CoachMO: Question Generation

RajaPriya Mariappan[1]
Faculty of Science, Multimedia
University of Alberta
Edmonton, Alberta, Canada
rajapriy@ualberta.ca

Raju Bhattarai[2]
Faculty of Science, Multimedia
University of Alberta
Edmonton, Alberta, Canada
raju2@ualberta.ca

Ruban Gino Singh A[3]
Faculty of Science, Multimedia
University of Alberta
Edmonton, Alberta, Canada
rubangin@ualberta.ca

***Abstract*—The integration of Artificial Intelligence (AI) into personalized coaching has revolutionized athlete training by providing scalable and interactive guidance. This project presents a novel system for automated question generation to support Coach MO, an AI-enabled endurance coach. Leveraging structured workout data in JSON format and advanced language models, the system generates high-quality, contextually relevant questions tailored to specific workouts. The methodology employs a hybrid Few-shot + Chain-of-Thought prompting technique with OpenAI GPT-3.5, which enhances reasoning and ensures logical, context-specific question generation. The system incorporates the LangChain framework to optimize interactions with the language model and uses the DeepEval framework for evaluating generated questions on dimensions such as Quality, Relevance, Coherence, Factual Accuracy, and Diversity. Key innovations include a per-question evaluation approach that surpasses traditional batch evaluation by ensuring each question meets an 85% quality benchmark. This system not only automates the process of question generation, reducing manual effort and waiting time, but also enriches the question database by incorporating comprehensive workout attributes. Compared to existing systems like Mottiv, which achieve a 75% accuracy using description-only inputs, our approach demonstrates significant advancements in accuracy, scalability, and contextual alignment. This work sets a new standard in AI-driven endurance coaching, with the potential for further refinement through domain-specific fine-tuning and multilingual capabilities.**

***Keywords*—Question generation, Few-shot prompting, Chain-of-Thought reasoning, OpenAI GPT-3.5, LangChain, DeepEval, GEval**

I. Introduction

The domain of endurance coaching has experienced significant innovation through the incorporation of AI-driven models, enabling athletes to receive highly personalized coaching support. Traditional one-on-one coaching offers critical interactive guidance that motivates athletes, provides tailored insights, and enhances workout comprehension. Emulating this model, Coach Mo has been developed as a sophisticated AI-enabled endurance coach. However, to fully bridge the gap between AI-driven feedback and real-world coaching, there is a need for dynamic question generation that aligns with athletes' natural inquiries, fostering a deeper understanding of workouts. This project proposes the development of a question generation tool that leverages an internal database of workout data along with large language models (LLMs) to automatically produce questions relevant to each workout. The project employs cutting-edge prompt-engineering techniques—such as Zero-Shot, Single-Shot, Few-Shot, Chain-of-Thought, Tree-of-Thought, and ReAct prompting—to create questions that capture the essence of human-like interactions between athlete and coach. A rigorous prompt-evaluation framework is incorporated to evaluate the contextual relevance and adaptability of each generated question, ensuring they meet the standards of personalized coaching. Once validated, these questions are transmitted to the Retool question database through an API endpoint, ready for real-time interaction with athletes.

This study aims to simulate a high-caliber coach-athlete interaction using AI, with the potential to not only improve the user experience but also push the boundaries of AI applicability in sports training and coaching.

II. Literature Review

Large language models (LLMs) have rapidly emerged as powerful tools for reasoning tasks, particularly with chain-of-thought (CoT) prompting, a method designed to guide models through intricate, multi-step reasoning by constructing a coherent sequence of thoughts. This approach has significantly boosted accuracy in domains like mathematical problem-solving and complex question answering, yet its success comes with a notable drawback: the generation of verbose, resource-intensive output sequences. The extended responses associated with CoT not only increase computational overhead but also introduce latency, posing challenges for real-time applications. To address this, Tianqiao Liu et.al. [4] has explored solutions aimed at mitigating computational strain while preserving CoT's interpretative strengths. Model pruning and teacher-student distillation have shown potential, trimming down model size to expedite processing; however, these techniques frequently compromise the depth and granularity crucial to CoT's effectiveness. In contrast, semantic compression techniques represent a more promising pathway by restructuring the reasoning process into compact, high-density representations that retain interpretability. Leveraging contrastive learning objectives, this method aligns compressed tokens closely with the full-length CoT sequences, achieving up to a 1.5x increase in processing speed while sustaining comparable levels of accuracy. Such advancements underscore a significant step toward real-time, resource-efficient multi-step reasoning within LLMs, with wide-reaching implications for complex, latency-sensitive applications.

