Final Report: Benchmarking LLMs for Summarization & Flashcard Generation

Task Objective:

Evaluate multiple language models (API-based and Hugging Face) on the task of:

1. Summarizing a detailed biography of **Nikola Tesla**
2. Generating 5 flashcard-style questions and answers

Models Benchmarked

Hugging Face (local execution):

- *T5-base*
- *T5-large*
- *mT5-base*

API-based via OpenRouter:

- *DeepSeek*
- *Qwen3-8B*
- *Mistral-Small-3.2*

Prompt Used

> You are the most creative teacher on earth.
> Read the following: *(full Tesla biography text)*
>
> Now:
>
> 1. Summarize this text in 3 sentences.
> 2. Generate 5 flashcard questions WITH answers.

Results Summary

Model	Time (s)	Memory (MB)	Summary Accuracy	Flashcard Quality	Relevance	Notes
T5-base	3.90	472.03	Low	Poor	Low	Generic summary, no questions
T5-large	4.59	0.9	Low	Very Poor	Low	Repetitive text, failed generation
mT5-base	0.38	1.56	Very Poor	None	Very Low	Output: “Tesla. Tesla.” only
DeepSeek	6.87	0	High	Excellent	High	Very clear, structured, engaging
Qwen3-8B	7.66	0	High	Very Good	High	Balanced and fact-rich
Mistral-Small-3.2	6.24	0	High	Very Good	High	Concise, readable, highly relevant

Analysis by Category

Accuracy of Summary

- **DeepSeek**, **Qwen3-8B**, and **Mistral-Small-3.2** provided faithful, readable, and complete 3-sentence summaries.
- **T5 models** struggled, with mT5-base failing completely.

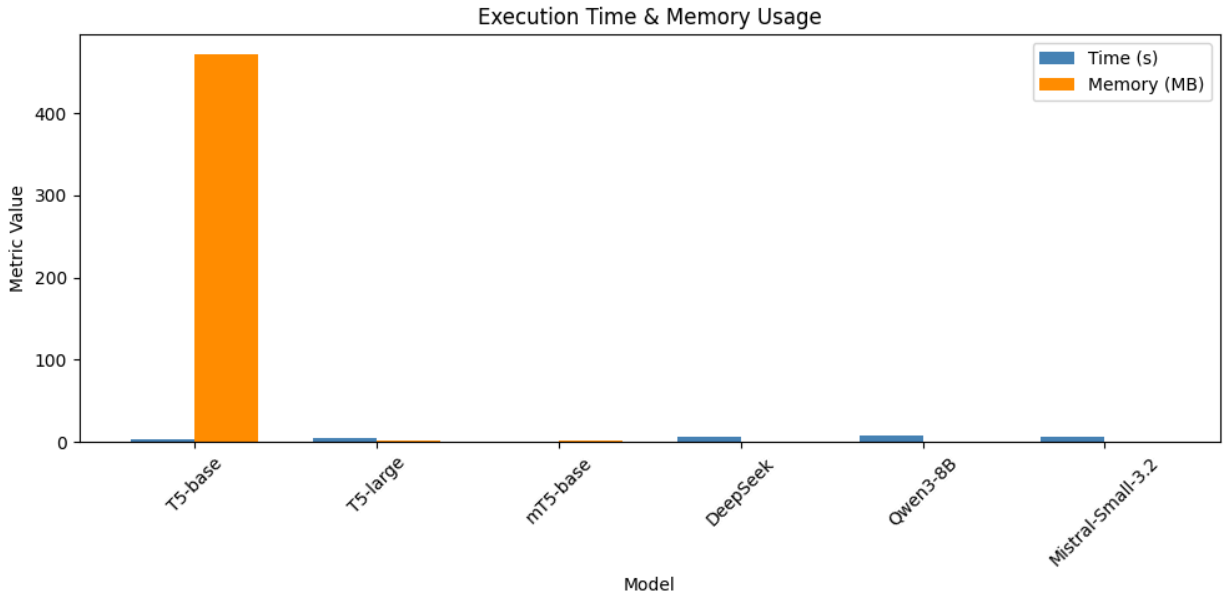
Flashcard Relevance

- **DeepSeek** and **Mistral-Small-3.2** produced engaging and well-aligned Q&A pairs.
- **Qwen3-8B** showed thoughtful question formulation grounded in the text.

Runtime & Memory

- **HF models** used significant local memory (T5-base ~472MB) but were fast.
- **API models** returned slightly slower responses but no local memory impact.

Chart: Time & Memory Usage



Note: API models show 0 memory usage because that load is offloaded to OpenRouter.

Conclusion & Recommendation

Based on output quality, consistency, and performance:

Best Overall:

DeepSeek – creative, clear, and pedagogically sound.

Best Balance:

Qwen3-8B – well-written, fact-rich, and technically solid.

Avoid for this task:

- T5-large, mT5-base – unreliable, malformed or empty outputs

Files Available

- Abdullokhon's_week5_research_model_metrics.csvv: Raw metrics table
- Abdullokhon's_week5_research.ipynb: Notebook where tests were conducted