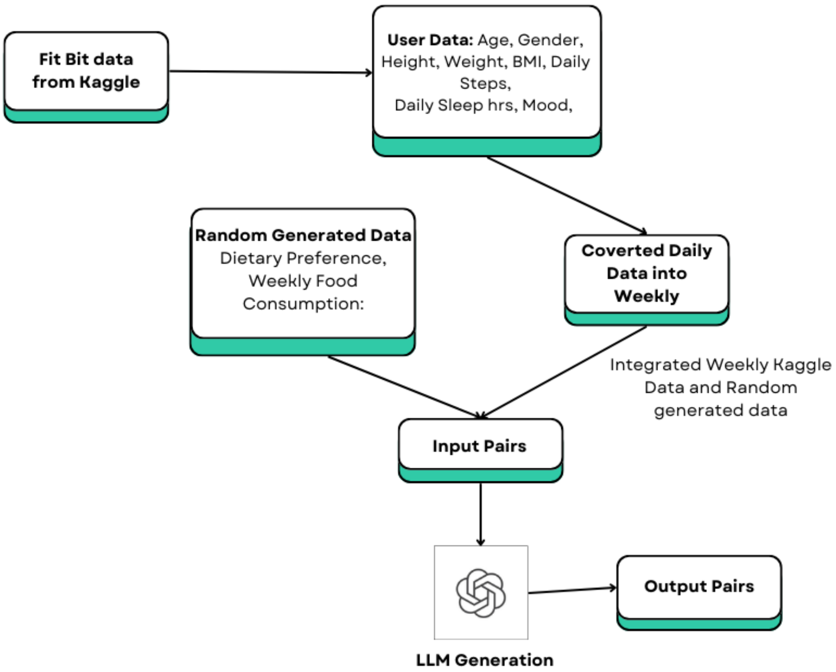

Fit Bit data from Kaggle
User Data: Age, Gender, Height, Weight, BMI, Daily Steps, Daily Sleep hrs, Mood,
Random Generated Data Dietary Preference, Weekly Food Consumption:
Coverted Daily Data into Weekly
Integrated Weekly Kaggle Data and Random generated data
Input Pairs
Output Pairs
LLM Generation

Figure 4.1: Flow of data Argumentation from Kaggle

general knowledge but was not able to follow the template, so supervised techniques were the best way to fine-tune the model for template learning.

However, it is often challenging to find sufficient high-quality datasets for personalized recommendations, especially in domains where user-specific information is limited or sensitive. To address this limitation, synthetic data was created for the training process. Synthetic data can be generated using generative models, which help simulate diverse user scenarios without relying solely on real-world datasets. Some common techniques for generating synthetic training data include:

- **Data augmentation:** Modifying existing data (e.g., by adding noise, rephrasing text, or scaling values) to increase diversity.
- **LLM-based generation:** Using large language models to generate realistic input–output pairs based on a template or seed examples. For this task, **GPT-4o-mini** was employed via API to produce the examples, some of which are illustrated in Appendix 7.6.

As shown in Figure 4.1, the input–output pairs were created using the FitLife

Health and Fitness Tracking Dataset from Kaggle KaggleData . From this dataset, we extracted key features such as age, gender, height, weight, BMI, daily steps, sleep hours, and stress level. The raw records were then converted into weekly data using Python by date-based grouping. To the dataset, we also introduced additional variables, such as dietary preferences and food choices, generated programmatically in Python. These combined features formed the input pairs, which were then passed to the LLM (GPT-4o-mini in our case) along with structured prompts to generate the corresponding output pairs for our well-being domains.

4.3.1 Books and Research Documents for Training

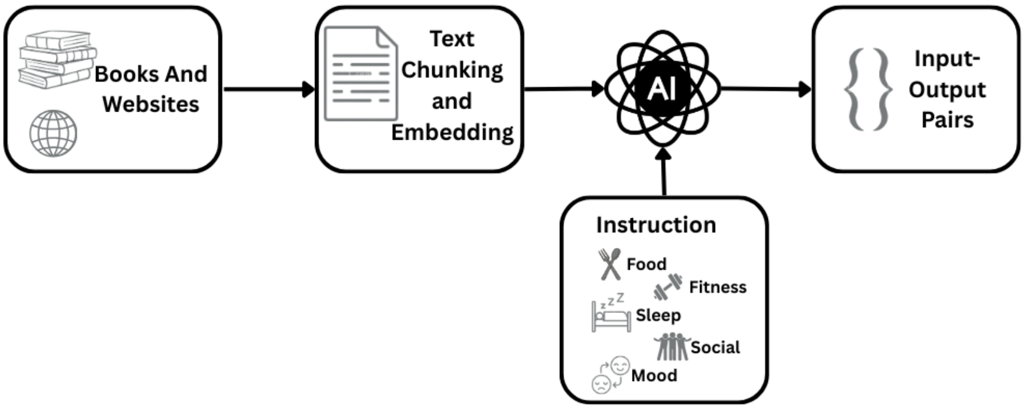


Figure 4.2: Flow of data Argumentation from Books and Websites

Other than the Fitbit dataset, various high-quality books and research articles mention in section 4 covering nutrition, fitness, mental health, social well-being, and longevity were carefully reviewed and used. The book content was first embedded through LangChain embeddings, then split into smaller chunks to preserve context. Each chunk was given as the input to GPT-4o-mini with explicit instructions to generate questions and answers in relation to the main well-being categories: fitness, food/nutrition, sleep, and social/mood. Approximately 4,000 such examples were created to fine-tune the model; a few examples are shown in table 4.1. This dataset's main goal is to help the model learn foundational guidelines and domain knowledge, which would be useful to generate structured, accurate, and personalized responses during inference.

Table 4.1: Supervised Training Data

Input	**Output**
What are some tips for better sleep hygiene?	Establish a regular sleep schedule, create a restful environment, limit screen time before bed, and practice relaxation techniques to improve sleep quality. Prioritize a comfortable sleeping environment and ensure your bedroom is dark and quiet.
What are the benefits of intermittent fasting for health and longevity?	Intermittent fasting can lead to weight loss, improved metabolic health, and reduced risk of chronic diseases. It may also enhance cellular repair processes and promote longevity. Consider the 16:8 method, where you fast for 16 hours and eat during an 8-hour window.
What role does physical exercise play in cognitive health?	Regular physical exercise can help prevent cognitive decline and reduce the risk of dementia, contributing to overall brain health.
How does nutrition play a role in emotional health?	Nutrition can greatly impact emotional health; a diet rich in omega-3 fatty acids, antioxidants, and vitamins can help improve mood and reduce symptoms of anxiety and depression. Foods like fatty fish, nuts, fruits, and vegetables are beneficial for mental health.
What is the recommended protein intake for metabolic health benefits?	A moderate reduction in protein intake, close to 7–9% of total caloric intake, is recommended for metabolic health benefits, as it has been shown to improve metabolic markers and promote weight loss while reducing fasting blood glucose levels.
How can incorporating whole grains into my diet benefit my health?	Regular consumption of whole grains can induce a 20–40% reduction in the risk of cardiovascular disease and a 20–30% decrease in the risk of type 2 diabetes. Whole grains help slow the absorption of carbohydrates, reducing hunger and the risk of obesity.
What lifestyle changes can enhance fitness and overall health?	Engaging in regular physical activity, combined with a diet rich in fruits, vegetables, and whole grains, can significantly enhance fitness levels and overall health.

Chapter 5

Fine Tuning

In this chapter, we describe our fine-tuning of our selected base model to be used with our personalized well-being coach. Fine-tuning is the act of taking a pre-trained model and adapting it to use in our specific domain, so that it may generate structured, accurate, and personalized recommendations based on user data. In our effort, we did not start from a scratch model but took a strong pre-trained medical-domain LLM and adapted it to our curated wellness dataset. This utilizes the knowledge that the model already possesses and fits it to the precise needs of our application. Figure 5.1 can shows architecture of how models are pre-trained and then further fine-tuned.

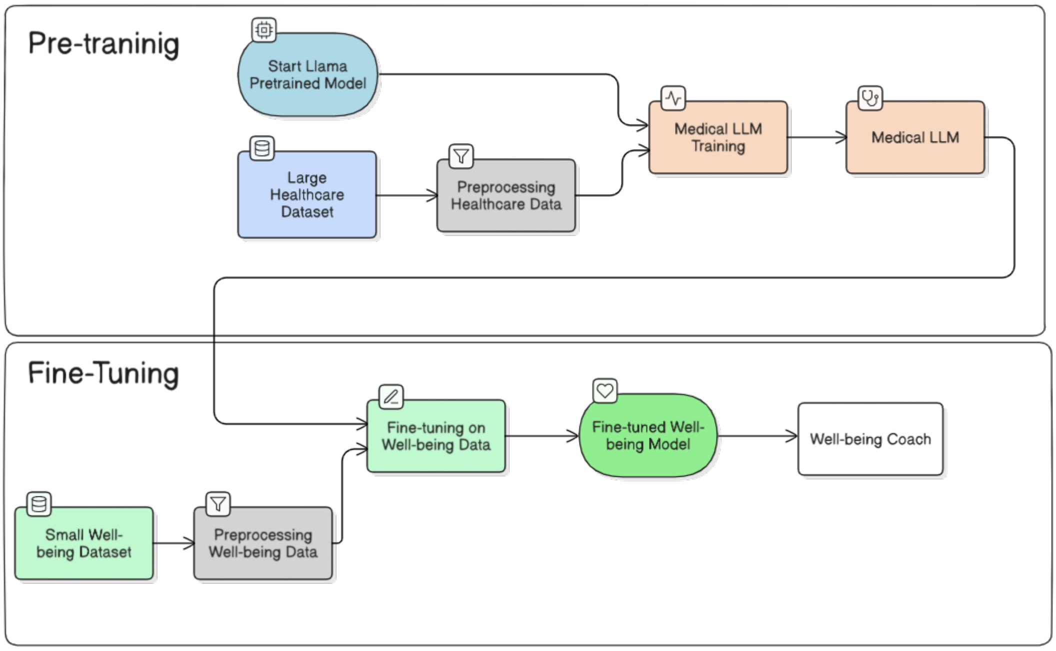

Pre-traninig
Start Llama Pretrained Model
Large Healthcare Dataset
Preprocessing Healthcare Data
Medical LLM Training
Medical LLM
Fine-Tuning
Fine-tuning on Well-being Data
Fine-tuned Well-being Model
Well-being Coach
Small Well-being Dataset
Preprocessing Well-being Data

Figure 5.1: Model architecture and fine-tuning pipeline

5.1 Model architecture and training setup

5.1.1 Base Model Selection

The first step towards fine-tuning begins with the selected base model. As discussed in Section 3 (Base Model Selection), we selected Contact Doctor, a pretrained medical-domain LLM, because of its strong foundational knowledge in health and wellness.