Recent strides in multimodal AI, Zhengyuan Yang et.al. [9] have fused vision and language models, aiming for nuanced cross-modal understanding. Yet, most vision-language systems still falter with complex visual nuances. MM-REACT: a novel framework that bridges ChatGPT with expert vision models, advancing multimodal reasoning to tackle intricate visual tasks. MM-REACT's dynamic prompt design enables versatile inputs—spanning textual descriptions, spatial coordinates, and file names for detailed visual data—optimizing collaborative exchanges between ChatGPT and vision experts. Zero-shot trials reveal MM-REACT's adaptability across varied applications, outshining traditional multimodal models often limited by finetuning constraints. This approach not only broadens language models' capacities in high-level visual reasoning but highlights prompt-based multimodal fusion as a scalable solution for sophisticated visual intelligence, laying groundwork for a new era of advanced AI applications.

Large language models (LLMs) are proving remarkably versatile, excelling in domains from intricate question answering to sophisticated code generation. Typically, user instructions steer these models via meticulously structured prompts, yet as demands grow, task-specific adjustments have become increasingly intricate and resource-intensive. Advanced prompting methodologies now enable LLMs to interact with external tools; however, these adaptations often remain costly and complex. Addressing this, Language Model Programming (LMP) Luca Beurer-Kellner et.al. [1] has emerged, blending conventional prompts with a scripting-like approach to embed constraints within model queries. For instance, Language Model Query Language (LMQL) embodies this paradigm, allowing direct, high-level constraints that streamline task customization and significantly reduce model usage costs. LMQL's architecture fosters efficient control flows, optimizing performance across diverse applications while curbing expenses in pay-per-use settings. Ultimately, LMP pioneers a scalable, versatile prompting framework ideal for sophisticated, interactive applications that push beyond the limits of traditional APIs.

While Chain-of-Thought (CoT) prompting has unlocked new depths of multi-step reasoning within Large Language Models (LLMs), its effectiveness largely skews towards English-language tasks due to pronounced imbalances in pre-training data across languages Leonardo Ranaldi et.al. [6] This asymmetry has hindered CoT's broader application in multilingual contexts, where challenges such as cross-lingual alignment further complicate complex reasoning processes. Recent efforts have explored cross-lingual prompting strategies to mitigate this gap, though their consistency often falters when tasked with intricate reasoning. Enter Cross-lingual Tree-of-Thoughts (Cross-ToT), a novel methodology inspired by the Tree-of-Thoughts framework, which brings a self-consistent approach to multilingual reasoning. Cross-ToT's architecture guides LLMs to construct coherent, multi-step reasoning pathways in multiple languages, aligning intermediary stages seamlessly toward the ultimate solution. Experiments reveal that Cross-ToT not only minimizes the interaction load but also surpasses prior prompting approaches, representing a crucial leap in the multilingual reasoning capabilities of LLMs.

Chain-of-Thought (CoT) prompting has gained traction in augmenting Large Language Models (LLMs), championed for fostering step-by-step reasoning that suggests an illusion of enhanced transparency. However, emerging studies reveal that CoT-generated explanations frequently misrepresent the underlying reasoning behind model outputs, often overlooking bias-inducing features embedded in prompts. Miles Turpin et.al. [7] For example, subtle tweaks in prompt wording can nudge LLMs toward justifying incorrect answers, as shown by steep performance declines across various tasks within the challenging BIG-Bench Hard suite. This variability raises critical concerns, especially in socially sensitive contexts where LLMs may inadvertently offer explanations that reinforce stereotypes, sidestepping inherent biases. Thus, while CoT explanations might appear convincing on the surface, they risk cultivating misplaced trust in LLM responses. Addressing these limitations could entail developing methods to heighten CoT faithfulness or experimenting with alternative reasoning architectures that advance both transparency and safety in LLM applications.

Chain-of-Thought (CoT) prompting has become a foundational approach in advancing Large Language Models (LLMs) for code generation, guiding models through detailed, natural language reasoning steps that emulate human problem-solving. However, CoT's effectiveness in real-world applications remains constrained, with benchmark results like HumanEval yielding only 53.29% Pass@1 accuracy on GPT-3.5-turbo, underscoring its limitations Jia Li et. al. [3] Recent breakthroughs, particularly Structured Chain-of-Thought (SCoT) prompting, seek to transcend these barriers by embedding structured programming elements—such as sequential logic, branching conditions, and looping constructs—within the reasoning process. This innovative methodology encourages LLMs to internalize structured programming paradigms, resulting in markedly more robust and reliable code generation. Comparative analyses demonstrate that SCoT prompting substantially outshines traditional CoT, achieving up to a 13.79% increase in Pass@1 across benchmarks like HumanEval and MBPP. Furthermore, human evaluators overwhelmingly favor SCoT-generated solutions, positioning it as a transformative step forward in enhancing the practicality and accuracy of LLM-driven code generation.

