Report: Analysis of AI Models for Text Summarization and Question Generation

Abstract

This study evaluates the performance of four AI models — *qwen/qwen3-8b*, *google/gemini-flash-1.5-8b*, *deepseek/deepseek-r1-0528-qwen3-8b:free*, and *openai/gpt-4.1-nano* — in two key tasks: text summarization and multiple-choice question generation. The analysis focuses on metrics such as clarity, relevance, and creativity. Additionally, the impact of prompt engineering on question generation quality is explored.

Introduction

Background

AI language models are increasingly used for educational and research purposes, including summarization and question generation. Evaluating their effectiveness helps identify optimal models for specific tasks.

Study Questions

1. How do different AI models perform in summarizing scientific text?
2. Which model generates the most effective multiple-choice questions for learning?
3. Does prompt engineering improve question quality?

Hypothesis

- Models with larger parameter sizes (e.g., *gemini-flash-1.5-8b*) will produce clearer and more relevant summaries.
- Creative prompting will enhance question generation by encouraging deeper engagement with the text.

Approach

- Summarization Task: Compare outputs of four models on a fixed astronomy passage.
- Question Generation: Test models in two modes:
  - Standard prompting.
  - Teacher-style prompting (creative, non-memorization-focused).

Methods

Experimental Design

1. Summarization:
   - Input: A paragraph about supernovae.
   - Output: Summaries from each model, rated for length, clarity, and relevance.
2. Question Generation:
   - Phase 1: Standard prompt ("Generate 5 MCQs").
   - Phase 2: Enhanced prompt ("Act as a teacher creating creative questions").
   - Evaluation criteria: Creativity (1–5), clarity, relevance.

Data Analysis

- Qualitative comparison of outputs.
- Tabular scoring (e.g., *gemini-flash-1.5-8b* scored 4/5 for creativity).

Results

Summarization Performance

Model	Length	Clarity	Relevance
qwen3-8b	Mid	Partially clear	Mid
gemini-flash	Mid	Clear	High
deepseek-r1	Long	Clear	Mid
gpt-4.1-nano	Mid	Clear	Mid

Key Finding: *gemini-flash* produced the clearest and most relevant summary.

Question Generation

Standard Prompting

- Best Model: *deepseek-r1* (creativity: 5/5).
- Weakness: *qwen3-8b* questions were partially unclear.

Enhanced Prompting

- All models improved, especially *gemini-flash* and *gpt-4.1-nano* (creativity: 5/5).
- Example: Gemini’s question linked supernovae to climate effects, fostering critical thinking.

Discussion

Support for Hypothesis

- Larger models (*gemini-flash*, *gpt-4.1-nano*) excelled in clarity and relevance.
- Creative prompts significantly improved question quality (e.g., deeper analytical questions).

Unexpected Observations

- *deepseek-r1* generated overly long summaries but highly creative questions.

Future Studies

- Test models on diverse text genres (e.g., humanities, technical).
- Quantify improvements from prompt engineering statistically.

Conclusion

For summarization, *gemini-flash-1.5-8b* is optimal. For question generation, creative prompting with *gemini-flash, deepseek/deepseek-r1* or *gpt-4.1-nano* yields the best results.

Appendices

- https://github.com/nuriddinovN/hogwarts-edai-project/blob/main/research/Abdullokhon's_API_week2_research.ipynb