5.1.2 Model Architecture

The first step towards building this well-being coach begins with loading the base model. So for that, we used the Hugging Face Transformer library, which provided access to our base pre-trained medical LLM (Contact Doctor). The base model was loaded along with its tokenizers to handle the input text processing. On top of that, we can either create or load an existing LoRA adapter layer to capture domain-specific knowledge. It allows an architecture to combine the general medical knowledge from the base model along with the new task-specific refinements introduced through the adapter.

```
from transformers import AutoTokenizer, AutoModelForCausalLM
from transformers import DataCollatorForLanguageModeling
from peft import PeftModel, LoraConfig, get_peft_model
from perft import prepare_model_for_kbit_training
from datasets import load_dataset
import torch
from transformers import BitsAndBytesConfig,  Trainer, TrainingArguments,

# --- Paths ---
base_model_id = "ContactDoctor/Bio-Medical-Llama-3-8B" #path to base LLM
adapter_path = "./adpter_path"  # previously trained LoRA adapter
)
```

The new Lora adapter can be created as following:

```
lora_config = LoraConfig(
    lora_alpha=16,
    lora_dropout=0.1,
    target_modules=["q_proj", "v_proj"],
    r=64,
    bias="none",
    task_type="CAUSAL_LM",
)

# Creating a PEFT model - Attaching PEFT configs to the base_model
base_model = prepare_model_for_kbit_training(base_model)
base_model = get_peft_model(base_model, lora_config)
```

Our well-being coach model is built using transformer-based architecture, which nowadays is commonly and widely used for NLP and generation. Transformers rely on multi-head self-attention mechanisms and feed-forward neural networks to process sequences of text, allowing the model to capture complex relationships between word and context information over long spans.

Transformers are made up of multiple layers that include multi-head self-attention mechanisms, feed-forward networks, positional encodings, and residual connections with layer normalization, enabling the model to understand complex relationships between tokens, sequential data, and generate new output.

```
bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_use_double_quant=True,
    bnb_4bit_compute_dtype=torch.bfloat16,
)

model = AutoModelForCausalLM.from_pretrained(
    base_model_id,
    quantization_config=bnb_config,
    device_map="auto",
    torch_dtype=torch.bfloat16,
)
```

In the above provided code snippet, **AutoModelForCausalLM.from_pretrained** loads the pre-trained transformer model with causal language modeling capabilities, allowing it to predict the next token in a sequence given previous context.

The **quantization_config** argument specifies that the model should be loaded in 4-bit precision; this technique is useful for optimizing memory usage while preserving performance.

device_map="auto" is used for automatic device selection depending on availability.

torch_dtype=torch.bfloat16 enables mixed-precision computations for faster training and lower memory consumption.

The below piece of code below freezes most of the base model's parameters, and updates the parameters of the adapter layer. This process ensures that the base model preserves its pre-trained knowledge and just learn the specific task related to our wellbeing domain, also, this helps reduce computational requirements and prevent catastrophic forgetting.

```
model = prepare_model_for_kbit_training(model)
model = PeftModel.from_pretrained(model, adapter_path)
)
```

The next step is to prepare a training dataset. Preparation of datasets is and will always be the most important part of fine-tuning a machine learning model. The dataset can be in various formats, so we have to reform it before using it for training.

```
dataset = load_dataset("./well-being-dataste", split="train")
dataset = dataset.train_test_split(test_size=0.2, seed=42)
train_dataset = dataset['train']
val_dataset = dataset['test']
```

```
def format_prompt(example):
    return {
        "text": f"<|user|>\n{example['input']}\n<|assistant|>\n{example['output']}"
    }

train_dataset = train_dataset.map(format_prompt)
val_dataset = val_dataset.map(format_prompt)
)
```

After loading the dataset for fine-tuning, the next important step is careful selection of the hyperparameters to ensure effective training. This includes choosing an appropriate learning rate, batch size (based on available memory), and number of epochs. Selecting too many epochs could lead the model to overfitting, while too few could result in underfitting, preventing the model from adequately learning the task-specific patterns.

```
# --- Training arguments ---
training_args = TrainingArguments(
    output_dir="./finetuned-model",
    per_device_train_batch_size=2,
    per_device_eval_batch_size=2,
    num_train_epochs=5,
    learning_rate=2e-4,
    bf16=True,
    logging_dir="./logs",
    save_strategy="epoch",
    save_total_limit=2,
    gradient_checkpointing=True,
    report_to="none",
)

# --- Trainer ---
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=val_dataset,
    tokenizer=tokenizer,
    data_collator=data_collator
)
```

5.2 Experimental Setup and System Requirements

5.2.1 Hardware Environment

The whole experiment, including model loading, dataset preprocessing, fine-tuning, and evaluation everything was carried out using Google Colab's cloud-based environment. All the specifications of the hardware environment are mentioned in table 5.1

Component	**Specification**
Platform	Google Colab (Pro)
GPU	NVIDIA L4 (22.5 GB VRAM)
CPU	2 vCPUs
System RAM	53 GB
Disk Storage	~112.6 GB

Table 5.1: System Specifications

5.2.2 Software Environment

The following tables 5.2 show all the software environments that were needed to complete the hardware setup.

Component	**Specification / Version**
Operating System	Ubuntu 20.04 LTS (Colab backend)
Python	3.10
PyTorch	2.1.0 (with CUDA 12.1 support)
Transformers	4.36.0
PEFT	0.6.0
Datasets	2.14.0
BitsAndBytes	0.41.1
Accelerate	0.23.0
CUDA Toolkit	12.1 (Google Colab GPU runtime)

Table 5.2: Software Environment Specifications

5.2.3 Tools and frameworks used

Hugging Face Transformers provides pre-trained open source models, tokenizers, and other training utilities required for NLP. Hugging Face was used to load the base model and store the fine-tuned models and adapters.

PEFT (Parameter-Efficient Fine-Tuning) was used as an adaptation technique that supports LoRA, and which updates only a part of model parameters, freezing the rest. This reduces memory and computational load during fine-tuning.

Datasets Library (Hugging Face) Again used for storing, splitting, and processing the dataset used for fine-tuning.

BitsAndBytes Provides quantization support (4-bit NF4) for large models, enabling lower GPU memory consumption while maintaining accuracy.

PyTorch The primary deep learning framework used for model operations, training, and integration with the above tools.

Training Utilities Includes `Trainer` and `TrainingArguments` from Hugging Face for handling the fine-tuning process, gradient checkpointing, logging, and model saving.

Python Libraries General-purpose Python libraries (e.g., NumPy, pandas) were used for data manipulation, preprocessing, and weekly summarization of user activity and dietary information.

5.2.4 Justification on resources used

The mentioned hardware and software resources were selected because they provide the perfect balance between performance, cost, and accessibility. Google Colab is a cloud-based service that provides cloud-based GPUs, enabling efficient model training without the need for expensive hardware. The NVIDIA L4 GPU, with 22.5 GB of VRAM, was sufficient for executing and fine-tuning the quantized LLM model.

Because the base model was around a 16GB model, optimization was important to reduce the computational requirements. The techniques, such as LoRA, allowed updating only a small subset of parameters during fine-tuning, which significantly reduced memory usage and training time. use of 4-bit quantization further minimized the GPU memory usage, making it possible to fine-tune the model further on a single GPU without compromising the performance.

Chapter 6

Evaluation

Evaluation of the model is one of the important steps towards the efficiency and the reliability of the machine learning system, specifically when we talk about domain-specific applications such as a well-being coach. Since the model has been fine-tuned on a curated dataset and structured templates, it is essential to rigorously evaluate how well it performs in practical applications. This includes testing accuracy, correctness, and clarity it provides to the users.

In this chapter, we talk about the techniques and methods we used to evaluate the performance of our well-being coach. First, we start with assessing the model's ability to generate accurate and relevant guidance in tasks such as food, fitness, social, and mental lifestyle management. Then we examine if the model is able to provide a structured response template, ensuring the consistency of output so it can be easily understood by the user. Later, we evaluate the personalization capability of the model, checking whether it is appropriately adapting its advice based on individual user profiles, including factors such as age, gender, dietary preferences, fitness activity, mood and sleep patterns.

6.1 Model Comparison (Base Model vs Fine Tuned Model)

In this section, we compare the evaluation of the base model and our well-being coach, which is further fine-tuned on the base medical LLM. This step is must for evaluating the accuracy and learning of the model for our domain. And to understand the impact of fine-tuning, we had to compare the base model with our fine-tuned well-being coach. The goal is to evaluate how fine-tuning improved the ability of the model towards personalized responses, and how well it can follow the structure and domain-specific recommendations for food, fitness, mental, and social health, considering individual unique input data.

6.1.1 Quantitative Comparison

The performances of both the models were evaluated using various metrics such as Sematic search, BERTScore, and ROUGE metrics, considering different tasks the well-being coach must perform, considering the input data collected from the fitness band.

The base model had strong general medical knowledge, but it struggled to provide personalized recommendations. It was not able to identify the contradictions in its generated output. For example, if the user's preferred meal was vegetarian, then the model should have understanding of not to recommend the meal containing meat, but it was having difficulty and was still recommending meals with meat. Although the model had good general knowledge about vegetarianism, it was not able to use its knowledge to provide personalized responses considering users' input data. This highlights a key limitation and its inability of the responses.

The fine-tuned model consistently outperformed the base model across all metrics. It outperformed the base model when we provided the data of FitBit, indicating that the fine-tuned had made the model adapt the structured input-output templates and user-specific data.