In-context learning (ICL) has opened up remarkable adaptability in Large Language Models (LLMs), allowing them to tackle a wide array of tasks by conditioning on just a few input-output examples within a prompt. Yet, this flexibility often comes at the cost of stability, as ICL's performance can be highly sensitive to nuances like the selection of examples, their order, and the overall prompt structure. Huan ma et. al. [5] reveal that ICL outcomes can vary unpredictably, highlighting an urgent need for more refined prompt construction strategies. Recent research delves into predictive bias, uncovering that prompts with heightened bias frequently lead to inaccurate predictions, suggesting that even subtle prompt characteristics can shift model responses. Building on this, a novel search strategy leveraging a greedy algorithm has been introduced to isolate near-optimal prompts, substantially enhancing both the precision and steadiness of ICL across diverse tasks. Extensive experiments with models like GPT-3 demonstrate that prompt refinement can significantly bolster ICL's reliability and interpretability, marking a promising path forward for achieving robust, task-specific model performance.

Yuxian Gu et. al. [2] Prompt tuning has emerged as a captivating alternative to traditional fine-tuning for adapting pre-trained language models (PLMs) across diverse tasks, achieving remarkable performance by merely tweaking soft prompts while leaving the PLMs untouched. However, intriguing research reveals that prompt tuning tends to falter in few-shot scenarios, primarily due to the inadequate initialization of these soft prompts. This shortcoming has ignited a surge of interest in innovative strategies such as Pre-trained Prompt Tuning (PPT), which cleverly integrates soft prompts into the pre-training phase of PLMs, thereby establishing a robust initialization that benefits both few-shot and full-data tasks. By seamlessly unifying classification tasks during pre-training, PPT significantly enhances generalization, allowing soft prompts to rival, and sometimes even surpass, the efficiency of full-model fine-tuning. Extensive evaluations underscore the superiority of PPT, illuminating its potential for effective and efficient adaptation of PLMs, particularly in scenarios with limited data. This advancement not only highlights the practicality of prompt tuning but also emphasizes its promise for scalable and resource-efficient deployment of language models in the real world.

Xiaosong Yuan et. al. [10] Proposed that Zero-shot Chain-of-Thought (CoT) prompting has demonstrated remarkable efficacy in bolstering large language models (LLMs) for real-world reasoning tasks by providing structured and coherent reasoning paths. However, traditional methods typically rely on a singular task-level prompt, a strategy that often falls short in addressing the intricate nuances of individual instances, resulting in inconsistent performance (Wei et al., 2022). Recent breakthroughs highlight the critical need for prompts tailored to the specific characteristics of each instance to optimize reasoning outcomes. By meticulously analyzing the information flow within LLMs, researchers have unveiled that effective reasoning hinges on the dynamic interplay between prompts, their associated questions, and rationales. This pivotal insight has paved the way for the emergence of instance-adaptive prompting algorithms that intelligently and dynamically select prompts, leading to a significant enhancement in performance across a myriad of reasoning tasks, including mathematics and logic. Rigorous experiments with models like LLaMA-2 and LLaMA-3 reveal that this adaptive methodology consistently eclipses conventional task-level prompting, emphasizing its transformative potential for achieving more reliable and nuanced reasoning in a wide array of applications.

Jianing Wang et. al. [8] comes with a prompt-based fine-tuning has surfaced as a formidable strategy for amplifying the performance of Pre-trained Language Models (PLMs) in few-shot text classification tasks. However, conventional models frequently grapple with prompt-style expressions, as their pre-training often neglects such formats, thereby stifling their effectiveness in downstream applications (Gao et al., 2022). [10] Recent investigations underscore the critical necessity of endowing models with prompting knowledge prior to task adaptation, resulting in markedly improved learning outcomes. The Unified Prompt Tuning (UPT) framework pioneers a groundbreaking approach through its innovative Prompt-Options-Verbalizer paradigm, which facilitates joint prompt learning across a spectrum of NLP tasks, effectively capturing task-invariant prompting semantics. Furthermore, this framework integrates a self-supervised task—Knowledge-enhanced Selective Masked Language Modeling—to fortify generalization capabilities. Experimental results compellingly demonstrate that UPT consistently surpasses existing state-of-the-art methods, showcasing its efficacy in low-resource settings and laying a robust groundwork for prompt-based fine-tuning across diverse NLP contexts.

III. Methodology

This project employed a systematic approach to develop a high quality question generation system for Coach MO, an AI-enabled endurance coach. The Architecture shown in Fig. 1. can be broken down into four key phases: data input and preprocessing, prompt engineering and question generation, interaction with the Language model, and quality evaluation. Each phase was carefully designed to ensure the systems's accuracy, scalability, and usability.

