



Provider	Model	Available in askme	Status	Comments	Free?
Google	gemini-1.0-pro	Yes	DEPRECATED	REMOVE - Returns 404	Yes
Google	gemini-pro	Yes	DEPRECATED	REMOVE - Returns 404	Yes
Google	gemini-1.5-flash	Yes	Active	Keep - Stable	Yes
Google	gemini-1.5-flash-8b	Yes	Active	Keep - Stable	Yes
Google	gemini-1.5-pro	Yes	Active	Keep - Deprecating Sept 2025	Yes
Google	gemini-2.5-pro	No	NEW	ADD - #1 LMArena model	Yes
Google	gemini-2.5-flash	No	NEW	ADD - Replace deprecated	Yes
Google	gemini-2.0-flash	No	NEW	ADD - Multimodal outputs	Yes
Mistral	mistral-small-latest	Yes	Active	Keep - No issues	Yes
Mistral	open-mistral-7b	Yes	Active	Keep - Open source	Yes
Mistral	open-mixtral-8x7b	Yes	Active	Keep - Good performance	Yes
Mistral	open-mixtral-8x22b	Yes	Active	Keep - High capacity	Yes
Mistral	mistral-medium-latest	Yes	Active	Optional: upgrade to mistral-medium-3	Yes
Mistral	mistral-medium-3	No	NEW	ADD - 90% Claude performance	Yes
Mistral	magistral-small	No	NEW	ADD - Reasoning model	Yes
Mistral	magistral-medium	No	NEW	ADD - Advanced reasoning	Yes
Llama	Meta-Llama-3-8B-Instruct-Turbo	Yes	Active	Keep - Fast inference	Yes
Llama	Llama-3-8b-chat-hf	Yes	Active	Keep - Standard chat	Yes
Llama	Meta-Llama-3-70B-Instruct-Turbo	Yes	DEPRECATING	REPLACE with Llama-3.3-70B-Instruct	Yes
Llama	Llama-2-7b-chat-hf	Yes	Active	Optional: upgrade to Llama-3-8b-chat-hf	Yes
Llama	Llama-2-13b-chat-hf	Yes	Active	Optional: upgrade to newer Llama-3 models	Yes
Llama	Llama-3.3-70B-Instruct	No	NEW	ADD - Replace 70B-Turbo	Yes
Llama	Llama-4-Maverick	No	NEW	ADD - 400B MoE, 9-23x cheaper	Yes
Llama	Llama-4-Scout	No	NEW	ADD - 10M context length	Yes

Priority Actions:

- **URGENT:** Replace Google deprecated models:
 - gemini-1.0-pro → gemini-2.5-flash
 - gemini-pro → gemini-2.5-pro

-  **PLAN:** Replace Llama-3-70B-Instruct-Turbo → Llama-3.3-70B-Instruct (by June 30, 2025)
-  **OPTIMIZE:** Add new high-performance models for cost/capability gains