Metric	**Contact Doctor (Base)**	**Well-Being Coach (Fine-Tuned)**
Semantic Similarity (Sentence-BERT cosine)	0.0834	0.7678
BERTScore F1	0.7693	0.8612
BLEU	0.0000	0.1574
ROUGE-1	0.0029	0.5329
ROUGE-2	0.0000	0.2099
ROUGE-L	0.0029	0.2526

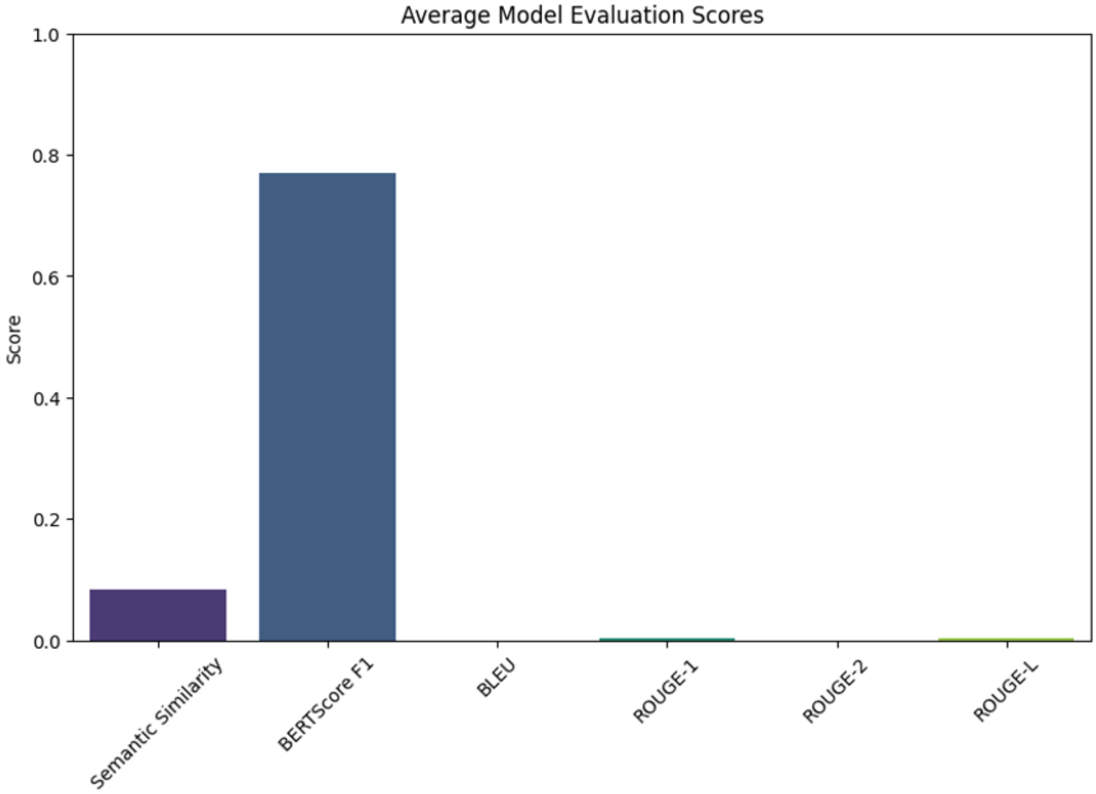
Table 6.1: Evaluation Results: Base Model vs Fine-Tuned Well-Being Coach

Table 6.1 clearly shows the evaluation between the base model(Contact Doctor) and our fine-tuned (Well-being coach) across different metrics. The table highlights the improvements it has achieved after fine-tuning. The Table shows the evaluation of personalized output generated for the same user data by different models.

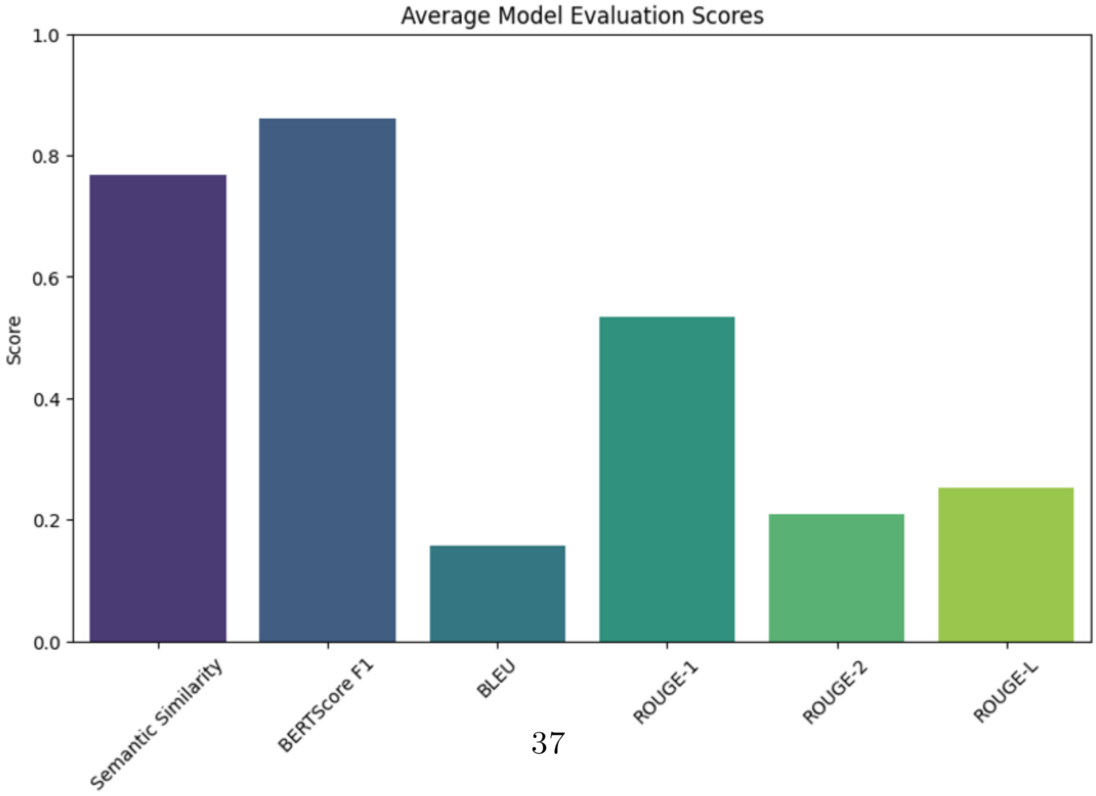
The Semantic Similarity(Sentence-Bert cosine) score increased dramatically from 0.0834 to 0.7678 for the fine-tuned model. This indicates that the fine-tuned model is able to generate responses that are much closer in meaning to the expected outputs. Similarly, the BERTScore F1 has also improved from 0.7693 to 0.8612, showing improved alignment between the generated response and reference response from the test dataset in terms of semantic and lexical similarities.

Metrics commonly used in NLP evaluation, such as BLEU, ROUGE-1, ROUGE-2, and ROUGE-L, also show substantial improvement. As BLEU, which was 0.0000 for the base model and increased to 0.1574, showing that the fine-tuned model better replicates the reference n-gram, showing that it is following the structure. ROUGE metrics, which measure overlap of unigrams, bigrams, and longest common subsequences between generated and reference texts, improved significantly. As we can refer to the table and visualize that ROUGE-1 increased from 0.0029 to 0.5329 and ROUGE-2 moved from 0.0000 to 0.2099, and ROUGE-L from 0.0029 to 0.0526.

Overall, these results indicate that the fine-tuned well-being coach not only preserves the medical domain knowledge from the base model, but it also learns to structure and infer personalized recommendations following the defined template. The model demonstrates better coherence, relevance, and semantic alignment with the expected outputs, validating the effectiveness of the fine-tuning approach.

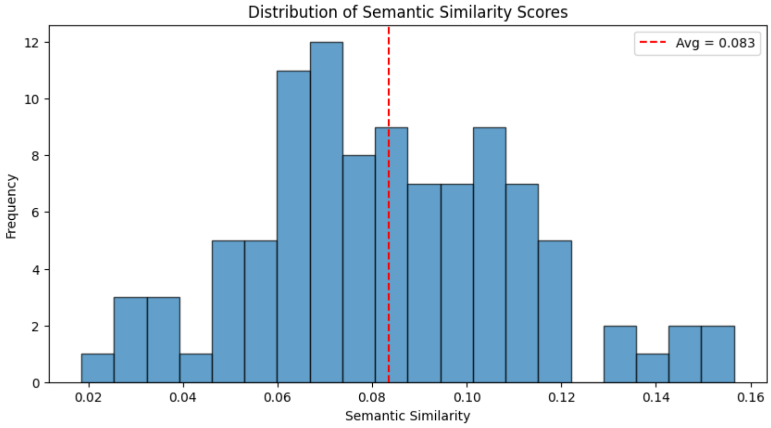


(a) Plot 1: Contact Doctor

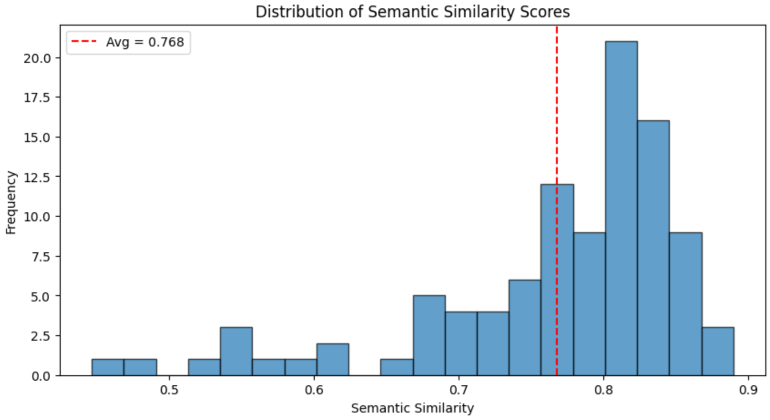


(b) Plot 2: Well-being coach

Figure 6.1: Quantitative Evaluation Comparison of both models

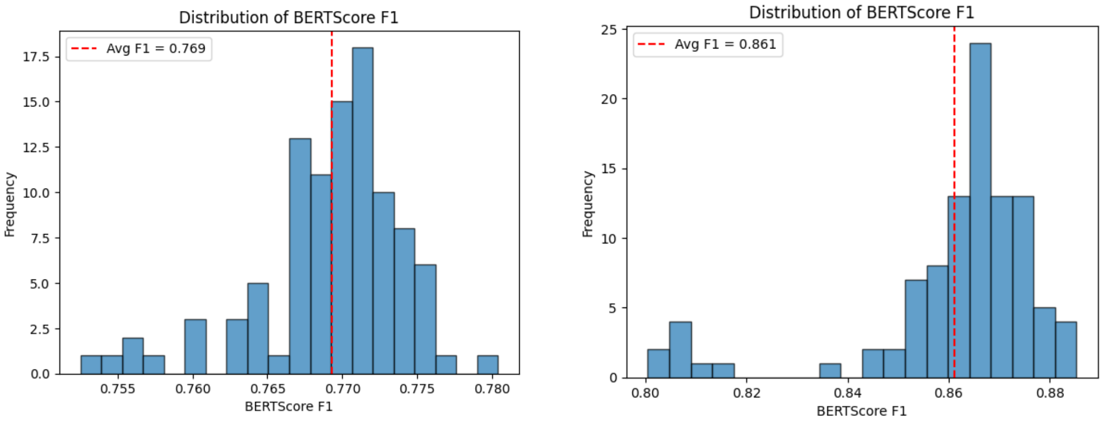


(a) Contact Doctor



(b) Well Being Coach

Figure 6.2: Semantic Similarity comparison



(a) Contact Doctor

(b) Well Being Coach

Figure 6.3: BERTScore comparison

Task based evaluation

Table 6.2: Comparison of Contact Doctor and Fine-Tuned Well-Being Coach Across Domains

