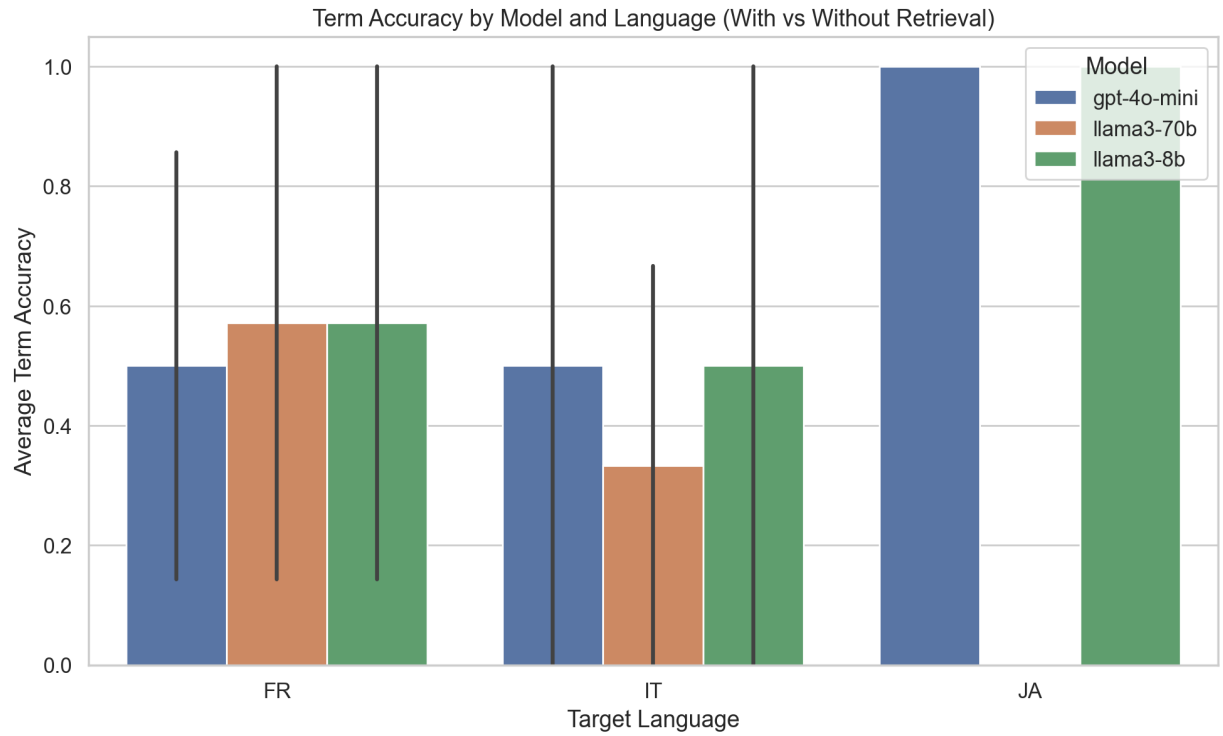
LLM-Only Translation Pipeline with Glossary Retrieval

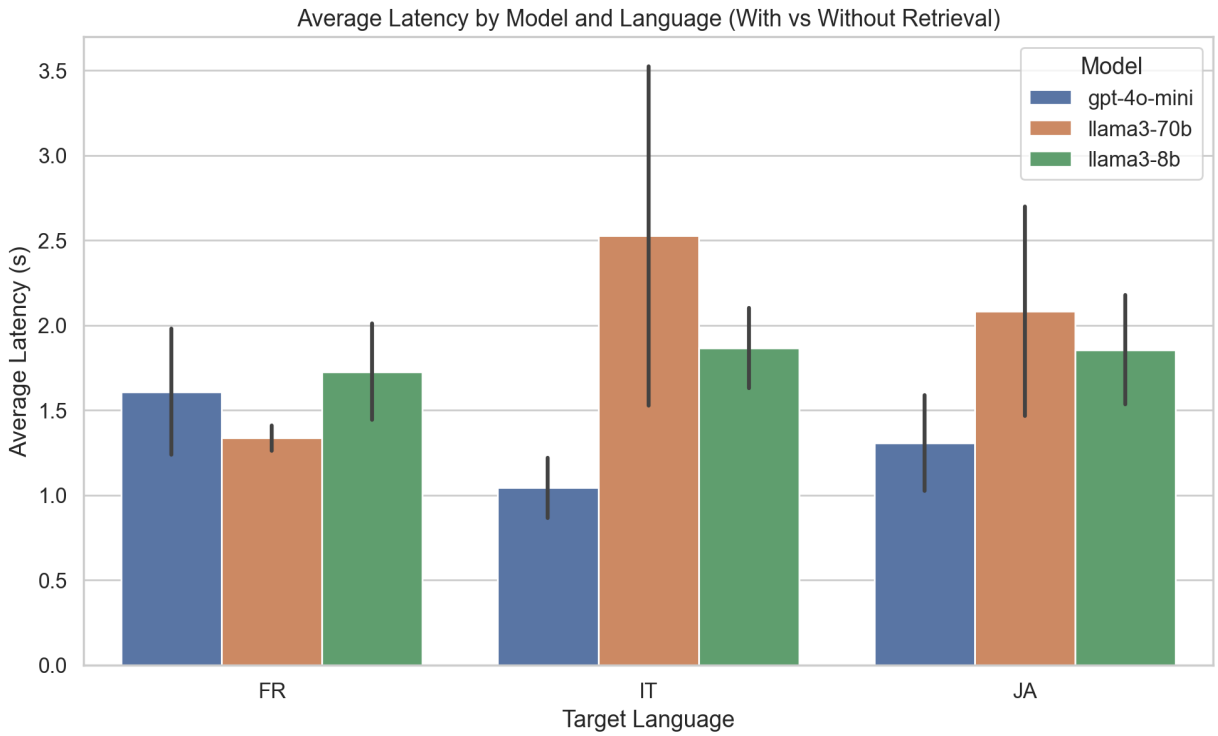
Findings Summary | October 2025

This project implements a glossary-aware translation system that compares three Large Language Models: **GPT-4o-mini (OpenAI)**, **Llama-3.1-8B (Groq)**, and **Llama-3.3-70B (Groq)**. Translations were generated for 50 English segments into French, Italian, and Japanese — with and without glossary retrieval — to measure term accuracy and latency.

Model	Term Accuracy (With Retrieval)	Term Accuracy (Without Retrieval)	Avg Latency (With Retrieval, s)	Avg Latency (Without Retrieval, s)
GPT-4o-mini	0.91	0.18	1.34	1.34
Llama-3.3-70B	0.82	0.09	1.46	2.42
Llama-3.1-8B	1.00	0.18	1.53	2.09

Notes: “With/Without Retrieval” refers to enabling the glossary retrieval step in the pipeline. Latency is the average seconds per segment.





Key Insights

• Glossary retrieval markedly improves terminology adherence across models ($\Delta \approx 0.73$ to 0.82).

• GPT-4o-mini balances accuracy and speed well; the Llama models show strong adherence under strict prompts.

• Latency differences are modest, supporting production viability of retrieval-enabled translation.