# MM811 Presentation

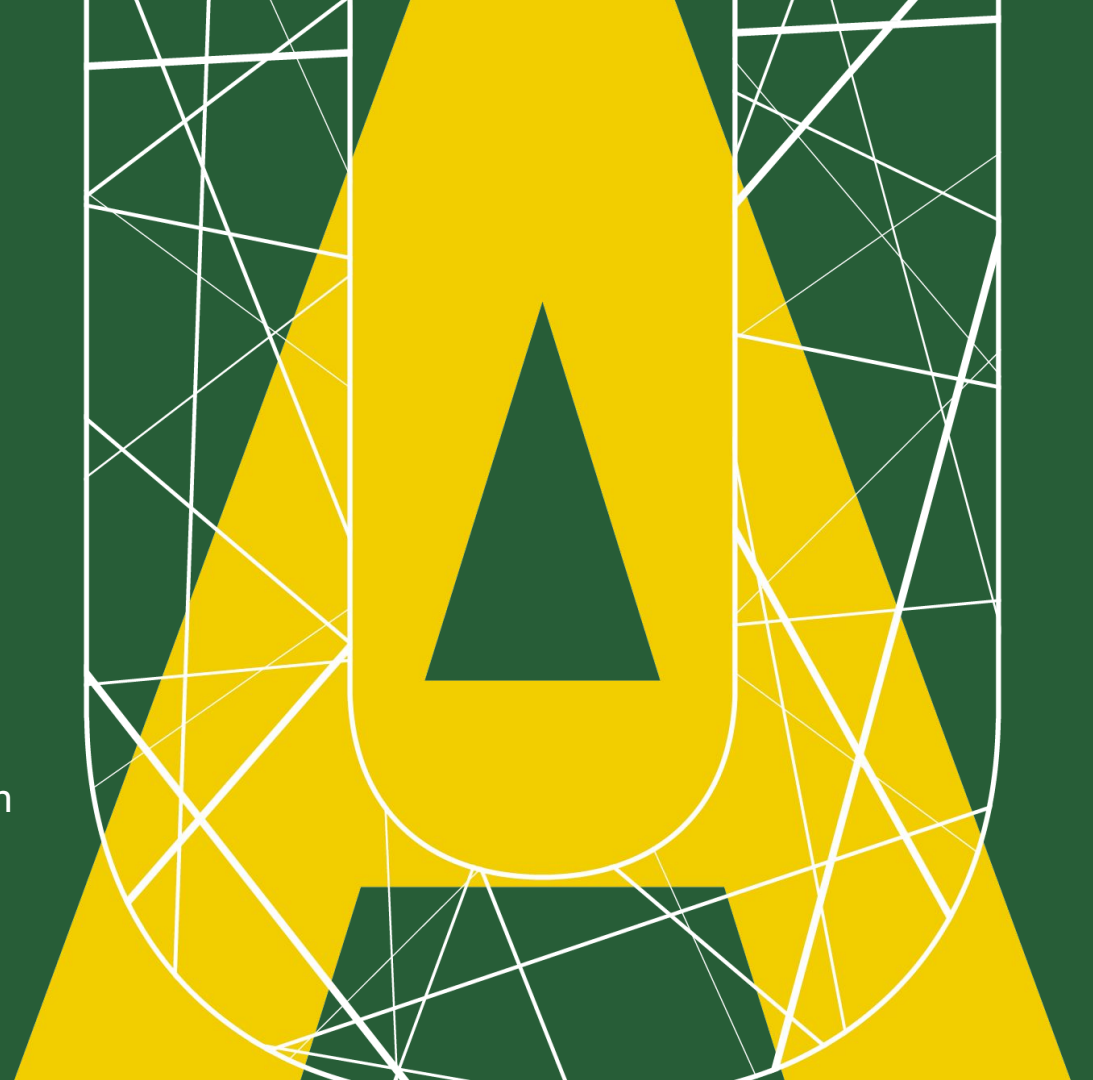
**Topic: Coach MO: Question Generation**