Metric	**Contact Doctor**			**Well-Being Coach**		
	Food	**Fitness**	**Mental Health**	**Food**	**Fitness**	**Mental Health**
Semantic Search	0.6435	0.5020	0.6805	0.7657	0.6824	0.6585
BERTScore	0.8058	0.8249	0.8677	0.8523	0.8469	0.8698
BLEU	0.0000	0.0114	0.0296	0.0000	0.0289	0.0307
ROUGE-1	0.1736	0.1638	0.2505	0.1472	0.2892	0.2822
ROUGE-2	0.0058	0.0248	0.0604	0.0033	0.0695	0.0773
ROUGE-L	0.1712	0.1130	0.1567	0.1431	0.1701	0.1819

Table 6.2 shows the comparison between the performance of the base and the fine-tuned model across three key aspects of the well-being domain:- Food, fitness, and mental health. Various evaluation metrics have been used to assess the model's accuracy, relevance, and its ability to generate the appropriate responses.

From the table, it is evident that the fine-tuned model consistently outperforms the base model in most of the domains, preserving its base knowledge without catastrophic forgetting.

Overall, the table demonstrates that the fine-tuned model has improved its performance over the base model in all key domains. This shows that the well-being domain is suited to provide personalized, actionable, and context-aware recommendations compared to the base model.

6.1.2 Prompt based evaluation

Now, to further evaluate the model, on top of quantitative evaluation with automatic metrics, a prompt-based evaluation was conducted to see the model's performance in realistic scenarios. For this, both the Contact Doctor and the Well-Being Coach were tested by providing the same user data prompt along with a template used across all evaluations, considering four main pillars of well-being: Food, Physical Activity, Sleep, Mood and Lifestyle, and Overall Weekly Summary and Goal.

So initial evaluation of the base model(Contact Doctor) revealed significant limitations following the structured template for the output. As we show, even though the model has great general knowledge, it could not provide a response based on the user's demographic data. And this limitation highlighted the need for the fine-tuning of the model to align with the structured template.

So after fine-tuning, the results demonstrated a noticeable improvement. Now the Well-Being Coach can provide structured and clear recommendations, advising actionable advice for each aspect of well-being.

Examples of prompt-response pairs confirmed that the fine-tuned model not only followed the template but also provided individualized recommendations aligned with user activity, dietary habits, and sleep patterns. A few such examples are mentioned below, which show the response generated by our well-being coach by providing user data input.

INSTRUCTION: You are a well-being coach. Based on the provided user data, generate personalized well-being recommendations. Follow the exact template provided below and do not deviate from its structure.

1) Food Recommendation
- Overall Assessment: []
- Areas of Improvements: []
- Suggested Meals: []

2) Physical Activity
- Activity Assessment: []
- Zone Minutes and intensity Feedback
- Strength/Cardio Tips: []
- Weekly Goals: []

3) Sleep, Mood, Mental Health, and Lifestyle
- Sleep Review: []
- Suggestions: []

4) Mood and Lifestyle
- Mood Analysis: []
- Suggestions : []

5) Weekly Summary
- Summary: []
- Goals: []

Input (User Data)

Demographics: Gender(F), Age(50), Height(164.2 cm), Weight(57.4 kg), BMI(21.2)
Physical Activity: Weekly Steps(38539), Zone Minutes(150), Exercise Sessions(2)
Sleep: Average hours per night(6.8)
Dietary Preference: Vegetarian, Allergies(eggs)
Weekly Food Consumption: Dairy(2.6 L), Legumes(531 g), Meat(1490 g), Fruits(639 g), Vegetables(1276 g), Grains(499 g), Nuts(93 g), Water(8.0 L)

Output (Food Recommendations)

Food Recommendation:
- Overall Assessment: Your current weekly food consumption shows a high intake of meat (1490 g), which is inconsistent with your vegetarian dietary preference. Additionally, while you consume an adequate amount of fruits (639 g) and vegetables (1276 g), legumes (531 g) could be increased to better align with your nutritional needs as a vegetarian.

- Areas of Improvement: - Reduce meat intake significantly or eliminate it entirely to adhere to your vegetarian diet.
- Increase legume consumption to at least 700 g per week for additional protein and fiber.
- Aim for around 800 g of grains per week to ensure sufficient energy levels.

- Suggested Meals: 1. Chickpea Salad Bowl: Combine chickpeas (200 g), mixed greens, cherry tomatoes, cucumber, avocado, and a lemon-tahini dressing. This meal provides plenty of plant-based proteins and healthy fats.
2. Quinoa Vegetable Stir-fry: Cook quinoa (100 g) and stir-fry with bell peppers, broccoli, carrots, and tofu. Serve over brown rice (100 g). This dish offers a balanced mix of carbohydrates and nutrients.
3. Lentil Soup: Prepare lentils (300 g) in vegetable broth with spinach, carrots, celery, and spices. Pair this hearty soup with whole-grain bread for added fiber.

Output (Fitness Recommendations)

Physical Activity:
- **Activity Assessment:** You have been quite active with a total of 38,539 steps per week, indicating that you're doing great in terms of daily activity. However, only two exercise sessions suggest there's room for improvement in structured workouts that can enhance strength and cardiovascular fitness.

- **Zones Minutes and Intensity Feedback:** Currently, you're achieving about 150 zone minutes per week, which indicates moderate-intensity exercise but may need some increase for optimal health benefits given your age and weight management goals. Consider aiming for 210 minutes by incorporating more vigorous activities like cycling or swimming.

- **Strength/Cardio Tips: - Bodyweight exercises:** Incorporate squats (15 reps x 3 sets), push-ups (10 reps x 3 sets), and lunges (20 reps x 3 sets) twice a week to improve muscle strength and endurance.
- **Cardiovascular workouts:** Add brisk walking (30 minutes, 3 times a week) or cycle (45 minutes, 2 times a week) to boost heart rate and overall fitness.

- **Weekly Goals:** 1. Step Target: Maintain or slightly increase step count to reach 40,000 steps per week.
2. Exercise Sessions: Increase to three exercise sessions per week focusing on both strength and cardio.
3. Zone Minutes: Aim for 210 minutes of moderate-intensity exercise each week.

Output (Sleep Recommendations)

Sleep:

- Sleep Review: You average about 6.8 hours of sleep per night, which is just above the recommended minimum but still suggests potential fatigue. Given your lifestyle, ensuring consistent quality sleep will greatly benefit your overall well-being.

- Sleep Suggestions: 1. Establish a routine: Go to bed and wake up at the same time every day to regulate your body's internal clock.
2. Relaxation techniques: Practice gentle stretches before bedtime and consider reading a book instead of screens to wind down.
3. Comfortable environment: Ensure your bedroom is dark, quiet, and cool; invest in a comfortable mattress if needed.

Output (Mood and Lifestyle Recommendations)

Mood & Mental Health:

- Mood Review: While you maintain a positive outlook generally, your reliance on meat may affect your alignment with your values as a vegetarian, potentially leading to feelings of inconsistency. It's essential to focus on nourishing yourself according to your preferences to support emotional well-being.

- Mental Health Suggestions: 1. Mindfulness practices: Set aside 10–15 minutes daily for meditation, deep breathing, or yoga to reduce stress and increase self-awareness.
2. Connect with community: Engage in social activities centered around vegetarianism or wellness groups to build supportive relationships and share knowledge.
3. Reflective journaling: Write down gratitudes and reflect on personal growth regularly to stay motivated and focused on long-term goals.

Output (Weekly overview)

Weekly Overview:
- Weekly Summary:
- Summary: Your current weekly food consumption reflects a good balance of fruits and vegetables, but a significant excess in meat intake contradicts your vegetarian preference. Increasing legumes and grains would help meet protein and energy needs effectively, while maintaining physical activity through a combination of strength training and cardio will enhance your overall fitness and mental clarity.
- Goals: 1. Reducing meat intake to align with vegetarian choices (ideally eliminating).
2. Increasing legumes to at least 700 g per week for improved protein sources.
3. Maintaining a target of 800 g of grains per week for sustained energy.
4. Achieving 210 minutes of moderate-intensity

6.2 LLM as a Judge (G-EVAL)

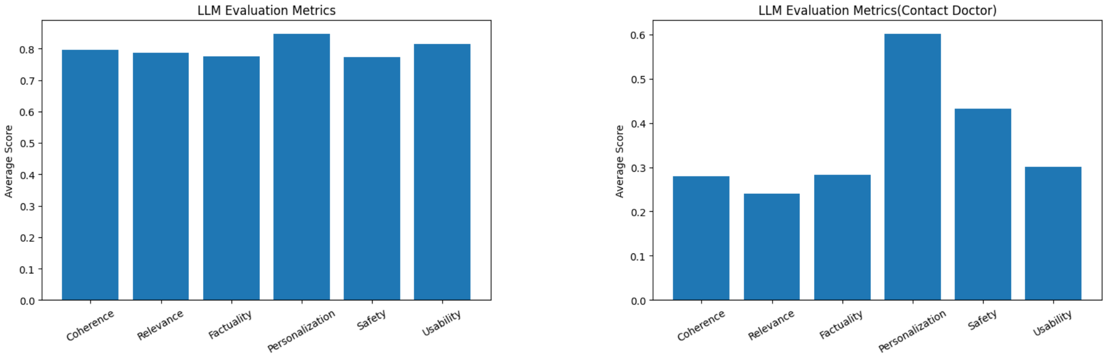
Evaluating the quality of texts generated by natural language generation (NLG) systems is difficult, even with the rise of powerful large language models (LLMs) that produce human-like text. Traditional evaluation methods like BLEU, ROUGE, and METEOR are common but often don't align well with human judgment, particularly for creative or open-ended tasks. A major drawback of these metrics is their need for a human-written "reference" text to compare against, which is expensive and time-consuming to obtain.

