3437766 - Availability of Generative AI Models

Component: CA-ML-AIC (Cross-Application Components > Machine Learning > SAP AI Core), Version: 86, Released On: 22.05.2025

Symptom

This SAP Note outlines the available models and their versions in different data centers, along with their deprecation dates. Furthermore, it details the conversion rates for GenAI Tokens for each model, and provides information on rate limits.

Solution

The information in the following table is subject to changes, in particular with respect to general availability and price. Futher information is provided after the table.

Executable ID (Access Type)	Model	Version	GenAI Input Tokens (for 1,000 Model Tokens)	GenAI Output Tokens (for 1,000 Model Tokens)	Available in Orchestration	Deprecated	Suggested Replacement	Retirement Date	Region Availability (SAP AI Core)													Rate Limit (req/min/tenant)
									AP10	AP11	AP20	AP30	EU10	EU11	EU20	EU30	JP10	US10	US21	US30	SA30	
aicore-mistralai (sap-hosted)	mistralai--mistral-large-instruct	2407 (latest)	0.00112	0.00320	yes			not earlier than 2025-06-30	Yes*				Yes	Yes	yes	yes	Yes*	Yes*	yes*	yes*		100
aicore-mistralai (sap-hosted)	mistralai--mistral-small-instruct	2503 (latest)	0.00005	0.00015	yes			not earlier than 2025-09-30	Yes*	Yes*	Yes*	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes*	Yes*	Yes*	Yes*	100
aicore-ibm (sap-managed)	ibm--granite-13b-chat	2.1.0 (latest)	0.00032	0.00032	yes			not earlier than 2025-06-30	Yes*				Yes	-	-	-	Yes*	Yes	-	-		100
aicore-opensource (sap-managed)	meta--llama3.1-70b-instruct	202409 (latest)	0.00115	0.00217	yes			not earlier than 2025-06-30	-				Yes	-	-	-	-	-	-	-		100

aicore-opensource (sap-managed)	mistralai--mixtral-8x7b-instruct-v01	null (latest)	0.00023	0.00038	yes	true		not earlier than 2025-05-30	-				Yes	-	-	-	-	-	-	-	yes	100
aicore-nvidia (sap-managed)	nvidia--llama-3.2-nv-embedqa-1b	2 (latest)		-				not earlier than 2025-09-30	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		138
aws-bedrock (AWS)	amazon--titan-embed-text	1.2	0.00014	-					Yes*				Yes*	Yes	Yes*	Yes*	Yes*	Yes	Yes	Yes		50
aws-bedrock (AWS)	amazon--titan-embed-text	2 (latest)	0.00014	-					Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		138
aws-bedrock (AWS)	amazon--titan-text-express	1 (latest)	0.00086	0.00158	yes				Yes		Yes		Yes	Yes	Yes	Yes	Yes*	Yes	Yes	Yes		50
aws-bedrock (AWS)	amazon--titan-text-lite	1 (latest)	0.00033	0.00040	yes				Yes		Yes		Yes	Yes	Yes	Yes	Yes*	Yes	Yes	Yes		50
aws-bedrock (AWS)	amazon--nova-pro	1 (latest)	0.00312	0.00920	yes				-				-	-	-	-	-	Yes	Yes	Yes		50
aws-bedrock (AWS)	amazon--nova-lite	1 (latest)	0.00312	0.00920	yes				-				-	-	-	-	-	Yes	Yes	Yes		50
aws-bedrock (AWS)	amazon--nova-micro	1 (latest)	0.00086	0.00158	yes				-				-	-	-	-	-	Yes	Yes	Yes		50
aws-bedrock (AWS)	anthropic--claude-3-haiku	1 (latest)	0.00024	0.00089	yes				Yes		Yes		Yes*	Yes	Yes*	Yes*	Yes*	Yes	Yes	Yes		50