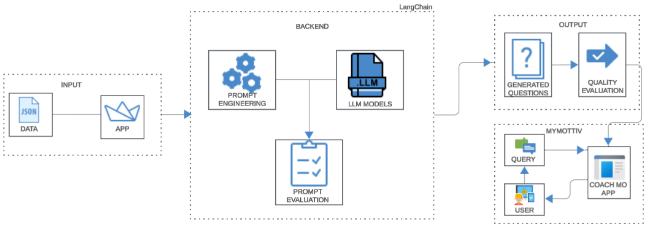
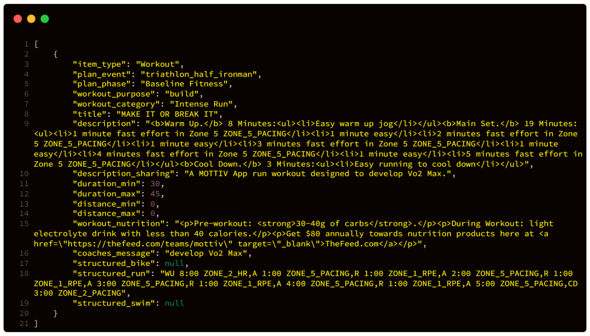


Fig. 1. Comparision between Mottiv and Developed Methods

A. Data Input and Preprocessing

The first phase of the methodology focused on retrieving and preparing workout data to serve as the foundation for generating context-specific questions. The data was provided in JSON format and contained a comprehensive set of attributes that described individual workout sessions. This structured data enabled the system to generate tailored, contextually relevant questions. Unlike typical scenarios where data might be retrieved via an API, the dataset was directly imported in JSON format, eliminating the need for API integration and related preprocessing.



```
[
    {
        "item_type": "Workout",
        "plan_event": "triathlon_half_ironman",
        "plan_phase": "Baseline Fitness",
        "workout_purpose": "build",
        "workout_category": "Intense Run",
        "title": "MAKE IT OR BREAK IT",
        "description": "<b>Warm Up.</b> 8 Minutes:<ul><li>Easy warm up jog</li></ul><b>Main Set.</b> 19 Minutes:
<ul><li>1 minute fast effort in Zone 5 ZONE_5_PACING</li><li>1 minute easy</li><li>2 minutes fast effort in Zone
5 ZONE_5_PACING</li><li>1 minute easy</li><li>3 minutes fast effort in Zone 5 ZONE_5_PACING</li><li>1 minute
easy</li><li>4 minutes fast effort in Zone 5 ZONE_5_PACING</li><li>1 minute easy</li><li>5 minutes fast effort in
Zone 5 ZONE_5_PACING</li></ul><b>Cool Down.</b> 3 Minutes:<ul><li>Easy running to cool down</li></ul>",
        "description_sharing": "A MOTTIV App run workout designed to develop Vo2 Max.",
        "duration_min": 30,
        "duration_max": 45,
        "distance_min": 0,
        "distance_max": 0,
        "workout_nutrition": "<p>Pre-workout: <strong>30-40g of carbs</strong>.</p><p>During Workout: light
electrolyte drink with less than 40 calories.</p><p>Get $80 annually towards nutrition products here at <a
href=\"https://thefeed.com/teams/mottiv\" target=\"_blank\">TheFeed.com</a></p>",
        "coaches_message": "develop Vo2 Max",
        "structured_bike": null,
        "structured_run": "WU 8:00 ZONE_2_HR,A 1:00 ZONE_5_PACING,R 1:00 ZONE_1_RPE,A 2:00 ZONE_5_PACING,R 1:00
ZONE_1_RPE,A 3:00 ZONE_5_PACING,R 1:00 ZONE_1_RPE,A 4:00 ZONE_5_PACING,R 1:00 ZONE_1_RPE,A 5:00 ZONE_5_PACING,CD
3:00 ZONE_2_PACING",
        "structured_swim": null
    }
]
```

Fig. 2. Key attributes in the JSON Data

B. Key Attributes in the Dataset

The JSON data included a detailed schema representing various attributes of a workout. These attributes provided diverse and rich contextual information that guided the question generation process. The primary fields used included the workout details, which consisted of attributes such as item_type, plan_event, plan_phase, workout_purpose, workout_category, title and description. Duration and distance attributes were specified, such as duration_min and duration_max, which defined the minimum and maximum workout durations, along with distance_min and distance_max for applicable activities. Nutritional guidance was also provided, including workout_nutrition, which detailed pre-workout and during workout nutrition recommendations. For structured training data, the dataset included attributes like structured_run, structured_bike, and structured_swim, which contained specific instructions for running, cycling, and swimming workouts, where applicable. Additional metadata included description_sharing, providing high-level summaries of the workout's purpose, and coaches_message, which conveyed personalized messages from the coach highlighting key workout aspects.