*-Mottiv*

**Team #6**

Raju Bhattarai, Rajapriya, Ruban Gino Singh

UNIVERSITY OF ALBERTA



# Coach Mo : Problem Statement & Challenges

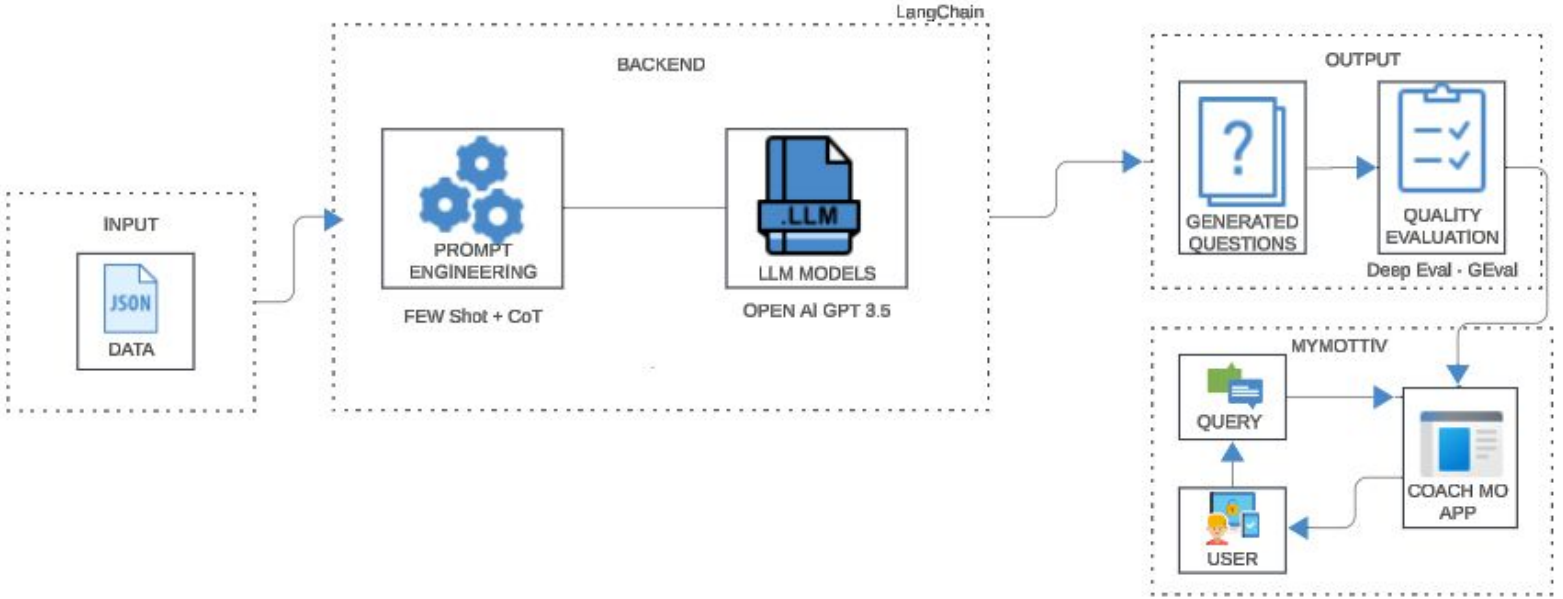
**Problem Statement**

- Athletes and fitness enthusiasts ask detailed, specific questions about workout plans.
- They expect **instant, personalized answers** that align with their fitness goals.
- **Current System Limitation:**
  - Relies on a pre-built database for responding to queries.
  - Provides generic responses if an answer is unavailable, causing user dissatisfaction.

**Challenges**

- **Incomplete Database:**Limited scope of responses leads to **delays and reduced user engagement**.
- **Scalability Issues:**As the user base grows, updating and maintaining the database becomes increasingly difficult.
- **Personalization Gap:**Generic responses fail to meet **user expectations for tailored coaching**.
- **User Retention Risk:**Frustration with delays or generic answers may lead users to switch to alternative platforms.