To address this, researchers have started using LLMs themselves as a new way to evaluate generated text, without needing a reference. The thinking is that LLMs can score the quality of a text based on how likely it is to be generated, with the assumption that high-quality, fluent text will have a higher probability. It is evaluated with three main components
1) prompt: define task and evaluation criteria
2) a chain-of-thoughts (CoT): detailed step by step instructions genertaed my LLMS and
3)Scoring function: scores the output based on the calculated score of the model's output.

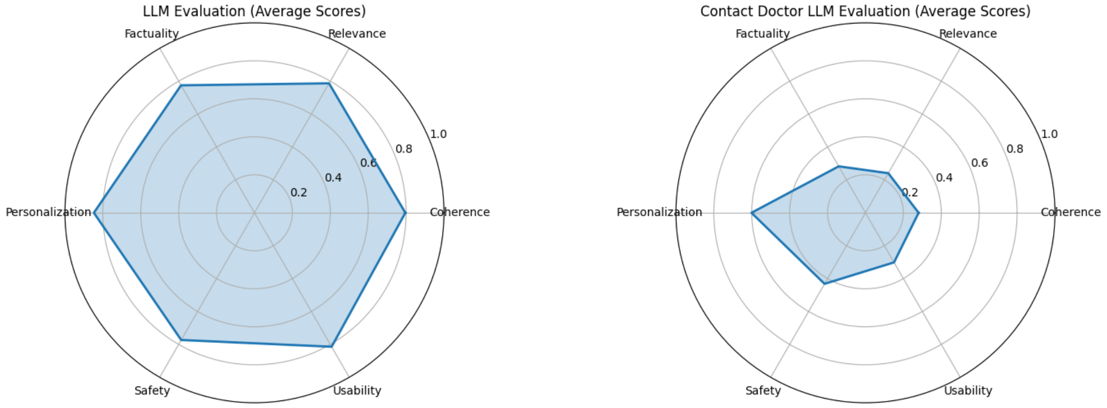
6.2.1 Results Visualizations

From the figure 6.4, we can see that the evaluation of a well-being coach using G-EVAL shows consistently strong performance across various dimensions compared to Contact Doctor. The model achieves high coherence (0.79), meaning the responses generated by it are logically structured with smooth flow. Relevance also scores well (0.78), indicating that the recommendations directly address user inputs without drifting off-topic. Factuality is slightly lower at 0.77 but still robust, showing that most outputs are grounded in accurate health and wellness knowledge with only minor gaps or inconsistencies. The strongest result is in personalization (0.85), demonstrating the model's ability to adapt recommendations to individual user data such as diet, activity levels, and lifestyle factors. Safety (0.77) is also satisfactory, suggesting the outputs avoid harmful or misleading advice and remain within reasonable health guidance. Finally, usability (0.81) confirms that the responses are not only understandable but also practical and actionable for users.

Taken together, these scores indicate that the fine-tuned Well-Being Coach provides reliable, user-centered, and safe recommendations, with particular strengths in personalization and clarity of presentation. The slightly lower factuality and safety scores highlight areas where future refinements, such as stricter grounding in medical guidelines could further enhance trustworthiness.



(a) G-EVAL Plot Well Being (b) G-EVAL Plot Contact Doctor



(c) G-EVAL Web Well Being (d) G-EVAL Web Contact Doctor

Figure 6.4: G-EVAL results of the well-being coach

Chapter 7

Summary

7.1 Motivation and Problem Statement

Nowadays, wearable devices such as fitness bands or fitness tracking applications have shown increasing use in the health monitoring aspects. While these devices provide rich streams of data, transforming the information into actionable personalized recommendations remains significantly challenging. As LLMs have shown incredible capabilities in various domains, specifically in health, there are no personalized wellness coaches. The motivation behind this thesis was to bridge this gap by developing a Well-being Coach LLM that can leverage wearable devices data into generating personalized recommendations considering four main pillars of well-being: food, fitness, sleep, mood and mental well-being.

7.2 Approach Summary

For this work, we started out by first selecting the base model to design Well-Being Coach, with the aim of transforming wearable data into practical, personalized health recommendations. The process began by fine-tuning the base model on medical and diagnostic datasets to ensure it had a strong grounding in health-related knowledge. To make the model truly personal and context-aware, we then created a custom dataset that combined Fitbit wearable data with carefully selected health guidelines, nutrition books, and lifestyle resources. This dataset was formatted into question–answer style entries and structured recommendations so that the model could learn not just facts, but also how to deliver them in an actionable way.

The final system was able to take multimodal inputs, such as daily steps, heart rate, dietary patterns, and sleep data, and could produce output in a clear four-part template. Each response included the recommendation related to food, fitness, sleep, mood & mental wellness, and overall weekly summary. For the evaluation of the model performance, we used various metrics along with G-EVAL, an LLM-based evaluator. This technique allowed us to not only measure accuracy but also more human-centred qualities such as coherence, relevance, personalization, safety, and usability.

7.3 Evaluation and Key Findings

To understand how well the well-being coach performs after the fine-tuning. It was necessary to compare it against the base model(Contact doctor) medical LLM. To do so, we used both quantitative and human-centered evaluations. The results clearly represented that the base model struggles to follow the structured template and often hallucinates or generates incomplete or irrelevant responses. On the other hand, the fine-tuned model showed coherence, context-awareness, and personalized recommendations.

While conventional metrics such as BLEU, ROUGE, and BERTSCORE verified that the fine-tuned model generated outputs such that, they were more reliable and personalised. However, the real value of the evaluation came from the G-EVAL, which provided the potential to deeper reflections on models' quality like coherence, relevance, personalisation, safety, and usability. On average, the Well-Being Coach scored well across all of these measures, with some notable results in personalisation (0.84) and usability (0.81).

One major finding and contribution of our study is that the model consistently outperformed the base model in handling complex user input. For instance, it was able to provide vegetarian-friendly meal plans, along with workable modifications to step count and sleep habits. At the same time, the safety scores indicated that the model was generally safe and that it was very unlikely to suggest harmful or medically inappropriate recommendations, which is a crucial requirement for any AI system in healthcare.

All these evaluations suggest that the fine-tuned Well-Being Coach is technically not only more accurate than the base model, but in addition, offers much more relevant, secure, and actionable well-being advice for real users.

7.4 Contribution

The field of personalized health monitoring and the wider application of large language models (LLMs) in well-being applications both benefit greatly from the significant contributions made by this thesis. The first significant contribution is the development of a refined version of the Meta LLaMA-3 model called the Personal Well-Being Coach. The refined model consistently produced individualized, safe, and actionable recommendations across domains like diet, physical activity, sleep, and lifestyle, in contrast to the base model, which frequently had trouble following directions or producing structured, pertinent outputs. The model was able to go beyond general recommendations and instead offer customized suggestions that took into account each person's unique needs by integrating user health data from wearable technology.

The process of developing the dataset is the second contribution. Real-world wearable data, mostly from Fitbit, was combined with information gleaned from scientific sources, wellness books, and medical guidelines to create a meticulously curated dataset. The model was able to contextualist wearable signals (such as heart rate, sleep patterns, and steps) within evidence-based frameworks for health recommendations thanks to this hybrid dataset. The dataset also improved the model's capacity to consistently respond to a wide range of user queries by incorporating structured Q&A pairs from domain literature. This work offers future researchers interested in improving LLMs for healthcare a reusable foundation in addition to a resource for this thesis.

Third, this work offers assessment techniques that surpass conventional metrics. A more human-centered evaluation along dimensions like coherence, relevance, personalisation, safety, and usability was made possible by the adoption of G-Eval, even though standard measures like BLEU or ROUGE were still used. These findings demonstrated the strengths and weaknesses of the refined model, providing a guide for more comprehensive evaluations of LLMs with a health focus.

Lastly, the thesis shows how LLMs can be used practically for individualized well-being. This work demonstrates the potential for AI-powered coaches to serve as approachable, scalable partners for people looking to improve their health by effectively mapping wearable data to structured health recommendations. The Personal Well-Being Coach is an example of how LLMs can offer early-stage insights, motivation, and customized lifestyle recommendations, but it is not a substitute for expert medical advice.

Collectively, these contributions enhance our knowledge of how LLMs can be modified and assessed for contexts involving personalized health, and they offer tangible resources, such as evaluation rubrics and datasets, that can hasten further advancement in this new field.

7.5 Limitations

There are several important limitations that must be acknowledged, although this thesis demonstrates how large language models (LLMs) can be used to create individualised well-being coaching for clients. Since the recommendations by the model were built on the data from the fitness band and medical knowledge, they should not replace instead of medical advice from qualified experts. As the models provided recommendations are solely based on the pattern learning from the training data, it can be misleading or unique to each specific patient, which a medical expert would consider.

Second, although it was thoughtfully and extensively selected, the fine-tuning dataset is relatively small compared to the huge corpora used to train base LLMs. In particular, the wearable and health guidelines data consisted of approximately 1,500-2000 synthetic input-output pairs. The dataset does not account for all possible user scenarios, dietary preferences, or unusual medical conditions; however, it is sufficient to train the model to adhere to structured templates and make sensible recommendations. As expected, the model may perform differently when exposed to edge cases or unexpected user inputs.

Third, despite the systematic evaluation, it relies on automated metrics G-Eval, BERTScore, ROUGE, and BLEU. Although the above metrics encompass coherence, relevance, and factuality, they are not able to capture the more subjective components of user experience, such as motivation, trust and adherence to recommendations. Validating these aspects would require testing in the field with real users with varied user populations.

Finally, decisions about model selection and tuning were constrained by computational limits. We were only able to look at potentially better architectures we excluded some models with more domain knowledge (for example, MedGemma) due to their computational demands. Also, although quantisation and LoRA adapters allowed fine-tuning on computationally limited hardware, they possibly had a negligible impact on the accuracy of some recommendations.

In conclusion, even though the Personal Well-Being Coach is a big step towards AI-powered personalised health advice, its results should be seen as helpful recommendations rather than firm medical orders. To improve reliability and applicability, future research should concentrate on growing datasets, integrating more complex multimodal sensor data, and incorporating actual user feedback.

7.6 Future Work and Recommendations

Based on the findings and contributions from this thesis, there are a number of opportunities for future research and future improvements. One area of opportunity involves the expansion and diversification of the training dataset. Especially because while the handcrafted dataset allowed the model to follow structured templates (as developed in Chapter 5), using larger-scale, real-world wearable, and broader population-related data would make the models better at providing more robust and personalized recommendations for participants regarding gear and exercise. This could include training more on the extent of data available from other age groups, cultures/populations, and medical and activity levels.