aws-bedrock (AWS)	anthropic--claude-3-sonnet	1 (latest)	0.00204	0.00988	yes				Yes		Yes		Yes	Yes	Yes	Yes	Yes*	Yes	Yes	Yes		50
aws-bedrock (AWS)	anthropic--claude-3.5-sonnet	1	0.00204	0.00988	yes				Yes	Yes*			Yes	Yes	Yes*	Yes*	Yes*	Yes	Yes	Yes		50
aws-bedrock (AWS)	anthropic--claude-3.5-sonnet	2 (latest)	0.00204	0.00988	yes				-				-	-	-	-	-	Yes	Yes	Yes		50
aws-bedrock (AWS)	anthropic--claude-3.7-sonnet	1	0.00204	0.00988	yes								yes		yes	yes		yes	yes	yes		100
aws-bedrock (AWS)	anthropic--claude-3-opus	1 (latest)	0.00988	0.04913	yes				Yes*				Yes*	-	Yes*	Yes*	Yes*	Yes	Yes	Yes		50
azure-openai (Azure)	text-embedding-3-large	1 (latest)	0.009	-				not earlier than 2026-01-25	Yes*	Yes*	Yes*	Yes	Yes*	Yes	Yes*	Yes*	Yes	Yes	Yes	Yes	Yes*	138
azure-openai (Azure)	text-embedding-3-small	1 (latest)	0.002	-				not earlier than 2026-01-25	Yes*				Yes*	Yes	Yes*	Yes*	Yes*	Yes	Yes	Yes		138
azure-openai (Azure)	text-embedding-ada-002	2 (latest)	0.007	-		yes	text-embeddding-3-small text-embeddding-3-large	not earlier than 2025-10-03	Yes	Yes*	Yes*	Yes*	Yes	Yes	Yes	Yes	Yes	Yes	Yes*	Yes*	Yes*	600 (US10, EU10), 138* all other landscapes
azure-openai (Azure)	gpt-35-turbo	1106 (latest)	0.00094	0.00122	yes	yes	gpt-4o-mini	not earlier than 2025-05-31	Yes*				Yes	Yes	Yes	Yes	Yes*	Yes	Yes	Yes		120

azure-openai (Azure)	gpt-35-turbo-0125	0125 (latest)	0.00037	0.00097	yes	yes	gpt-4o-mini	not earlier than 2025-05-31	-				Yes*	-	-	-	-	Yes	-	-		120
azure-openai (Azure)	gpt-4	0613	0.01735	0.03462	yes	yes	gpt-4o	2025-06-06	Yes	Yes*	Yes*	Yes*	Yes	Yes	Yes	Yes	Yes	Yes	Yes*	Yes*	Yes*	78
azure-openai (Azure)	gpt-4-32k	0613 (latest)	0.03462	0.06917	yes	yes	gpt-4o	2025-06-06	Yes	Yes*	Yes*	Yes*	Yes	Yes	Yes	Yes	Yes	Yes	Yes*	Yes*	Yes*	78
azure-openai (Azure)	gpt-4o	2024-05-13	0.00312	0.00920	yes			not earlier than 2025-06-30	Yes*	Yes*	Yes*	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	Yes	Yes	Yes*	78
azure-openai (Azure)	gpt-4o	2024-08-06 (latest)	0.00159	0.00616	yes			not earlier than 2025-08-06	-				Yes	Yes	Yes	Yes	-	Yes	Yes	Yes		78
azure-openai (Azure)	gpt-4o	2024-11-20			yes								Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		78
azure-openai (Azure)	gpt-4o-mini	2024-07-18	0.00009	0.00039	yes			not earlier than 2025-07-18	Yes*			Yes	Yes	Yes	Yes	Yes	Yes*	Yes	Yes	Yes	Yes*	120
azure-openai (Azure)	gpt-4	turbo-2024-04-09 (latest)	0.00616	0.01833	yes			not earlier than 2025-06-06	Yes*				Yes	Yes	Yes*	Yes*	Yes*	Yes	Yes	Yes		78
azure-openai (Azure)	gpt-4.1	2025-04-14	0.00129	0.00494	yes				Yes*	Yes*	Yes*	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	Yes	Yes		78
azure-openai (Azure)	gpt-4.1-mini	2025-04-14	0.00026	0.00099	yes				Yes*	Yes*	Yes*	Yes*					Yes*	Yes	Yes	Yes		120