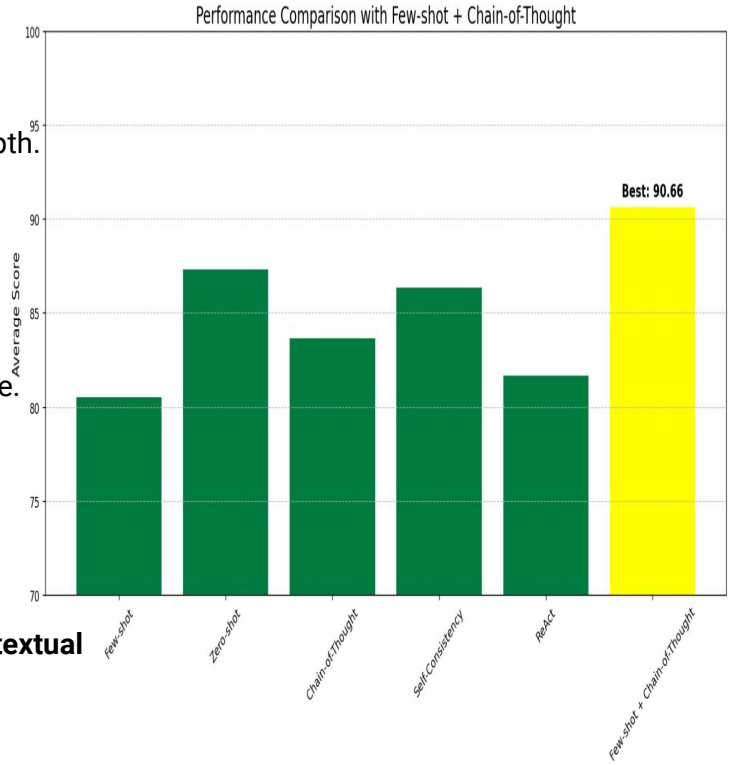
# Implementation



# Evaluation of Prompting Techniques

**Prompting Techniques Overview**

- **Zero-shot Prompting:**
  - Simple task description; generated questions were generic and lacked depth.
- **Few-shot Prompting:**
  - Examples improved relevance but lacked reasoning depth.
- **Chain-of-Thought Prompting:**
  - Enhanced logical structure; occasionally verbose or redundant.
- **ReAct (Reasoning + Acting):**
  - Combined reasoning with action; promising but computationally expensive.




**Selected Approach: Few-shot + Chain-of-Thought Combination**

- Tailored questions align with workout context.
- Effectively addresses user-specific goals.

The **Hybrid Prompting Technique** provides the best balance for **high-quality, contextual question generation**, making it the most effective methodology for Coach Mo.

UNIVERSITY OF ALBERTA

# Introduction to G-Eval Metrics by DeepEval Framework

**What is DeepEval?**

- To evaluate large language models (LLMs) and generative AI systems.
- Includes a set of metrics designed to measure the performance of these systems

**Focus on G-Eval Metrics:**

- Core components of DeepEval.
- In-depth assessment of model-generated content.
- Focuses on natural language understanding, coherence, relevance, and factuality.

**Uses of G-Eval?**

- Provides robust evaluation for AI and LLMs.
- Identifying strengths and weaknesses of the generated content.
- Ensures alignment with real-world use cases.

# Key Features of G-Eval Metrics

**Quality of Generated Content:**

- Assesses the fluency, grammar, and readability of the output.

**Relevance & Coherence:**

- Measures how relevant and contextually appropriate the responses are to the given input.
- Evaluates logical flow and consistency within the generated text.

**Factual Accuracy:**

- Assesses the factuality of the model's response.
- Compares the output against reliable data sources.

**Diversity & Creativity:**

- Measures the level of diversity in the output to avoid repetitive and predictable responses.
- Encourages creativity without sacrificing relevance.

# Why Use G-Eval for This Evaluation?

**Comprehensive Evaluation:**

- Multi-dimensional approach to assessing the accuracy and quality of generated questions.
- Evaluates not just grammar and syntax but also factual correctness, relevance, and alignment.

**Customization for Specific Use Cases:**

- Evaluate specific metrics like correctness and relevance tailored to a particular use case.

**Improving Model Performance:**

- Helps in pinpointing areas of improvement for the AI system.

**Objective and Automated Evaluation:**

- Offers an automated way to evaluate AI-generated questions without requiring manual input, providing a more consistent and scalable approach to evaluation.
- Scoring mechanism in the code allows for consistent grading of generated content.

