

Fig 15. LLM Model Specifications

Model	Provider	Release	Architecture	Context	API
Gemini 2.5 Flash	Google	May 2025	MoE	1M	Gemini API
DeepSeek V3.2	DeepSeek	Jun 2025	671B MoE	128K	DashScope
Qwen3 235B	Alibaba	Jul 2025	235B MoE	128K	DashScope
GPT-5 mini	OpenAI	Aug 2025	Undisclosed	128K	OpenAI API