gAIn

AI-enabled Health and Fitness Assistant

Tomas Arevalo (tarevalo@college.harvard.edu), Jake Pappo (jakepappo@college.harvard.edu), Vincent Hock (vhock@college.harvard.edu), Mads Groeholdt (madsgroeholdt@college.harvard.edu)

Background

With health and fitness research and insights being published at a higher rate than ever across a wide variety of platforms, it is becoming increasingly difficult for the average person to stay up to date on how to keep their body in the best shape possible. gAIn seeks to consolidate newly published cutting-edge research and expert advice into a simple UI to provide the best suggestions for workouts, healthy meals, and other aspects of a person's fitness journey. Now it is well-known that generic fitness advice is a thing for the past, as the heterogeneity of users' lives requires individual advice; these outdated methods do not take advantage of our unique position in the information age. To accommodate this, gAIn will seek to seamlessly integrate with the user's personal fitness devices (such as WHOOP, Apple Health, Strava, etc.), and leverage the context of each user's historical and recent activity to provide the optimal advice for each individual.

Problem Statement

The primary deep learning model underlying gAIn's platform will: 1) be fine-tuned on a large corpus of text about health and fitness topics; and 2) have access to a user's personal data. An existing pre-trained large language model (e.g., LLaMa, Gemini, ChatGPT) will serve as the foundation for our interactive chatbot. It will be fine-tuned to gain an expert-level understanding of health topics, allowing it to engage in informative and helpful conversations with the user. We will also implement a retrieval augmented generation (RAG) pipeline through which the LLM can access the user's personal records and integrate it into personalized responses. By combining domain-specific specialization with user-specific personalization, gAIn will serve as an incredibly valuable all-in-one health and fitness coach.

Data Sources and Usage

There are two forms of data required for this project. First, we will need a large amount of fitness and health content to fine-tune our model into a reliable trainer. By consuming the guidance of professional health experts (blogs, articles, regimens), the model will learn how to interpret user data and turn it into successful feedback and coaching. Additionally, the model will need access to user health data (heart rate, sleep patterns, step counts) at the time of operation in order to provide the most relevant fitness advice.

Sources of Data

We will scrape the training data for fine-tuning from a variety of fitness pages and health databases, including but not limited to https://www.health.com/, https://www.webmd.com/, https://www.healthline.com/, https://www.strava.com/, and https://www.ncbi.nlm.nih.gov/pmc/.

Personal user data will come from Apple Watch, Apple Health, Whoop, and other wearables. Almost all of these products have APIs that allow developers to read and write their data (i.e., Apple HealthKit). Users will also be able to manually provide various data, including their meals, workouts, and general goals. Past conversations with gAIn's chatbot will be used to further enrich personalized user profiles.

Description of Dataset & Key Attributes
Articles should be verified and trusted, coming largely from accredited sources and academic institutions. We should have a diverse array of topics covered, including, but not limited to: diet, sleeping habits, exercise, and injury detection / prevention. User health data should be labeled and comprehensive, as the model's performance will benefit from historical trends.

Relevance to the Project
The articles will allow the model to gain an in-depth understanding of existing and new health recommendations and become a reliable expert on the subject matter. User health data is necessary for customizing fitness plans and goals for each user.

Data Quality Concerns
There will undoubtedly be unreliable health articles, so we must verify the sources from which we scrape data. An example of this would be to require any article used for training to be peer-reviewed. User data also has the potential to be sparse or ill-formatted, so we will need to perform preprocessing before feeding this information to the model, and potentially include some rules for whether or not to include the user data context for each individual, depending on its quality.

Scope and Preliminary Design

Minimum Components for Good Project

- **Large or Heterogeneous Data:** Described above
- **Scalability:** Our product is most definitely scalable. While our MVP will use data simply from common health apps like Apple Health, Whoop, and Fitbit, we can easily expand upon this by allowing us to partner with plenty of other apps that specifically track calories, runs, sleep cycles, or any other health related activity. With more data, we expect the chatbot to be much more specialized. On the product side, we can eventually begin to recommend informative videos from Youtube or Tik Tok along with maybe even motivational videos. Additionally, we could also add a friends feature that allows users, if desired, to see their friends activity and share any content or recommendations that they get. Long term, this could turn into a small social media app for fitness enthusiasts where there may even be a public section where people can post progress updates, new routines, and new studies.
- **Complex Models:** As explained in the problem statement, we hope to train a pre-existing LLM and fine-tune it on our data to be able to have an interactive chatbot that can give us informative and detailed fitness advice that will improve upon simply asking ChatGPT.
- **Computationally Expensive Inference:** Implementation of efficient algorithms and data storage techniques to minimize latency in both the retrieval and generation component of the chatbot.

Application Mock Design

Our app will have the following features:

1. Ability to connect existing health apps to gAIn and directly input their own data
2. A dashboard displaying personal health and fitness information
3. Interactive chatbot that informs and makes suggestions to users about workout plans, diets, and general fitness advice

Fun Factor

We were inspired to pursue this topic as we all share a background as athletes, and have all experienced the struggle of setting a training and nutrition plan that is attainable (and suitable) for our lifestyle. Professional health and fitness coaches are most certainly outside our budget, and we believe there is a substantial hole in the market for an accurate, affordable, and customized AI agent to provide everyday people with advice on how to improve themselves.

Limitations and Risks

- There are challenges to ensuring that all health-related training data come from accurate and verified sources.
- Our user data must be private, secure, and inaccessible to third parties.

Milestones/Objectives

1. Data collection and preprocessing for a large body of expert-level text on health and fitness topics.
2. Data collection and preprocessing for user-specific data via wearable device APIs and user inputs.
3. Fine-tuning an LLM model on the expert-level text to improve its ability to generate educated and helpful recommendations and advice.
4. Implementing a RAG pipeline to enable the chatbot to retrieve user-specific data for generating personalized responses.
5. Designing an intuitive, user-friendly interface with a strong back- and front-end infrastructure.
6. Testing and deployment
7.