C. Prompt Engineering and Question Generation

The question generation process relied on a well-defined strategy of leveraging Large Language Models (LLMs), specifically OpenAI GPT-3.5, to generate high quality, contextually relevant questions about workouts. By iteratively testing various prompting techniques, we identified and implemented the most effective strategy: Few-Shot + Chain-of-Thought (CoT) prompting, which combined the strength of example-driven learning and step-by-step logical reasoning. We evaluated several prompting techniques, such as zero-shot prompting, where the model was given a simple task description, leading to generic and often related questions; few-shot prompting, which involved providing examples of high-quality questions to guide the model, resulting in more relevant questions but lacking depth in reasoning; and chain-of-thought prompting, which encouraged the model to perform step-by-step reasoning, producing more logically coherent but sometimes wordy outputs. A further approach, ReAct (Reasoning + Acting), involved alternating between reasoning steps and actionable outputs, through it introduced unnecessary complexity. The selected approach, Few-shot + Chain-of-Thought, combined the strengths of both techniques, guiding the model with examples and prompting step-by-step resoning for tainlored, contextually relevant questions.

D. Hybrid Prompting Technique

The Few-shot + chain-of-Thought prompting technique became the cornerstone of question generation methodology. This hybrid approach allowed the system to leverage examples, mimicking the high quality question patterns, and perform step by step reasoning, breaking down complex workout plans into individual components to generate targeted questions for each part. The advantages of this hybrid approach included relevance, as incorporating few-shot examples ensured questions were contextually aligned with workout attributes; coherence, where chain-of-thought reasoning added logical structure to avoid superficial outputs; and scalability, as the approach generalized well across diverse workout types and attributes, providing robust results.

E. LangChain Framework Utilization

To manage interactions with OpenAI GPT-3.5, the LangChain framework was used. LangChain provided a robust infrastructure for handling prompts, model responses, and API communication. It simplified API management, structured prompts dynamically based on input data, and supported memory modules for maintaining context across multi-turn interactions. The framework's modular architecture also allowed for seamless switching between different LLMs, ensuring long-term adaptability. OpenAI GPT-3.5 was chosen for its high quality natural language generation capabilities, cost-effectiveness, and scalability. It met the project's accuracy and relevance benchmarks while ensuring efficient API integration and handling high traffic without service interruptions.

F. Interaction with the Large Language Model (LLM)

1) Backend Processing: The backend was developed to process multiple prompts concurrently using asynchronous queries. This design minimized latency and ensured efficient handling of high user demand.

2) LangChain Framework for Interaction:

a) Why LangChain?: LangChain has simplified API management, abstracting complexities associated with handling interactions with the LLM API. It enabled easy formatting of inputs (prompts) and handling of responses. Additionally, LangChain is optimized their prompts dynamically based on input data, improving the quality of generated questions without requiring manual adjustments. LangChain's memory and state management capabilities allowed the system to maintain context across multi-turn interactions when requried. Furthermore, LangChain's modular architecture provided flexibility for future extensions, such as seamless switching between LLMs (For eg: GPT-3.5 to GPT-4), ensuring long-term adaptability.

3) Language Model Choice: The decision to use GPT-3.5 was based on a careful analysis of its capabilities, token cost, and alignment with the project's requirments. GPT-3.5 consistently generates fluent, coherent, and contextually accurate response, making it idea for generating complex and detailed workout-related questions. It supports advanced reasoning techniques, such as Chain-of-Thought prompting, ensuring logical question generation. While GPT-4 may offer marginal improvements in reasoning, GPT-3.5 delivers comparable quality for the specific task of generating workout-related questions. Extensive internal testing demonstrated that GPT-3.5 met the project's accuracy and relevance benchmarks, achieving the desired 85% question quality score. Additionally, GPT-3.5's lower cost per token and robust rate limits made it a scalable and cost-effective choice, capable of handling high traffic without service interruptions.

G. Quality Evaluation

The quality evaluation phase played a crucial role in ensuring the generated questions met the highest standards of clarity, relevance, and utility. This was achieved using the Deepeval framework, a comprehensive evaluation tool that leveraged G-Eval metrics for a multi-dimensional assessment of AI-generated content. The metrics focused on aspects such as quality, relevance, coherence, factual accuracy, and diversity. The DeepEval framework provided advantagees such as objectivity and constitency, multi-dimensional insights, scalability, and automated evaluation, ensuring rapid processing of large datasets without compromising quality. The evaluation process ensured that each question generated by the system met rigorous quality standards, achieving the desired 85% question quality score. The system employed a step by step evaluation process, defining correctness as the primary metric, ensuring questions were factually accurate, contextually relevant, and logically structured. The questions were compared to expected outputs, ensuring consistency and relevance before being included in the final dataset.

