FitFuture Bot-Project Report

Muhammad Aman
Computer Science
Habib University
ma07727@st.habib.edu.pk

Muhammad Tahir Ghazi
Computer Science
Habib University
mg07593@st.habib.edu.pk

Mahwish Ahmed
Computer Science
Habib University
ma07@st.habib.edu.pk

***Abstract*—Can artificial intelligence revolutionize fitness advice? This study explores the potential of fine-tuned language models to deliver personalized fitness solutions, bridging the gap between generic guidance and user-specific needs. We fine-tuned GPT-2 to generate fitness-related Q&A responses and DistilBERT to classify fitness descriptions into exercises or supplements. GPT-2 achieved promising text generation quality, with BLEU and ROUGE scores significantly surpassing baseline performance. Meanwhile, DistilBERT reached an impressive classification accuracy of 75%, a remarkable improvement from the baseline accuracy. These results highlight the models' ability to comprehend and generate contextually relevant fitness content.**

Despite challenges like limited datasets, our findings underscore the potential of AI to transform the fitness domain, paving the way for more advanced, user-centric applications.

***Index Terms*—Fitness,Chatbot, DistilBert, Fitness, Bleu, Gpt-2.**

I. Introduction

Fitness is more than simply a way of life in the modern world; it is a necessity for overall health and wellbeing. Yet for many, the process comes with challenges: Which exercises fit my objectives? Which supplement do I need? There has never been a greater demand for clear, trustworthy guidance.

FitFuture Bot, a virtual fitness assistant driven by cutting-edge artificial intelligence, appears as an answer for these issues. It was created as a Q&A chatbot and uses cutting-edge models to deliver accurate, approachable responses regarding workouts and supplements. FitFuture Bot will simplify decision-making and assist users in reaching their fitness objectives by combining the strength of artificial intelligence and natural language processing with domain expertise in the fitness industry.

This report looks upon the development of FitFuture Bot, considering related work, the methodology behind its fine-tuned models, the results achieved, and potential avenues for future enhancement.

II. Related Work

AI-powered chatbots have advanced significantly, especially within Natural Language Processing, enabling personalized assistance across various domains, including health and fitness. While many systems address general queries, few specifically focus on fitness-related information and recommendations. This section discusses key papers that influenced the decisions and approaches in this project. These works shaped the design of FitFuture Bot, particularly in integrating pretrained models like GPT-2, using evaluation metrics like ROUGE and BLEU, and implementing task-oriented dialogue systems. The insights from these papers guided our choices in developing an adaptive chatbot for fitness and health.

A. *Pondera: A Personalized AI-Driven Weight Loss Mobile Companion with Multidimensional Goal Fulfillment Analytics*

Pondera is an AI-driven mobile application designed to assist users in achieving weight loss and fitness goals through personalized recommendations. Utilizing NLP techniques, Pondera analyzes user inputs to offer advice related to nutrition, exercise, and mental health, while tracking progress across multiple dimensions. [1] However, its focus is primarily on weight loss, and it lacks the ability to classify queries into specific categories, such as "exercise" or "supplement." This is a key distinction from FitFuture Bot, which uses advanced NLP models like BERT for query classification and GPT-2 for generating tailored fitness responses. Unlike Pondera, FitFuture Bot aims to provide precise, category-specific responses by using fine-tuned NLP models.

B. *Hello, Its GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems*

The paper "*Hello, Its GPT-2 How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems* introduces the application of GPT-2, a transformer-based PLM to task-oriented dialogue systems. As a result of using GPT-2 the system is able to produce smooth and semantically correct responses to performed tasks without being restricted by the rules or trained only for one particular domain. [2]

This approach encouraged us to integrate GPT-2 to FitFuture Bot because it showed how pretrained models can improve the versatility and effectiveness of conversational interfaces. Besides, the paper also focuses on the evaluation metrics such as ROUGE and BLEU for the assessment of the generated responses quality. These metrics are critical to check the accuracy and fluency of the generated text with the reference response and are based on the degree of text overlap. Therefore, we used both ROUGE and BLEU in the course of the project to assess the effectiveness of the FitFuture Bot.

C. *Few-Shot Natural Language Generation for Task-Oriented Dialogue*

The paper *Few-Shot Natural Language Generation for Task-Oriented Dialogue* reviews the methods of using few-shot learning for generating the best replies in the task-oriented dialogue systems. This was achieved by using a small number of labelled examples which would allow the model to learn across the various tasks especially when the training data is limited. The authors also show that with few-shot learning, it is possible to decrease the amount of domain specific data required while still producing good quality dialogues. [3]

This paper is relevant to our work as it emphasizes the importance of efficient training techniques for dialogue systems, an issue that we faced when designing FitFuture Bot. Our project, much like the one discussed in the paper, benefits from using pretrained models such as GPT-2, which can generalize effectively with limited task-specific data. Additionally, the concept of few-shot learning aligns well with our approach to providing personalized workout and supplement recommendations, allowing the system to adapt quickly to new user inputs without requiring extensive retraining.