Another key extension might be the use of multimodal data sources, rather than simply adding multiple wearable metrics. An example of this would be the addition of continuous heart rate, sleep phases, stress levels, and diet to contextual data like calendar events, location, or weather, which theoretically would allow the model to offer decision support with better specificity and awareness of context. If we could also bring in real-time monitoring and adaptive feedback loops, the coach would be able to react to a variety of changes in the user's behavior or health status as they occur.

In addition, better evaluation protocols are important in safeguarding the safety, usability, and effectiveness of health guidance produced by AI. Future research could involve user studies to assess how individuals interact with the model, whether their recommendations are able to be followed, and the change in adherence over time. Ideally, combining evaluation metrics that are automated, like G-Eval, with evaluation metrics defined by users with human-in-the-loop processes would provide a more complete picture of model performance.

Finally, exploring larger or more specialized base LLMs could enhance the depth and accuracy of domain knowledge, especially for medical and nutritional guidance. Techniques such as incremental fine-tuning, active learning, and reinforcement learning from human feedback (RLHF) may further improve model performance while maintaining efficiency.

In conclusion, while the Personal Well-Being Coach developed in this thesis demonstrates the potential of LLMs for personalized health guidance, the future holds exciting opportunities to improve scalability, reliability, and user-centricity. By leveraging more comprehensive datasets, multimodal inputs, and continuous evaluation, next-generation AI well-being coaches can become even more effective, trustworthy, and widely applicable tools for promoting health and lifestyle optimization.

Appendix

Appendix A

Sample Prompts used for evaluation

Health and nutritional fact base prompts

- What to eat before and after workout? [25]
- How does a person's diet affect their cardiovascular health?[26]
- What foods should I be integrating into my diet to improve my cardiovascular health?[26]
- What food should i avoid to keep my heart healthy?[26]
- Can mediterranean diet really improve my cardiovascular health?[27]
- How many calories should an average adult female consume per day?[28]
- I do intermittent fasting (16:8), sleep 8 hours, and do yoga. Am I missing any nutrients?
- If a person is running for 15 km everyday and his BMI is 20. So can you make a diet plan for his physical health?
- Diet can have a positive effect on a sporting performance. Explain why a diet high in carbohydrates could lead to an improvement in a sportsperson's performance.

Fitness and General Lifestyle

- Exercise is only helpful if you do 30 minutes every day.[29]
- Daily activities such as vacuuming, mopping, and sweeping can help build up an individual's endurance.[29]

- Being physically active is expensive and you need a lot of money for equipment.
- Staying active is good for dealing with stress
- People with high blood pressure should not be physically active.
- Muscle weighs more than fat.
- Stretching is not important before and after you complete your workout.
- Exercises to strengthen your bones includes doing push-ups and climbing.
- Aerobic exercises include walking, running, or anything that makes your heart beat faster.
- Physical activity can increase your risk for developing type 2 diabetes.
- Exercise can relieve body aches such as headaches by releasing natural pain killers into your brain called endorphins.
- What time of day should I exercise?[29]

Age and Population specific guidelines

- How much physical activity does the children between age 3 to 5 needs?[15]
- How much physical activity does the Children and adolescents 6 to 17 needs?[15]
- How can my child meet the physical activity recommendations?[15]
- What kind of activities adult people above age 60 should perform for better health?[16]
- If an adult is suffering from chronic conditions and disabilities what kind of physical activities they should perform for their better well being?[16]

Stress and mood

- What are Stress Management and Resiliency Exercises?
- I'm feeling really stressed about work lately. What are some simple ways to relax?
- What are the common symptoms of anxiety?

- I've been feeling really overwhelmed with my job lately, and it's making me constantly on edge. What are some simple ways to de-stress after a long day?
- I'm feeling low and haven't exercised for a week. What should I do?
- How does fitness affect the mental health?
- How do i stay motivated?

Personalized Assessments Based on User Data

- I'm 42, walk 5,000 steps daily, eat vegetarian. How's my wellness?
- I'm a 25-year-old male, BMI 22, running 10 km daily but still feel fatigued. What can I do?
- I'm 40, female, sedentary job, walk 3,000 steps/day. I want to lose weight. Help?
- I'm a 30-year-old vegetarian male, sleep 6 hours/day, exercise 3x/week. Suggest a high-protein diet plan.
- I'm 42F, diabetic, eat 3 meals a day, mostly carbs. What should I change?
- I'm a 29-year-old woman sleeping only 4 hours/night due to stress. How is this affecting me?
- I sleep 9 hours but still feel tired and unproductive. I'm 36, male. Why?
- I'm 55, walk 7,000 steps daily, average heart rate is 88 bpm. Is this healthy?
- User Profile: - Age: 28 - Gender: Female - Height: 165 cm - Weight: 60 kg - Sleep: 6.5 hours - Mood: "Anxious" - Food Intake: Balanced meals, 2L water - Exercise: 45 min yoga - Steps: 8,500 - Resting Heart Rate: 78 bpm. Suggest personalized wellness recommendations for today, addressing mood, recovery, and physical performance. Keep it supportive and evidence- based.

Prompts with responses of selected LLMs

- **Prompt 1:** Staying active is good for dealing with stress.

  > **Contact Doctor:** Yes, staying active is an excellent way to manage and reduce stress. Exercise has been shown to decrease the production of stress hormones, improve your mood, boost your self-esteem, and even help you sleep better at night. Here are some reasons why

exercise can be so beneficial in reducing stress:
Reduces cortisol levels
Boosts endorphins
Distracts from worries
Provides a sense of accomplishment

Deep Seek: Yes, staying active is indeed good for managing and dealing with stress. Physical activity can help reduce stress in several ways.
It releases endorphins.
It reduces levels of stress hormones like cortisol and adrenaline.
It improves sleep quality.
It boosts self-esteem.
It distracts your mind from worrying thoughts.

MedAlpaca: Yes. Exercise releases hormones that make you feel happier and more relaxed. It also provides a chance for you to clear your mind and forget about problems for a while. The key is finding an exercise you enjoy so you will look forward to doing it.

BioMistral: Yes, this is true. Physical activity has been shown to reduce stress levels by releasing endorphins, which are natural mood boosters. Exercise also provides an outlet to redirect negative thoughts and helps to clear the mind. Additionally, regular exercise can serve as a preventive measure against anxiety and depression.

- **Prompt 2:** Aerobic exercises include walking, running, or anything that makes your heart beat faster

  Contact Doctor: .Yes, that's correct. Aerobic exercises are physical activities that use oxygen to generate energy and can be sustained over a relatively long period of time (usually 4 minutes or more). These types of exercises are designed to raise your heart rate and keep it elevated for an extended period of time, typically between 30 seconds and several minutes.
  Deep Seek True. Aerobic exercises are activities that raise the heart rate and respiratory rate above resting levels for a sustained period of time. They rely on aerobic metabolism to meet energy demands. The question stem states: "Aerobic exercises include walking, running, or anything that makes your heart beat faster?" This is true because these activities increase the heart rate and breathing rate through

aerobic (oxygen-dependent) metabolic pathways.

MedAlpaca: No. Aerobic exercise refers to any activity that requires the body to utilize oxygen to produce energy. This energy production occurs in the cells of the body, including the muscle cells, liver, and brain cells. Because muscular contractions occur during any type of exercise, all activities burn calories and are therefore useful for fat loss and improving body composition. Therefore, aerobic exercise does not refer to just cardiovascular exercise, but rather any activity that causes the body to break down glucose to create ATP.

BioMistral: Yes, this is correct. Aerobic exercises, sometimes called cardio exercises, are any type of physical activity that involves sustained or rhythmic movement of large muscle groups, such as brisk walking, jogging, running, cycling, swimming, or dancing. These types of exercises typically cause an increased heart rate and increased breathing and sweating rates.

- **Prompt 3:** What to eat before and after workout?

  Contact Doctor: What to eat before and after workout? Before a workout:
  Carbohydrates: Pasta, brown rice, whole wheat bread, fruits, and vegetables.
  Protein: Eggs, lean meats like chicken, fish, turkey, lentils, and legumes.
  Healthy fats: Nuts, seeds, avocados, and olive oil.
  Complex snacks: Apple slices with almond butter or banana with peanut butter.
  Hydrate with water or low-sugar sports drinks.
  After a workout:
  Replenish carbohydrates: Fresh fruits, energy bars, smoothies, or sports drinks.
  Protein: Greek yogurt, milk, protein shakes, hard-boiled eggs, or cottage cheese.
  Electrolytes: Bananas (potassium), dates (potassium), coconut water, or sports drinks.
  A balanced snack: Whole grain crackers with hummus, trail mix with nuts and dried fruits, or a small serving of granola.
  Stay hydrated: Drink water or low-sugar sports drinks within 30

minutes after exercise.

Deep Seek Before workouts:
Carbohydrates: To provide energy for exercise. Examples include whole grains, fruits, vegetables.
Protein: Helps preserve muscle tissue during intense activity. Fats: Provide sustained energy but avoid high-fat foods that can cause digestive issues.
After workouts:
Carbohydrates: Replenish glycogen stores used during training.
Protein: Crucial for repairing damaged muscles fibers.
Water: Hydrate to support metabolism and rehydration.

MedAlpaca: Before:
- Drink plenty of water. Hydration will help you perform your best during exercise and recover faster afterwards.
- Eat a light meal 2–3 hours before exercising. Avoid eating heavy, greasy foods, but choose something that will give you energy and won't weigh you down. Good options include fruit and yogurt, whole grain cereal and protein powder, or a smoothie.

After:
- Drink a glass of water immediately after exercising.
- Rehydrate with water or a sports drink if you exercised intensely enough to need electrolytes can replace fluids and minerals lost through sweat.
- Have a snack with protein powder can help recovery.
- Carbohydrate replacement
- Protein powder can help with your recovery
- Low-impact of water and muscle
- Muscle recovery
- You need to lose weight loss of muscle
- Lactic exercise can help your bodybuilding muscle and weight loss
- Low Impact loss. Building muscle isometric weight training helps you can also help impact

BioMistral: Before a workout, it is best to consume foods that are easy to digest and provide a quick source of energy, such as complex carbohydrates (whole grains, fruits, and vegetables) and lean protein sources (lean meat, fish, eggs, nuts, and seeds). Good pre-workout snacks include a piece of fruit or banana, oatmeal or granola bar, yogurt, or a small smoothie or juice.

