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

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*`—across three tasks: text summarization, multiple-choice question generation, and document text extraction and analysis. The analysis focuses on metrics such as clarity, relevance, creativity, and adaptability to different input formats (PDFs, images, text files).

Introduction

Background

AI language models are increasingly used for educational, research, and document-processing tasks. Evaluating their effectiveness helps identify optimal models for specific use cases, from summarization to interactive learning tools.

Study Questions

1. How do different AI models perform in summarizing scientific and personal narrative text?
2. Which model generates the most effective multiple-choice questions for learning?
3. Can models accurately extract and analyze text from uploaded documents (PDFs, images)?
4. Does prompt engineering improve output quality across tasks?

Hypothesis

- Larger models (e.g., `*gemini-flash-1.5-8b*`, `*gpt-4.1-nano*`) will excel in clarity and relevance.
- Creative prompting will enhance question generation by fostering critical thinking.
- Document processing pipelines (OCR + AI) will yield usable outputs but may vary in accuracy.

Approach

- Summarization Task: Compare outputs on fixed astronomy and personal narrative texts.
- Question Generation: Test models in standard and teacher-style prompting modes.
- Document Processing: Extract text from uploaded files (PDF/image) and analyze with AI-generated summaries/questions.

Methods

Experimental Design

1. Summarization:
   - Input: Astronomy passage + personal narrative (uploaded PDF).
   - Output: Summaries 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").

3. Document Processing:
   - Text Extraction: `pytesseract` (images), `pdfplumber` (PDFs).
   - Analysis: Generated summaries/questions from extracted text.

Data Analysis

- Qualitative comparison of outputs.
- Tabular scoring (e.g., creativity: 1–5, clarity: low/mid/high).

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	High

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*`, `*deepseek-r1*` and `*gpt-4.1-nano*` (creativity: 5/5).
- Example: Gemini's question linked AI limitations to practical learning.

Document Processing

- Text Extraction: Successfully parsed PDFs/images using `*pytesseract*` and `*pdfplumber*`.
- Analysis Outputs:
  - *Qwen3-8B*: Focused on narrative arc (e.g., "How did early failures shape your perspective?").
  - *Gemini Flash*: Highlighted practical tool applications (e.g., "How did PyTorch transform your understanding?").
  - *GPT-4.1-nano*: Balanced summary with actionable questions (e.g., "How do resource limitations motivate innovation?").

Discussion

Support for Hypothesis

- Larger models (`*gemini-flash*`, `*gpt-4.1-nano*`) excelled in clarity and relevance across tasks.
- Creative prompts significantly improved question quality (e.g., deeper analytical questions).
- Document processing worked reliably but required clean input files.

Unexpected Observations

- `*deepseek-r1*` generated overly long summaries but highly creative questions.
- OCR accuracy dropped with low-quality images, affecting downstream AI analysis.

Future Studies

- Test models on diverse text genres (e.g., technical manuals, humanities).
- Quantify improvements from prompt engineering statistically.
- Optimize OCR preprocessing for noisy inputs.

Conclusion

For summarization, `*gemini-flash-1.5-8b*` is optimal. For question generation, creative prompting with `*gemini-flash*` or `*gpt-4.1-nano*` yields the best results. Document processing pipelines are viable but depend on input quality.

Literature Cited

- OpenAI API documentation.
- Google Generative AI technical reports.

Appendices

https://github.com/nuriddinovN/hogwarts-edai-project/blob/main/research/Abdullokhon's_research_week3.ipynb