azure-openai (Azure)	gpt-4.1-nano	2025-04-14	0.00008	0.00026	yes				Yes*	Yes*	Yes*	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	Yes	Yes		120
azure-openai (Azure)	o1	2024-12-17	0.00920	0.03658	yes			not earlier than 2025-12-17	Yes*	Yes*	Yes*	Yes*					Yes*	Yes*	Yes*	Yes*		78
azure-openai (Azure)	o3-mini	2025-01-31	0.00069	0.00270	yes				Yes*	Yes*	Yes*	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	Yes	Yes		120
azure-openai (Azure)	o3	2025-04-16	0.00610	0.02436	yes				Yes*	Yes*	Yes*	Yes*					Yes*	Yes*	Yes*	Yes*		78
azure-openai (Azure)	o4-mini	2025-04-16	0.00069	0.00270	yes				Yes*	Yes*	Yes*	Yes*					Yes*	Yes*	Yes*	Yes*		120
gcp-vertexai (Google)	gemini-1.5-flash	001	0.00007 (<128k tokens) 0.00012 (>= 128k tokens)	0.00021 (< 128k context) 0.00040 (>= 128k context)	yes	yes	gemini-2.0-flash-001 gemini-2.0-flash-lite-001	2025-05-24	Yes*		yes		Yes	-	Yes	Yes	Yes*	Yes	Yes	Yes		100
gcp-vertexai (Google)	gemini-1.5-flash	002 (latest)	0.00007 (<128k tokens) 0.00012 (>= 128k tokens)	0.00021 (< 128k context) 0.00040 (>= 128k context)	yes				-				Yes	-	Yes	Yes	-	Yes	-	Yes		100

gcp-vertexai (Google)	gemini-1.5-pro	001	0.000087 (<128k tokens) 0.000167 (>= 128k tokens)	0.000327 (< 128k context) 0.000647 (>= 128k context)	yes	yes	gemini-2.0-flash-001 gemini-2.0-flash-lite-001	2025-05-24	Yes*		Yes		Yes	-	Yes	Yes	Yes*	Yes	Yes	Yes		100
gcp-vertexai (Google)	gemini-1.5-pro	002 (latest)	0.000087 (<128k tokens) 0.000167 (>= 128k tokens)	0.000327 (< 128k context) 0.000647 (>= 128k context)	yes				-				Yes	-	Yes	Yes	-	Yes	-	Yes		100
gcp-vertexai (Google)	gemini-2.0-flash	001 (latest)	0.000012	0.000040	yes								yes		yes	yes		yes	yes	yes		100
gcp-vertexai (Google)	gemini-2.0-flash-lite	001 (latest)	0.000007	0.000021	yes								yes		yes	yes		yes	yes	yes		100
aicore-alephalpha (sap-hosted)	alephalpha-pharia-1-7b-control	202411 (latest)	0.000018	0.000027	yes (orchestration-only)								Yes									100
aicore-opensource	deepseek-ai--deepseek-r1	202502 (latest)	0.000060	0.000305	yes(orchestration-only)					yes*							yes*					10

* Models with asterisks might be accessed from a cross-border location. This setup harmonizes the model offering across data centers and allows for load balancing and failovers, which help to mitigate data center outages. Data centers are selected based on model availability and data center proximity.

For example, AWS Bedrock Titan Embed Light in region EU10 is likely to be accessed within the EU10 region, but data centers outside of EU10 may be used for the purposes outlined.

The same model in region US10 uses only data centers within that region.