- **Prompt 4:** User Profile: - Age: 28 - Gender: Female - Height: 165 cm - Weight: 60 kg - Sleep: 6.5 hours - Mood: “Anxious” - Food Intake: Balanced meals, 2L water - Exercise: 45 min yoga - Steps: 8,500 - Resting Heart Rate: 78 bpm Suggest personalized wellness recommendations for today, addressing mood, recovery, and physical performance. Keep it supportive and evidence-based.

  > **Contact Doctor:** Firstly, regarding your reported anxious mood, incorporating activities that promote relaxation and reduce stress could be beneficial. Have you considered starting your day with a gentle morning walk or some light stretching? Research suggests that spending time in nature and engaging in low-intensity exercise can help alleviate symptoms of anxiety. For recovery purposes, since you practice yoga, I recommend continuing with your usual routine. Yoga has been shown to improve both mental and physical well-being. If you’re feeling fatigued or experiencing any discomfort during your practice, however, consider modifying your poses or taking breaks to avoid exacerbating any existing tension. Lastly, considering your high step count, it’s essential to stay hydrated throughout the day. In addition to drinking 2 liters of water, ensure you’re consuming electrolyte-rich foods or beverages to support fluid balance and optimize physical performance.
  >
  > **BioMistral:** Continue to work on improving diet quality by increasing intake of fruits, vegetables, lean protein, and whole grains. Drink water instead of sweetened drinks. Continue to track food intake. Consider consulting with a registered dietitian to discuss ways to meet nutritional goals.

Appendex B

Fitbit Data for Training

User Input 1

Input:
Demographics: Gender(M), Age(34), Height(180.4 cm), Weight(76.4 kg), BMI(23.4)|
Physical Activity: Weekly Steps(42400), Zone Minutes per week(134), Exercise Sessions per week(3))
Sleep: Average hours per night(8.0)

Dietary Preference: Declared Diet(Vegetarian), Allergies(Lactose Intolerant)
Weekly Food Consumption: Dairy(4.9 L), Legumes(722 g), Meat(0 g), Fruits(316 g), Vegetables(575 g), Grains(2355 g), Nuts(41 g), Water(12.7 L)

Output:

1. **Food Recommendations**
    - **Overall Assessment:** Your current weekly food intake reflects a significant amount of grains and a moderate intake of legumes and vegetables. However, your dairy consumption is notably high, which may exacerbate your lactose intolerance, while your fruit and nut intake is quite low, indicating a potential deficiency in essential vitamins and healthy fats.
    - **Areas of Improvement:**
        - **Fruits:** Target 700 g weekly (100 g daily) to increase your intake of vitamins and antioxidants.
        - **Vegetables:** Maintain or slightly increase to 600 g weekly (about 85 g daily) to ensure you're getting a variety of nutrients.
        - **Legumes:** Aim for 1000 g weekly (about 140 g daily) to boost your protein and fiber intake while adhering to your vegetarian diet.
        - **Grains:** Reduce to 1500 g weekly (around 215 g daily) to avoid excess carbohydrate consumption.
        - **Nuts:** Increase to 200 g weekly (about 30 g daily) for healthy fats and protein.
        - **Dairy:** Minimize to 1 L weekly (about 150 mL daily) or consider lactose-free alternatives to prevent digestive issues.
    - **Suggested Meals:**
        - Chickpea Salad: Combine chickpeas, diced cucumber, cherry tomatoes, red onion, parsley, and a dressing of olive oil and lemon juice. This meal is rich in legumes and vegetables, providing protein and essential nutrients.
        - Vegetable Stir-Fry: Sauté a mix of bell peppers, broccoli, and snap peas in sesame oil, served over brown rice or quinoa, topped with a sprinkle of nuts for added texture and nutrition.
        - Smoothie Bowl: Blend spinach, banana, and a scoop of plant-based protein powder, topped with mixed berries, a tablespoon of almond butter, and a sprinkle of chia seeds to enhance your fruit and nut intake.

  - Lentil Soup: A hearty soup made with lentils, carrots, celery, and spices, served with a side of whole grain bread. This meal is an excellent source of protein and fiber while keeping you satisfied.

2. **Physical Activity Recommendations**

   - **Activity Assessment:** Your weekly step count of 42,400 steps indicates that you are quite active, which is a fantastic strength. However, your structured exercise sessions are currently limited to just three per week, so there may be room for improvement in your workout routine to ensure a balanced approach to fitness. Given your dietary intake, especially with higher grain consumption, it's important to focus on a variety of workouts to enhance energy levels and overall well-being.
   - **Zone Minutes and Intensity Feedback:** With 134 zone minutes per week, you are engaging in some moderate-intensity activity, which is great for cardiovascular health. To gradually increase your intensity, consider incorporating short bursts of higher-intensity exercises during your sessions, but always listen to your body and adjust accordingly to avoid fatigue or injury.
   - **Strength/Cardio Tips:**
     - **Cardio Exercise (Brisk Walking or Jogging):** 3 sets of 10-15 minutes, 3 times per week. Start at a comfortable pace and gradually increase duration as you feel more confident.
     - **Strength Training (Bodyweight Squats):** 2 sets of 10-15 reps, 2 times per week. Maintain proper form by keeping your back straight and knees aligned with your toes.
     - **Core Exercise (Plank):** 3 sets of 20-30 seconds, 2 times per week. Ensure your body is in a straight line from head to heels.
     - Always include a 5-10 minute warm-up of dynamic stretching before workouts and cool down with static stretching afterward to enhance flexibility and prevent injuries.
   - **Weekly Goals:**
     - 1. Increase your exercise sessions to 4 times per week, incorporating at least one additional strength training day.
     - 2. Aim to include at least 80 grams of fruit and 100 grams of vegetables daily to support your energy needs.
     - 3. Focus on hydration by reaching the goal of 3 liters of water daily.
     - 4. Schedule at least 1 full rest day each week to allow for recovery and muscle repair.

3. **Sleep, Mood, and Lifestyle**
    - **Sleep Review:** Sleep duration is adequate at 8 hours per night, supporting your active lifestyle. Slight adjustments in diet timing and exercise scheduling may further improve sleep quality.
    - **Suggestions:** To enhance your sleep quality and emotional well-being, consider establishing a consistent bedtime routine that allows you to wind down effectively each night. Aim to go to bed and wake up at the same time every day, even on weekends, to regulate your body's internal clock. Incorporate calming activities such as reading or gentle stretching before bed, which can help signal to your body that it's time to relax.
4. **Overall Summary**
    - **Summary:** You are doing an excellent job maintaining an active lifestyle with a commendable step count and sufficient sleep each night. It's great to see your commitment to a vegetarian diet, but there are opportunities to balance your food intake, particularly by increasing your fruits and vegetables. Keep up the good work, and let's focus on making small adjustments that can enhance your overall well-being!
    - **Goals:**
        - 1. Increase your exercise sessions to 4 times this week by adding one strength training session.
        - 2. Aim to include at least 80 grams of fruits and 100 grams of vegetables in your daily meals.
        - 3. Stay hydrated by reaching a daily water intake of 3 liters.
        - 4. Schedule at least one full rest day this week to support recovery and muscle repair.
        - 5. Incorporate a 10-minute evening stretching routine before bed to promote relaxation and improve sleep quality.

User Input 2

Input:
Demographics: Gender(F), Age(20), Height(163.6 cm), Weight(55.7 kg), BMI(20.7)|
Physical Activity: Weekly Steps(62182), Zone Minutes per week(88), Exercise Sessions per week(1))
Sleep: Average hours per night(6.34)
Dietary Preference: Declared Diet(Keto), Allergies(Lactose Intolerant)
Weekly Food Consumption: Dairy(3.5 L), Legumes(130 g), Meat(1087 g),

Fruits(656 g), Vegetables(861 g), Grains(859 g), Nuts(13 g), Water(16.7 L)

Output:

1. **Food Recommendations**
    - **Overall Assessment:** Your current weekly intake shows a significant excess in meat consumption, alongside low amounts of nuts and legumes. Additionally, your dairy consumption is high despite being lactose intolerant, which can lead to discomfort. Your fruit and vegetable intake is decent, but both could be more varied to meet the nutritional needs of a keto diet.
    - **Areas of Improvement:**
        - **Meat:** Reduce to 700 g per week (100 g per day)
        - **Legumes:** Increase to 350 g per week (50 g per day).
        - **Nuts:** Increase to 70 g per week (10 g per day)
        - **Dairy:** Reduce to 0 L per week (avoid entirely due to lactose intolerance)
        - **Vegetables:** Maintain at 800 g per week (around 115 g per day)
        - **Fruits:** Reduce to 300 g per week (around 40 g per day, focusing on low-sugar options)
        - **Water:** Maintain at 2-3 L per day
    - **Suggested Meals:**
        - Breakfast: Scrambled eggs with spinach and avocado, topped with pumpkin seeds (provides healthy fats and protein).
        - Lunch: Grilled chicken salad with mixed greens, olives, cucumbers, and a dressing of olive oil and lemon (low-carb and nutrient-dense).
        - Snack: A small handful of mixed nuts (almonds, walnuts) and a few slices of cucumber (great for healthy fats and fiber).
        - Dinner: Zucchini noodles with beef stir-fry (using lean cuts) and a variety of colorful bell peppers and mushrooms (rich in vitamins and minerals, while keeping it keto-friendly).
2. **Physical Activity Recommendations**
    - **Activity Assessment:** You've been quite active with an impressive weekly step count of 62,182 steps, indicating a solid base of cardiovascular fitness. However, with only one dedicated exercise session per week, there is room for improvement in strength and flexibility training.

- **Zone Minutes and Intensity Feedback:** Your total of 88 zone minutes per week suggests a moderate level of intensity during your workouts, but it's essential to ensure that you are not pushing too hard without adequate recovery. Consider gradually increasing your exercise sessions to 2-3 per week, incorporating varied intensities while always listening to your body to prevent overexertion.
- **Strength/Cardio Tips:**
  - **Bodyweight Squats:** 3 sets of 10-15 reps, 2 times per week. Focus on maintaining proper form with your knees aligned over your toes..
  - **Push-ups (knee or standard):** 3 sets of 8-12 reps, 2 times per week. Keep your core engaged and your back straight.
  - **Walking or light jogging:** 20-30 minutes, 2-3 times per week. Start at a comfortable pace, gradually increasing intensity as you feel stronger.
  - **Stretching routine:** 5-10 minutes of dynamic stretches before and static stretches after each workout to enhance flexibility and recovery.
- **Weekly Goals:**
  - 1. Increase your exercise sessions to 2-3 times this week, focusing on a mix of strength and cardio.
  - 2. Aim to include at least 2-3 different strength exercises (like squats and push-ups) in your routine.
  - 3. Ensure you are taking at least 1-2 rest days to allow your body to recover and adapt.
  - 4. Increase your water intake to at least 3 liters daily to support hydration and recovery.