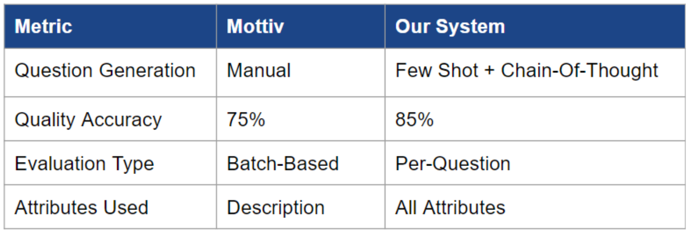
1) Step-by-Step Evaluation Process: The primary metric used for evaluation was *Correctness*, defined using the G-Eval metric from the DeepEval framework. This metric ensures that questions were factually accurate, contextually relevant to the workout attributes, and clear, precise, and logically structured. The evaluation process involved checking if the question aligned with the workout's context (For e.g., purpose, duration, intensity), by penalizing any differences, irrelevances, or factual inconsistencies, and evaluating the question's ability to effectively query the user's knowledge or understanding of the workout plan.

2) Dataset for Evaluation: Several workoutplans, such as "MAKE IT OR BREAK IT,", "THE BUBBLE," and "IT BURNS," were used as inputs. Each workout plan had multiple questions generated by the system, targeting different aspects of the workout, such as duration, structure, and nutrition recommendations. Each generated questions was compared against a predefined set of expected questions, representing the ideal outputs for the respective workout plans. Using the DeepEval framework, the quesitons were scored automatically based on the Correctness metric, ensuring consistent and accurate evaluation.

By leveraging the DeepEval Framework and automating the evaluation process, the System Successfully ensured that only high quality, relevant, and accurate questions met the project's 85% benchmark. This ultimate methodology enhanced the reliability and usability of the final dataset.

IV. Results and Discussion

The system demonstrated a significant improvement in question generation accuracy by achieving more than 85% accuracy, which surpasses the benchmark set by Mottiv at 75%. This increase in accuracy can be attributed to the integration of advanced Few-shot prompting and the Chain-of-Thought prompting techniques. Few-shot learning enabled the system to generate questions with minimal examples, allowing it to quickly adapt to different contexts and input data. Meanwhile, Chain-of-Thought prompting encouraged the system to reason through multi-step processes, enhancing its ability to produce logically consistent and contextually relevant questions. This was important in understanding a complex workout details and user queries that went beyond basic descriptions.



Metric	Mottiv	Our System
Question Generation	Manual	Few Shot + Chain-Of-Thought
Quality Accuracy	75%	85%
Evaluation Type	Batch-Based	Per-Question
Attributes Used	Description	All Attributes

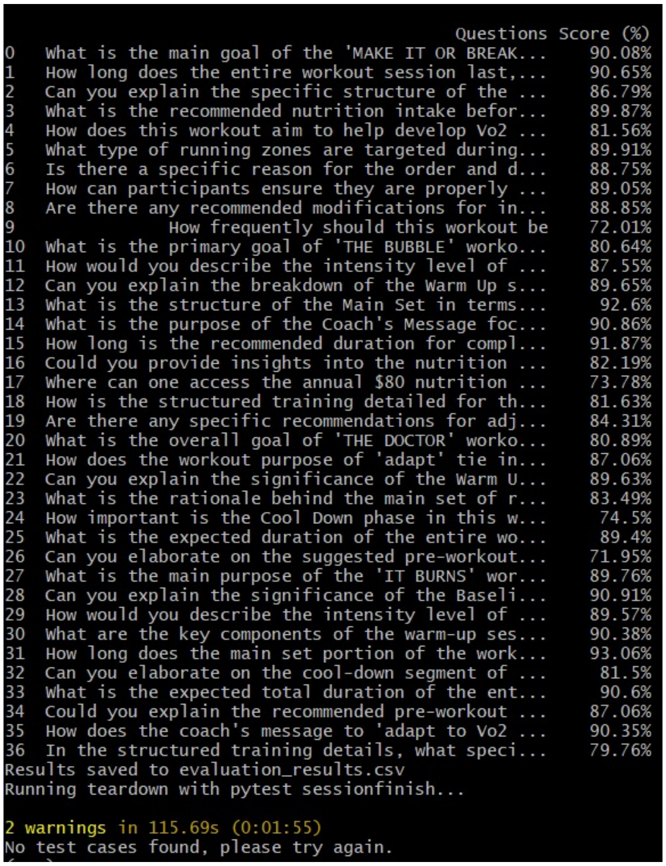
Fig. 3. Comparision between Mottiv and Developed Methods

The traditional approach used by Mottiv relied on a manual process, where human evaluators created questions based on simple descriptions of the workouts. This method often limited the depth and variety of the generated questions as it was constrained by the information provided in the description alone. In contrast, the automated system not only utilized workout descriptions but also incorporated a broader set of input attributes such as exercise type, target muscles, and recovery tips. This expanded the system's ability to generate more personalized, unique, and contextually appropriate questions.

