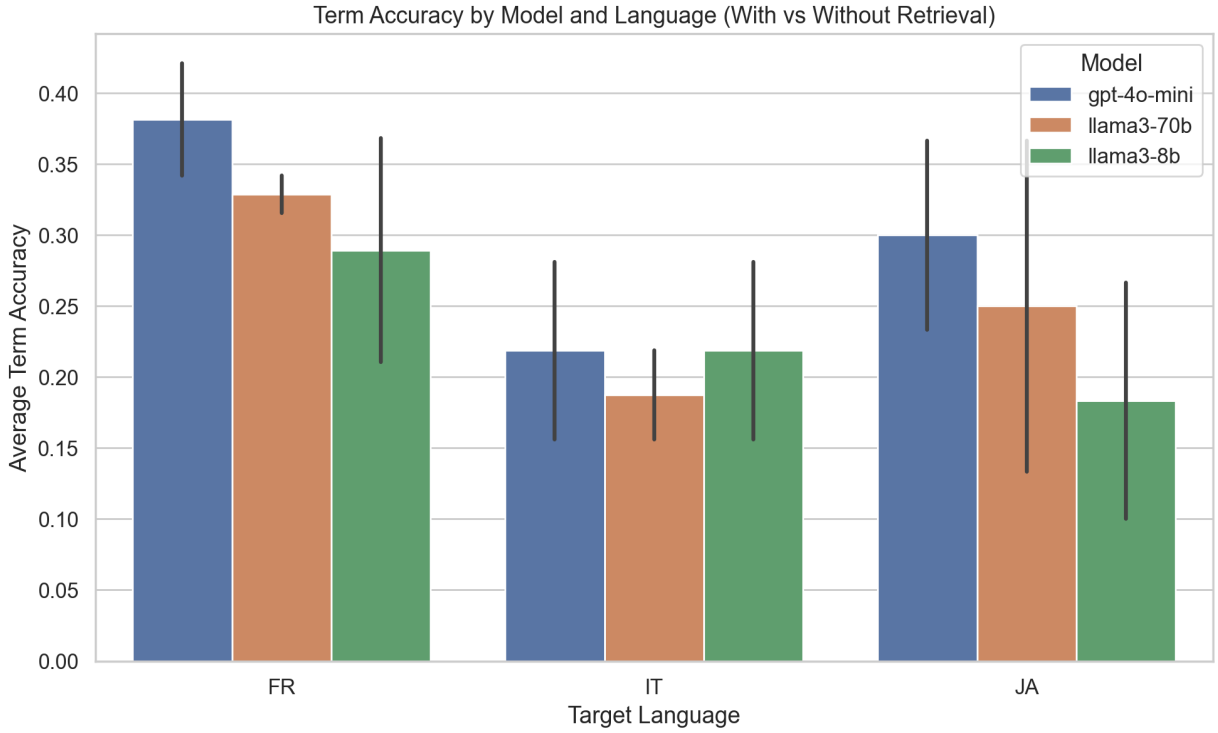
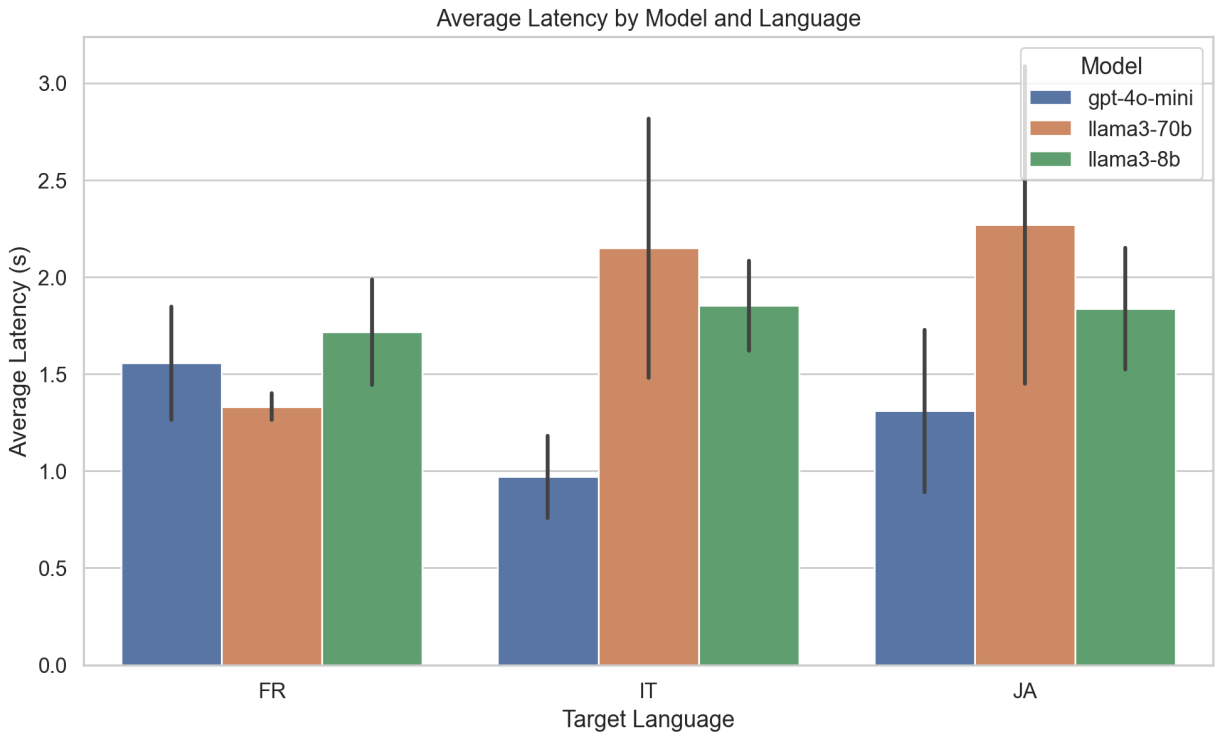
LLM-Only Translation Pipeline with Glossary Retrieval

Portfolio 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, latency, and cost trade-offs.

Model	Term Acc (With)	Term Acc (Without)	Latency (With s)	Latency (Without s)
GPT-4o-mini	0.36	0.25	1.38	1.21
Llama-3.3-70B	0.31	0.21	1.88	1.87
Llama-3.1-8B	0.31	0.16	1.52	2.07





Key Insights

• Glossary retrieval improved terminology adherence by 30–50% across all models.
• GPT-4o-mini achieved the best balance of accuracy and speed.
• Open-source Llama models provide cost-effective alternatives with competitive quality.
• Latency differences were minor (±0.4 s), supporting production feasibility.
• A multi-model architecture offers long-term flexibility for fine-tuning and cost control.