III. Methodology

A. *GPT-2 for Text Generation*

1) ***Objective****:* The primary goal for fine-tuning GPT-2 was to generate contextually relevant answers to fitness-related queries. We trained the model on a dataset consisting of questions and answers related to exercises and supplements, allowing it to generate responses that are informative and coherent.

2) ***Data Cleaning and Pre-Processing****:* We began with thorough research to identify websites that contained relevant data for our product categories and exercises. Initially, we web-scraped data from multiple websites to ensure a diverse and comprehensive dataset. However, after evaluating the requirements and the type of data needed, we narrowed it down to two key sourcesone for exercise details and another for product information. These websites provided detailed data on exercises, their descriptions, and instructions, as well as products with pricing, flavors, and categories.

B. *Data Cleaning*

Exercise Data:

- *Removing Redundant Links:* Any hyperlinks or unrelated URLs embedded in the raw data were filtered out to focus on exercise-specific information.
- *Extracting Key Details:* Essential fields like exercise name, muscle groups targeted, experience level, focus type, and equipment were systematically parsed and extracted.
- *Handling Missing Data:* Missing fields such as descriptions or equipment were filled with placeholder values (e.g., "N/A") to maintain consistency in the dataset.
- *Text Normalization:* Unstructured instructions were cleaned to extract actionable "How-to" steps or tutorials, ensuring usability in the analysis phase.
- *Duplicate Removal:* Duplicate exercise entries were identified and eliminated to ensure the uniqueness of the dataset.

Product Data:

- *URL Filtering:* Only product-related URLs were retained from the scraped data to remove irrelevant entries.
- *Price Standardization:* Product prices were extracted using regular expressions to ensure a consistent format (e.g., "$XX.XX") across all entries.
- *Description Parsing:* Product descriptions were extracted from nested HTML elements, retaining only relevant details while discarding extraneous content.
- *Flavor Listings:* Available flavors were extracted into structured lists for each product, improving accessibility and usability.
- *Duplicate and Null Handling:* Duplicate rows were removed, and missing fields were filled with placeholder values to ensure dataset completeness.

C. *Pre-Processing*

Exercise Data:

- *Structuring Data:* The cleaned exercise data was formatted into structured rows containing all key attributes, including exercise name, focus type, and tutorial steps.
- *Category Assignment:* Exercises were categorized based on their primary muscle groups (e.g., Back, Chest, Core) to enable focused analysis.
- *Consistent Formats:* All textual data was converted to lowercase and stripped of excess whitespace to ensure uniformity.

Product Data:

- *Categorization:* Products were categorized into predefined groups (e.g., Whey Protein, Multivitamins, Greens) using index ranges or product characteristics.
- *Null Value Handling:* Fields with missing data, such as directions or flavors, were filled with placeholder values or marked as "N/A" for clarity.
- *Consistent Data Formatting:* Text was normalized, and prices were standardized to ensure seamless integration into downstream analytical workflows.

D. *Output*

Both datasets (exercise data and product data) were saved as separate, cleaned, and pre-processed CSV files. These structured datasets ensured data accuracy and consistency, providing a robust foundation for analysis and insights generation.

1) ***Dataset****:* From the initial separate datasets for exercises and supplements, we created a combined dataset consisting of fitness-related questions and answers (Q&A pairs). The questions are about various exercises and supplements, while the answers provide detailed information about each topic.

Example Q&A pair:

Question: "What is the exercise Squats?"
Answer: "Squats are a lower body exercise that strengthens your legs."

a) Tokenization:: The next step was to tokenize the text using the GPT-2 tokenizer, which splits the text into smaller units called tokens. This is necessary because GPT-2 operates on token IDs rather than raw text. The tokenizer converts each question and answer into a sequence of token IDs.

```
def tokenize_gpt2(examples):
    # Tokenize function for gpt2
    encodings = tokenizer_gpt2(examples
    ['text'], truncation=True,
    padding=True,
    max_length=512)
    input_ids = encodings['input_ids']
    labels = input_ids.copy()
    for i in range(len(input_ids)):
        labels[i][1:] = input_ids[i][:-1]
    encodings['labels'] = labels
    return encodings
```

b) Encoding: After tokenization, each Q&A pair was encoded into token IDs that could be input into the model. Special tokens like ¡EOS¿ (end of sequence) were added to mark the ending boundaries of each sequence.

c) Padding and Truncation: To ensure that all sequences fed into the model have the same length, padding and truncation were applied to the sequences, standardizing them to a fixed maximum length (e.g., 512 tokens).