3. **Sleep, Mood, and Lifestyle**

- **Sleep Review:** Average sleep of 6.33 hours per night may be insufficient for your active lifestyle. Adjusting exercise timing and balancing diet can help improve sleep quality and overall well-being.
- **Suggestions:** To improve your sleep quality, consider setting a consistent bedtime and wake-up time, which helps regulate your body's internal clock. Aim for a wind-down routine that includes relaxing activities such as reading or gentle stretching about an hour before bed, allowing your body to transition into sleep mode. Lastly, incorporating mindfulness practices such as meditation or deep breathing exercises can help stabilize your mood and promote relaxation, leading to better sleep quality over time. Focus on making these adjustments gradually for sustainable results.

4. **Overall Summary**
    - **Summary:** You've demonstrated a commendable commitment to physical activity with a high step count, showcasing your active lifestyle. While your vegetable intake is impressive, there are opportunities to align your diet more closely with your Keto preferences and improve your hydration. It's great to see your dedication to well-being, and with some adjustments, you can enhance your overall health even further.
    - **Goals:**
        - 1. Increase your exercise sessions to 2-3 times this week, incorporating both strength and cardio activities.
        - 2. Aim to consume at least 150 grams of nuts throughout the week to boost healthy fat intake.
        - 3. Increase your daily water intake to 3 liters to support hydration and overall well-being.
        - 4. Establish a consistent sleep schedule by going to bed and waking up at the same time each day.
        - 5. Include a 10-minute evening stretching routine before bed to promote relaxation and improve sleep quality.

Appendix C

Data and Resources Availability

All datasets, prompts, and models used in this thesis are stored and maintained under my Hugging Face account. The resources are organized as follows:

- **Reference Data**The full set input-ouput pairs from books and references. Hugging Face
- **Models**: The fine tuned model can be assesed from Hugging Face
- **Syntehtic Data generated by LLM** All the input output pairs can be found at dataset Hugging Face
- **Kaggle Dataset** Dataset which was primary used to generte synthetic input-output pairs. Kaggle, Kaggel
- **Other Data** other useful dataset can be found at Hugging Face Hugging Face Kaggle

Bibliography

[1] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. «Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data». In: *Proceedings of Machine Learning Research* 1 (2024). Conference on Health, Inference, and Learning (CHIL) 2024, pp. 1–15. DOI: `10.48550/arxiv.2401.06866`. URL: `https://arxiv.org/pdf/2401.06866` (cit. on p. 4).

[2] kingabzpro. *DeepSeek-R1-0528-Qwen3-8B-Medical-Reasoning*. Accessed: 2025-08-28. 2025. URL: `https://huggingface.co/kingabzpro/DeepSeek-R1-0528-Qwen3-8B-Medical-Reasoning` (cit. on p. 14).

[3] BioMistral. *BioMistral-7B*. Accessed: 2025-08-28. 2025. URL: `https://huggingface.co/BioMistral/BioMistral-7B` (cit. on p. 14).

[4] ContactDoctor. *Bio-Medical Llama 3 8B*. Accessed: 2025-08-28. 2024. URL: `https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B` (cit. on p. 14).

[5] MedAlpaca. *MedAlpaca-7B*. Accessed: 2025-08-28. 2025. URL: `https://huggingface.co/medalpaca/medalpaca-7b` (cit. on p. 14).

[6] JijaGallery. *FitLife: Health & Fitness Tracking Dataset*. `https://www.kaggle.com/datasets/jijagallery/fitlife-health-and-fitness-tracking-dataset`. Accessed: YYYY-MM-DD. 2023 (cit. on pp. 21, 22).

[7] Dan Buettner. *The Blue Zones Solution: Eating and Living Like the World's Healthiest People*. 2015. URL: `https://www.bluezones.com/the-blue-zones-solution/` (visited on 08/28/2025) (cit. on p. 22).

[8] M.D. Michael Greger. *How Not to Die*. 2015. URL: `https://nutritionfacts.org/book/how-not-to-die/` (visited on 08/28/2025) (cit. on p. 22).

[9] Luigi Fontana. *The Path to Longevity*. 2019. URL: `https://www.amazon.com/Path-Longevity-Luigi-Fontana/dp/1426213781` (visited on 08/28/2025) (cit. on p. 22).

[10] MD Peter Attia. *Outlive: The Science and Art of Longevity.* 2023. URL: `https://peterattiamd.com/outlive/` (visited on 08/28/2025) (cit. on p. 22).

[11] Héctor García and Francesc Miralles. *Ikigai: The Japanese Secret to a Long and Happy Life.* 2017. URL: `https://www.amazon.com/Ikigai-Japanese-Secret-Long-Happy/dp/0143130722` (visited on 08/28/2025) (cit. on p. 22).

[12] National Heart, Lung, and Blood Institute. *Your Guide to Healthy Sleep.* Patient booklet. 2011. URL: `https://www.nhlbi.nih.gov/sites/default/files/publications/11-5800.pdf` (visited on 08/29/2025) (cit. on p. 22).

[13] World Health Organization. *Red and Processed Meat in the Context of Health and the Environment: Many Shades of Red and Green.* Information brief. 2023. URL: `https://www.who.int/publications/i/item/9789240074828` (visited on 08/29/2025) (cit. on p. 22).

[14] World Health Organization. *WHO Guidelines on Physical Activity and Sedentary Behaviour.* Geneva: World Health Organization, 2020. ISBN: 978-92-4-001512-8. URL: `https://www.who.int/publications/i/item/9789240015128` (visited on 08/29/2025) (cit. on p. 22).

[15] World Health Organization. *Guidelines on Physical Activity, Sedentary Behaviour and Sleep for Children Under 5 Years of Age.* Geneva: World Health Organization, 2019. ISBN: 978-92-4-155053-6. URL: `https://www.who.int/publications/i/item/9789241550536` (visited on 08/29/2025) (cit. on pp. 22, 54).

[16] World Health Organization. *Global Recommendations on Physical Activity for Health.* Geneva: World Health Organization, 2010. ISBN: 978-92-4-159997-9. URL: `https://www.who.int/publications/i/item/9789241599979` (visited on 08/29/2025) (cit. on pp. 22, 54).

[17] U.S. Department of Agriculture and U.S. Department of Health and Human Services. *Dietary Guidelines for Americans, 2020–2025: Make Every Bite Count.* 9th. Government Report. Washington, DC: U.S. Department of Agriculture, U.S. Department of Health, and Human Services, 2020. URL: `https://www.dietaryguidelines.gov/sites/default/files/2020-12/Dietary_Guidelines_for_Americans_2020-2025.pdf` (visited on 08/29/2025) (cit. on p. 22).

[18] National Institute on Aging. *Exercise & Physical Activity: Your Everyday Guide from the National Institute on Aging.* NIH Publication No. 09-4258. Bethesda, MD: U.S. Department of Health and Human Services, National Institutes of Health, 2009. URL: `https://www.nia.nih.gov/health/exercise-physical-activity` (visited on 08/29/2025) (cit. on p. 22).

[19] U.S. Department of Health and Human Services and U.S. Department of Agriculture. *2015–2020 Dietary Guidelines for Americans.* Accessed: 2025-08-29. 2015. URL: `https://odphp.health.gov/sites/default/files/2019-09/2015-2020_Dietary_Guidelines.pdf` (visited on 08/29/2025) (cit. on p. 22).

[20] Royal United Hospitals Bath NHS Foundation Trust – Dietetics. *Vegetarian and Vegan Diets: How to Get the Best Nutrition.* Patient Information Leaflet. 2020. URL: `https://ruh.nhs.uk/zz_content_include/services/clinical_depts/dietetics/documents/Vegetarian_and_Vegan_Diets.pdf` (visited on 08/29/2025) (cit. on p. 22).

[21] Mayo Clinic. *Exercise: 7 benefits of regular physical activity.* Accessed: 2025-09-12. 2023. URL: `https://www.mayoclinic.org/healthy-lifestyle/fitness/in-depth/exercise/art-20048389` (cit. on p. 22).

[22] Harvard Health Publishing. *Exercise: 7 benefits of regular physical activity.* Accessed: 2025-09-12. 2024. URL: `https://www.health.harvard.edu/topics/exercise-and-fitness` (cit. on p. 22).

[23] American Heart Association. *AHA Recommendations for Physical Activity in Adults.* Accessed: 2025-09-12. 2024. URL: `https://www.heart.org/en/healthy-living/fitness/fitness-basics/aha-recs-for-physical-activity-in-adults` (cit. on p. 22).

[24] Sleep Foundation. *Mastering Sleep Hygiene: Your Path to Quality Sleep.* Accessed: 2025-09-12. 2025. URL: `https://www.sleepfoundation.org/sleep-hygiene` (cit. on p. 22).

[25] UCLA Health. *What to Eat Before and After a Workout, Based on Your Workout Type.* `https://www.uclahealth.org/news/article/what-eat-before-and-after-workout-based-your-workout-type`. [Online; accessed DATE]. July 2025 (cit. on p. 53).

[26] National Heart, Lung, and Blood Institute (NHLBI). *Heart-Healthy Living: Choose Heart-Healthy Foods.* `https://www.nhlbi.nih.gov/health/heart-healthy-living/healthy-foods`. [Online; accessed DATE]. Mar. 2022 (cit. on p. 53).

[27] — Author(s) —. «Lipid Transport and Cardiovascular Disease: Translating Basic Science into Clinical Insights». In: *Circulation Research* 123.9 (2018), e123–e150. DOI: `10.1161/CIRCRESAHA.118.313348`. URL: `https://doi.org/10.1161/CIRCRESAHA.118.313348` (cit. on p. 53).

[28] NHS. *Understanding calories.* `https://www.nhs.uk/live-well/healthy-weight/managing-your-weight/understanding-calories/`. [Online; accessed DATE]. Apr. 2023 (cit. on p. 53).

[29] World Health Organization. *Physical activity*. [Online; accessed 12 September 2025]. June 2024. URL: `https://www.who.int/news-room/fact-sheets/detail/physical-activity` (cit. on pp. 53, 54).