Compliance

It is your responsibility to ensure that your use of AI models made available on SAP Generative AI Hub comply with applicable laws, regulations, rules and corporate policies.

Note: this statement does not modify SAP's obligations under your agreement for the use of SAP Generative AI Hub.

Access Type

Models available on Generative AI Hub are accessed via remote providers, such as AWS Bedrock, or hosted by SAP AI Core.

SAP Hosted (sap-hosted): Model(s) hosted on SAP's own infrastructure, managed by SAP AI Core.

SAP Managed (sap-managed): Model(s) hosted on SAP's tenant isolated Hyperscaler infrastructure and managed by SAP AI Core.

Remote (*provider e.g. AWS*): Model(s) hosted and managed by Providers, accessed via SAP AI Core.

Model Versions

You have the following options when working with a model:

- You choose to always have the "latest" version of the model. The "latest" model version can be found in the table. When a version is deprecated, you will automatically be moved to the latest version of the model. Be aware that model behavior could change from version to version.
- You choose to work with a specific version of a model. In this case, you must check when the version will be deprecated and take appropriate steps.

For more information, see Create a Deployment for a Generative AI Model.

Model Availability, Deprecation, and Retirement Dates

Models are deprecated when more recent models or model versions are available with better performance. There may be models without a deprecation period before retirement. Please note that models marked as deprecated could mean there are newer, better and/or cheaper model alternatives available on Generative AI Hub. See "Suggested Replacements".

Models will be removed on the retirement date. API calls to these models will result in error responses.

We recommend that you avoid using deprecated models. If a model is deprecated, switch to newer models or model versions.

Note that deprecation or retirement dates may be adjusted. Please monitor this SAP Note for deprecation and retirement date changes.

The date format is YYYY-MM-DD.

Rate Limits

The rate limits in the table are requests made each minute to a specific model under an SAP AI Core tenant.

- AI Core tenancy is BTP subaccount based, multiple deployments for same model under same tenant/subaccount shares quota numbers shared in table.
- Limits are applied at tenant level based on certain assumptions such as the utilization of your global quota at a given point in time.
- When the limits are reached, clients will receive a 429 response code (too many requests).

Conversion Rates for GenAI Tokens

GenAI tokens correspond to blocks of 1,000 tokens of each model. The amount of GenAI tokens varies depending on the model used and the type of token (input or output). You can refer to the conversion rates listed above to determine how many GenAI tokens are consumed for each model and token type.

Use of orchestration modules may incur costs, in addition to the costs associated with completion calls. For more information on conversion rates, see SAP Note 3505347.

Recently Retired Models

These models have recently been retired and are no longer available.

Executable ID (Provider)	Model	Version	Suggested Replacement
aicore-opensource (self-hosted)	tiiuae--falcon-40b-instruct	N/A (latest)	meta--llama3.1-70b-instruct mistralai--mixtral-8x7b-instruct-v01
azure-openai (Azure)	gpt-35-turbo	0613	gpt-4o-mini
azure-openai (Azure)	gpt-35-turbo-16k	0613	gpt-4o, gpt-4.1
aicore-opensource (sap-managed)	meta--llama3-70b-instruct	N/A (latest)	meta--llama3.1-70b-instruct
gcp-vertexai (Google)	textembedding-gecko	003	
gcp-vertexai (Google)	textembedding-gecko-multilingual	001	
gcp-vertexai (Google)	text-bison	002	gemini-2.0-flash
gcp-vertexai (Google)	gemini-1.0-pro	001	gemini-2.0-flash
gcp-vertexai (Google)	chat-bison	002	gemini-2.0-flash

SLA

Please note that the model response time depends on the token size, so we are unable to offer a latency SLA for Generative AI models.

This document is referenced by

SAP Note/KBA	Component	Title
3566760		Region mapping between SAP BTP ABAP Environment and SAP AI Core
3505347	CA-ML-AIC	Orchestration
3248365	CA-ML-AIC	Central SAP Note for SAP AI Core