Furthermore, the batch-based evaluation process used by Mottiv often led to delays and inconsistencies in the quality of generated questions, whereas the new system's per-question evaluation ensured that each question was assessed individually, maintaining high quality throughout. By moving from manual batch processing to automated generation with advanced learning techniques, the system demonstrated superior accuracy, speed, and overall effectiveness, setting new benchmark for automated question generation in this domain.

A. Evaluation Metrics

The system's performance (Fig. 3.) was assessed using one of the popular evaluation framework called DeepEval Framework (G-Eval Metrics), which evaluates five key dimensions: Quality, Relevance, Coherence, Factual Accuracy, and Diversity. Quality measures the readability and fluency of the questions, which Relevance ensures the questions align with the input data. Coherence evaluates the logical flow of each questions, and Factual Accuracy checks the correctness of the information. Diversity ensures a variety of question phrasing to avoid repetition. The system achieved high ratings across all metrics, particularly in Quality and Relevance, indicating that the generated questions were both contextually accurate and well-formed. The Diversity score was also strong, ensuring varied and engaging questions for users. Therefore the overall system was demonstrated its ability to produce its diverse, accurate and high-quality questions.



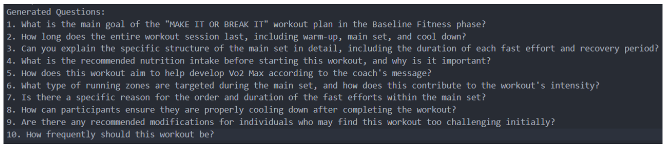
```
                                            Questions Score (%)
0   What is the main goal of the 'MAKE IT OR BREAK...    90.08%
1   How long does the entire workout session last,...    90.65%
2   Can you explain the specific structure of the ...    86.79%
3   What is the recommended nutrition intake befor...    89.87%
4   How does this workout aim to help develop Vo2 ...    81.56%
5   What type of running zones are targeted during...    89.91%
6   Is there a specific reason for the order and d...    88.75%
7   How can participants ensure they are properly ...    89.05%
8   Are there any recommended modifications for in...    88.85%
9                 How frequently should this workout be    72.01%
10  What is the primary goal of 'THE BUBBLE' worko...    80.64%
11  How would you describe the intensity level of ...    87.55%
12  Can you explain the breakdown of the Warm Up s...    89.65%
13  What is the structure of the Main Set in terms...     92.6%
14  What is the purpose of the Coach's Message foc...    90.86%
15  How long is the recommended duration for compl...    91.87%
16  Could you provide insights into the nutrition ...    82.19%
17  Where can one access the annual $80 nutrition ...    73.78%
18  How is the structured training detailed for th...    81.63%
19  Are there any specific recommendations for adj...    84.31%
20  What is the overall goal of 'THE DOCTOR' worko...    80.89%
21  How does the workout purpose of 'adapt' tie in...    87.06%
22  Can you explain the significance of the Warm U...    89.63%
23  What is the rationale behind the main set of r...    83.49%
24  How important is the Cool Down phase in this w...     74.5%
25  What is the expected duration of the entire wo...     89.4%
26  Can you elaborate on the suggested pre-workout...    71.95%
27  What is the main purpose of the 'IT BURNS' wor...    89.76%
28  Can you explain the significance of the Baseli...    90.91%
29  How would you describe the intensity level of ...    89.57%
30  What are the key components of the warm-up ses...    90.38%
31  How long does the main set portion of the work...    93.06%
32  Can you elaborate on the cool-down segment of ...     81.5%
33  What is the expected total duration of the ent...     90.6%
34  Could you explain the recommended pre-workout ...    87.06%
35  How does the coach's message to 'adapt to Vo2 ...    90.35%
36  In the structured training details, what speci...    79.76%
Results saved to evaluation_results.csv
Running teardown with pytest sessionfinish...

2 warnings in 115.69s (0:01:55)
No test cases found, please try again.
```