*2) **Training Procedure**:* The GPT-2 model was fine-tuned on the preprocessed data using the following procedure:

a) Model Initialization:: A pre-trained GPT-2 model was loaded from the Hugging Face Transformers library. The model was then adapted to the specific task of text generation by fine-tuning it on the fitness Q&A pairs.

b) Optimization:: The model was trained using the Adam optimizer with a learning rate of 5×10^{-5}. The batch size was set to 4 to fit within the available memory.

c) Epochs:: The model was trained for 5 epochs, ensuring that it had enough iterations to learn from the training data without overfitting.

d) Evaluation:: The performance of the fine-tuned GPT-2 model was evaluated on a validation set using BLEU and ROUGE scores, which are standard metrics for evaluating the quality of generated text.

```
model = GPT2LMHeadModel.from_pretrained
('gpt2')
training_args_gpt2 = TrainingArguments(
    output_dir="./gpt2_finetuned",
    num_train_epochs=5,
    per_device_train_batch_size=4,
    save_steps=1000,
    logging_steps=100,
    logging_dir="./logs",
    save_total_limit=2,
    learning_rate=5e-5,
)
trainer_gpt2 = Trainer(
    model=model_gpt2,
    args=training_args_gpt2,
    train_dataset=tokenized_gpt2_data,
)
trainer_gpt2.train()
```

E. DistilBERT for Text Classification

1) Objective: The objective for fine-tuning DistilBERT was to classify fitness-related descriptions into two categories: exercise and supplement. The model was trained to distinguish between product descriptions of fitness exercises and supplements.

*2) **Dataset**:* From the initial separate datasets for exercises and supplements, we created a combined dataset consisting of labeled descriptions from two categories:

- Exercise (1)
- Supplement (0)

Each text description was paired with a binary label indicating its category.

a) Tokenization:: The DistilBERT tokenizer was used to split the text into tokens suitable for input into the DistilBERT model. This tokenizer converts text into subword units and prepares it for model consumption.

```
tokenizer_bert = DistilBertTokenizer.from_
pretrained('distilbert-base-uncased')

# Tokenize function
def tokenize_function(examples):
    if isinstance(examples['text'], list):
        texts = examples['text']
    else:
        texts = [examples['text']]
    return tokenizer_bert(texts,
    truncation=True, padding=True,
    max_length=512)
```

b) Padding and Truncation:: The text descriptions were padded or truncated to a fixed maximum length (e.g., 512 tokens), ensuring consistent input size.

3) Training Procedure: The DistilBERT model was fine-tuned on the preprocessed dataset using the following approach:

a) Model Initialization:: A pre-trained DistilBERT model was loaded and adapted for binary classification by adding a classification head (a fully connected layer).

b) Optimization:: The model was fine-tuned using the AdamW optimizer, with a learning rate of 5×10^{-5}. We used a batch size of 8 for training.

c) Epochs:: The model was trained for 5 epochs.

d) Evaluation:: The model's performance was evaluated using accuracy, the percentage of correct classifications on the validation set.

```
training_args_bert = TrainingArguments(
    output_dir="./bert_finetuned",
    num_train_epochs=5,
    per_device_train_batch_size=8,
    save_steps=1000,
    logging_steps=100,
    evaluation_strategy="epoch",
    logging_dir="./logs_bert",
    save_total_limit=2,
    learning_rate=5e-5,
)

trainer_bert = Trainer(
    model=model_bert,
    args=training_args_bert,
```

```
    train_dataset=tokenized_data['train'],
    eval_dataset=tokenized_data['test'],
    compute_metrics=lambda p:
    {"accuracy": (p.predictions.argmax
    (axis=-1) == p.label_ids).mean()},
)

trainer_bert.train()
```

IV. Experiments and Results

A. GPT-2: Text Generation Results

The GPT-2 model was evaluated using BLEU and ROUGE scores. These metrics are widely used to measure the quality of generated text compared to reference text.

a) Metrics Used::

- **BLEU Score:** Measures the n-gram overlap between generated text and reference text.
- **ROUGE Score:** Measures the recall of overlapping n-grams between generated and reference text.

b) Results for GPT-2:: The GPT-2 model achieved the following scores after being fine-tuned on the dataset during evaluation:

Metric	Value
BLEU	0.225
ROUGE-1	0.327
ROUGE-2	0.173
ROUGE-L	0.284

TABLE I
GPT-2 Text Generation Results

B. DistilBERT: Text Classification Results

The DistilBERT model was evaluated on the binary classification task of categorizing fitness-related descriptions into either exercise or supplement. The model's performance was measured using accuracy.

a) Results for DistilBERT:: After being fine-tuned on the dataset, The DistilBERT model achieved **75%** accuracy in the text classification task.

