LLM-Based Summarization Pipeline

Long Document	Prompt Engineering	LLM Processing	Decoding Control	Summary Output
Input text (up to 100K tokens)	System prompt + few-shot examples	GPT-3.5/4, Claude, LLaMA, Mistral	Temperature, top-p, max_tokens	Concise summary (target length)