# G-Eval Compared to other Metrices

| Metrics | Coherence | | Consistency | | Fluency | | Relevance | | AVG | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ |
| ROUGE-1 | 0.167 | 0.126 | 0.160 | 0.130 | 0.115 | 0.094 | 0.326 | 0.252 | 0.192 | 0.150 |
| ROUGE-2 | 0.184 | 0.139 | 0.187 | 0.155 | 0.159 | 0.128 | 0.290 | 0.219 | 0.205 | 0.161 |
| ROUGE-L | 0.128 | 0.099 | 0.115 | 0.092 | 0.105 | 0.084 | 0.311 | 0.237 | 0.165 | 0.128 |
| BERTScore | 0.284 | 0.211 | 0.110 | 0.090 | 0.193 | 0.158 | 0.312 | 0.243 | 0.225 | 0.175 |
| MOVERSscore | 0.159 | 0.118 | 0.157 | 0.127 | 0.129 | 0.105 | 0.318 | 0.244 | 0.191 | 0.148 |
| BARTScore | 0.448 | 0.342 | 0.382 | 0.315 | 0.356 | 0.292 | 0.356 | 0.273 | 0.385 | 0.305 |
| UniEval | 0.575 | 0.442 | 0.446 | 0.371 | 0.449 | 0.371 | 0.426 | 0.325 | 0.474 | 0.377 |
| GPTScore | 0.434 | – | 0.449 | – | 0.403 | – | 0.381 | – | 0.417 | – |
| G-EVAL-3.5 | 0.440 | 0.335 | 0.386 | 0.318 | 0.424 | 0.347 | 0.385 | 0.293 | 0.401 | 0.320 |
| - Probs | 0.359 | *0.313* | 0.361 | *0.344* | 0.339 | *0.323* | 0.327 | *0.288* | 0.346 | *0.317* |
| G-EVAL-4 | **0.582** | **0.457** | **0.507** | **0.425** | **0.455** | **0.378** | **0.547** | **0.433** | **0.514** | **0.418** |
| - Probs | 0.560 | *0.472* | 0.501 | *0.459* | 0.438 | *0.408* | 0.511 | *0.444* | 0.502 | *0.446* |
| - CoT | 0.564 | 0.454 | 0.493 | 0.413 | 0.403 | 0.334 | 0.538 | 0.427 | 0.500 | 0.407 |

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, & Chenguang Zhu. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment.

# Workout Data


```
[
    {
        "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
    }
]
```

# Code Implementation

```python
def generate_questions(prompt, model="gpt-3.5-turbo", max_tokens=200):
    #few-shot + chain-of-thought
    response = openai.ChatCompletion.create(
        model=model,
        messages=[
            {"role": "system", "content": "You are a fitness enthusiast looking to get more details about workout plans."+
            "Use step-by-step reasoning to think about the key aspects of the workout plan and generate thoughtful questions about it."},
            # Few-shot + chain-of-thought example 1
            {"role": "user",
                "content": "Title: BEGINNER STRENGTH WORKOUT\nDescription: Warm Up: 5 minutes light jogging."+
                "Main Set: 3 rounds of push-ups (10 reps), squats (15 reps), and lunges (10 reps per leg). Cool Down: 5 minutes stretching."},
            {"role": "assistant", "content": "Let's break this workout into components: warm-up, main set, and cool-down."+
            "\n1. Warm-Up: Should the jogging pace be light for beginners? Are there alternative warm-ups?"+
            "\n2. Main Set: How can push-ups be modified for someone with limited upper body strength? How much rest should there be between rounds?"+
            "Are squats suitable for beginners with joint issues?\n3. Cool-Down: What stretches should be included in the cool-down?"+
            "How long should each stretch be held?"},
            # Few-shot + chain-of-thought example 2
            {"role": "user", "content": "Title: CARDIO BLAST\nDescription: Warm Up: 10 minutes of brisk walking."+
            "Main Set: 20 minutes alternating between 1 minute of sprinting and 2 minutes of slow jogging. Cool Down: 5 minutes of walking."},
            {"role": "assistant", "content": "I'll analyze this workout step by step:"+
            "\n1. Warm-Up: What pace is recommended for the brisk walking? Should beginners adjust the time?"+
            "\n2. Main Set: What intensity level should be aimed for during the sprinting? How can the jogging pace support proper recovery?"+
            "\n3. Cool-Down: Are stretches recommended after this workout? Can the cool-down walking time be extended?"},
            # Target prompt
            {"role": "user", "content": prompt}
        ],
        max_tokens=max_tokens,
        temperature=0.7
    )
    return response['choices'][0]['message']['content'].strip()
```

UNIVERSITY OF ALBERTA

# Results

| | 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.
```

UNIVERSITY OF ALBERTA

# Novelty

## Mottiv Vs Our System

| 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 |

# Discussion

## Key Contributions:

- **Few-shot + Chain-of-Thought prompting**: Enhances reasoning and quality.
- **85% Quality Benchmark**: Sets a new standard.
- **Per-question evaluation**: Ensures accuracy and reliability.

## Impact of our project:

- Higher accuracy, relevant and diverse questions.
- Enriched question database
- Cost savings
- Reduced waiting time

# Thank You