C. Performance Comparison Table

The results achieved in both text generation and text classification were highly encouraging especially when compared with the baseline performances of these models on the same tasks and keeping in mind the limitations of the project in terms of data and computational resources. The following table provides a comparison of the models' baseline performance and performance after fine-tuning :

	Accuracy	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	25.8%	0.037	0.095	0.03	0.083
Fine-tuned	75%	0.225	0.327	0.173	0.284

TABLE II
Comparison of Baseline and Fine-tuned Results

V. Discussion

The findings of using GPT-2 and DistilBERT for fitness text generation and classification tasks are more effective than the baseline models. For the model being discussed, GPT-2, the fine tuning was done on a dataset of Q&A pairs related to fitness, and the idea was to get the model to produce answers that are relevant to the given context. After evaluating the model using BLEU and ROUGE scores, the results were as follows:

- BLEU: 0.225
- ROUGE-1: 0.327
- ROUGE-2: 0.173
- ROUGE-L: 0.284

These metrics indicate that the fine-tuned GPT-2 model generates text that closely aligns with reference answers, achieving a meaningful level of coherence and relevance. While the BLEU and ROUGE scores were moderate, they are consistent with typical performance for models of this type, especially given the relatively small size of the dataset and computational constraints.

The DistilBERT model, trained for the task of classifying fitness-related descriptions into exercise or supplement categories, achieved a classification accuracy of 75%. This result is significant, demonstrating that DistilBERT effectively learned to distinguish between the two categories, even though the dataset was smaller than what would typically be used for such tasks.

When comparing these results to the baseline performance, both models exhibited substantial improvements:

- GPT-2:
  - BLEU: 0.037 (baseline) $\rightarrow$ 0.225 (fine-tuned)
  - ROUGE-1: 0.095 (baseline) $\rightarrow$ 0.327 (fine-tuned)
- DistilBERT:
  - Accuracy: 25.8% (baseline) $\rightarrow$ 75% (fine-tuned)

Overall, the fine-tuned models outperformed their baseline versions by a significant margin. Despite the limitations in data and computational resources, these results indicate the effectiveness of GPT-2 for text generation and DistilBERT for text classification in the fitness domain, suggesting further improvements with more data and optimization.

VI. Future Work and Conclusion

A. Future Work

FitFuture Bot's future iterations will focus extensive and thorough collection of data in order to significantly enhance model accuracy and contextual understanding. We hope to overcome the present model's weaknesses by collecting a large and diversified dataset that includes well-documented training plans, validated supplement information, and culturally appropriate instructions.

Improving the model's accuracy and BLEU scores will be the primary goal. This will entail fine-tuning on high-quality annotated datasets and applying advanced optimization techniques to improve the bot's linguistic coherence

and domain-specific knowledge. Experiments with larger pre-trained language models and transfer learning approaches will also be investigated to improve performance metrics further.

Furthermore, user feedback loops can also be used to iteratively enhance the chatbot's responses, ensuring that they line with user expectations and fitness goals. By combining higher data quality with focused model enhancements, we hope to develop FitFuture Bot as a dependable, accurate, and user-friendly virtual fitness assistant.

B. Conclusion

FitFuture Bot is a key step in incorporating artificial intelligence and NLP into customized fitness and nutrition plans. Through the use of a conversational model based on GPT-2 and DistilBert, this project shows how AI-driven solutions can streamline access to trustworthy fitness advice. Despite existing limits in data quality and performance metrics, the bot's platform demonstrates the potential for scalable, interactive, and user-centric fitness support.

However, the journey is far from over. The issues faced have revealed chances for improvement, particularly in terms of accuracy and coherence. With plans to increase the dataset and add advanced fine-tuning techniques, FitFuture Bot intends to be an inspiration in the Fitness Industry, bridging knowledge gaps and allowing people to easily attain their health goals.

REFERENCES

[1] G. Pashev and S. Gaftandzhieva, "Pondera: A Personalized AI-Driven Weight Loss Mobile Companion with Multidimensional Goal Fulfillment Analytics," Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024), Sofia, Bulgaria, Sep. 2024, pp. 264-271.

[2] Budzianowski, P. and Vuli, I., "Hello, Its GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems," Proceedings of the 3rd Workshop on Neural Generation and Translation, Hong Kong, Nov. 2019, pp. 15-22, doi: 10.18653/v1/D19-5602.

[3] B. Peng, C. Zhu, C. Li, X. Li, J. Li, M. Zeng, and J. Gao, "Few-shot Natural Language Generation for Task-Oriented Dialog," Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 172-182, Online, 2020.