Fig. 4. Overall G-Eval Output

B. Key Outputs

The system generated contextually relevant and personalized questions by utilizing a wide range of input attributes, beyond just workout descriptions. This approach allowed for the creation of more detailed and targeted questions tailored to the user's needs. Sample questions demonstrated the system's ability to address various aspects of the workout, from technique to recovery, providing users with insightful and customized guidance.



```
Generated Questions:
1. What is the main goal of the "MAKE IT OR BREAK IT" workout plan in the Baseline Fitness phase?
2. How long does the entire workout session last, including warm-up, main set, and cool down?
3. Can you explain the specific structure of the main set in detail, including the duration of each fast effort and recovery period?
4. What is the recommended nutrition intake before starting this workout, and why is it important?
5. How does this workout aim to help develop Vo2 Max according to the coach's message?
6. What type of running zones are targeted during the main set, and how does this contribute to the workout's intensity?
7. Is there a specific reason for the order and duration of the fast efforts within the main set?
8. How can participants ensure they are properly cooling down after completing the workout?
9. Are there any recommended modifications for individuals who may find this workout too challenging initially?
10. How frequently should this workout be?
```

Fig. 5. Generated Questions using Few-Shot + Chain-of-Thought

The developed system offers significant improvements over the traditional manual methods used by Mottiv. By integrating Few-Shot Prompting and Chain-of-Thought prompting, the system was able to generate more accurate, contextually relevant, and logically consistent questions as shown in Fig. 4. Unlike Mottiv's batch-based approach, which often limited the variety and depth of generated questions, the automated system's ability to evaluate each question individually allowed for higher quality outputs. Additionally, by leveraging a broader set of input attributes-such as exercise type, target muscles, and recovery phases, the system produced a rich and more personalized set of questions. This advancement enables the system to address more complex user queries, offering detailed guidance that enabled beyond simple workout descriptions.

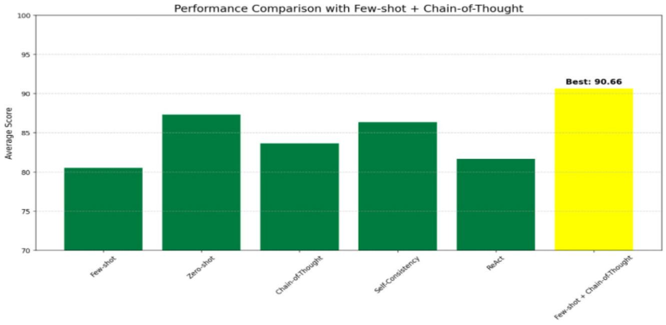


Fig. 6. Comparision between prompting techniques

With respect to these advancements, the system is not without its limitations. While the integration of Few-Shot Prompting and Chain-Of-Thought prompting enhanced performance, the system still depends on the inherent capabilities of the underlying large language models (LLM). In some workout contexts, the generated output questions might not always be optimal. Furthermore, the automated evaluation framework, while consistent, may overlook certain subjective nuances that a human evaluator might notice, such as tone or emotional context. These limitations suggest that further refinement, including domain-specific fine tuning of the LLM and a more advanced feedback loop, could enhance the system's accuracy and adaptability.

In conclusion, the developed system successfully outperforms Mottiv's manual approach in several key areas, including question generation accuracy, contextual relevance, and overall system performance. The integration of Few-shot prompting, Chain-of-Thought prompting, and per-question evaluation has led to significant improvements in question quality, scalability, and usability. While there are some limitations, such as dependence on the LLM's inherent capabilities, the system shows great potential for future expansion, including domain-specific fine-tuning and multilingual support. This system represents a significant step forward in automated question generations and can be further refined to meet the needs of a diverse user base.

References

[1] Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. Prompting is programming: A query language for large language models. *Proceedings of the ACM on Programming Languages*, 7(PLDI):1946–1969, June 2023.

[2] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. Ppt: Pre-trained prompt tuning for few-shot learning, 2022.

[3] Jia Li, Ge Li, Yongmin Li, and Zhi Jin. Structured chain-of-thought prompting for code generation, 2023.

[4] Tianqiao Liu, Zui Chen, Zitao Liu, Mi Tian, and Weiqi Luo. Expediting and elevating large language model reasoning via hidden chain-of-thought decoding, 2024.

[5] Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. Fairness-guided few-shot prompting for large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 43136–43155. Curran Associates, Inc., 2023.

[6] Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. Empowering multi-step reasoning across languages via tree-of-thoughts, 2024.

[7] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.

[8] Jianing Wang, Chengyu Wang, Fuli Luo, Chuanqi Tan, Minghui Qiu, Fei Yang, Qiuhui Shi, Songfang Huang, and Ming Gao. Towards unified prompt tuning for few-shot text classification, 2022.

[9] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action, 2023.

[10] Xiaosong Yuan, Chen Shen, Shaotian Yan, Xiaofeng Zhang, Liang Xie, Wenxiao Wang, Renchu Guan, Ying Wang, and Jieping Ye. Instance-adaptive zero-shot chain-of-thought prompting, 2024.