

Sarvam releases 30B and 105B open models trained from scratch in India

Source copy kept by wheresthe.ai for the story "Sarvam releases 30B and 105B open models trained from scratch in India" (<https://wheresthe.ai/d/in-sarvam-30b-105b-open-models-2026-02/>).

Source: <https://docs.sarvam.ai/api-reference-docs/models/sarvam-105b>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: f6edd6e77a617a95135bb9bd060f9229ca377a1d3dc7d1d16c4b1d960d22cf8d

Skip to navigation This page is for Model APIs. Sarvam-105B (Flagship Chat LLM)

Sarvam AI's flagship Mixture-of-Experts reasoning model trained from scratch, with Multi-head Latent Attention (MLA) for efficient long-context inference. Matches or outperforms most open and closed-source frontier models of its class across knowledge, reasoning, and agentic benchmarks.

Highlights:

105B+ total parameters : our most capable MoE model with Multi-head Latent Attention

Pre-trained on 12 trillion tokens across code, math, multilingual, and web data

98.6 on Math500 , 88.3 on AIME 25 (96.7 with tools), 49.5 on BrowseComp

State-of-the-art Indian language performance: wins 90% of pairwise comparisons

Powers Indus , Sarvam's AI assistant for complex reasoning and agentic workflows

OpenAI-compatible chat completions API | Apache 2.0 open-source

At a Glance

Model ID sarvam-105b (flagship) · sarvam-105b-conversations (conversational) What it does Flagship chat LLM, complex reasoning, coding, and agentic workflows (105B+ MoE with MLA). The conversational variant is post-trained for real-time dialogue and voice-agent use cases. Languages 10 most-spoken Indian languages + English; native script, romanized, and code-mixed input APIs Chat Completions (POST /v1/chat/completions ; sarvam-105b also on /v2/chat/completions , OpenAI-compatible, streaming supported) Input limits 128K-token context window (sarvam-105b); 32K-token context window (sarvam-105b-conversations), all limits Benchmarks Key Features and the model blog Pricing Pricing page Best for Multi-step reasoning, code generation, long-context document analysis, agentic tool use Known limitations See below

Key Features

Flagship Indian Language Support Wins 90% of pairwise comparisons across Indian language benchmarks and 84% on STEM, math, and coding. Trained extensively on native script, romanized, and code-mixed inputs across the 10 most-spoken Indian languages.

Advanced Reasoning 98.6 on Math500, 88.3 on AIME 25 (96.7 with tools), 85.8 on HMMT, and 69.1 on Beyond AIME, reflecting deep multi-step reasoning and complex mathematical problem solving.

Agentic Capabilities 49.5 on BrowseComp and 68.3 on Tau2 (avg.), highest among compared models. Optimized for tool use, long-horizon reasoning, and environment interaction in real-world workflows.

Efficient MoE Architecture Mixture-of-Experts Transformer with 128 sparse experts and Multi-head Latent Attention (MLA), a compressed attention formulation that reduces memory requirements for long-context inference.

Learn More

For detailed information on architecture, training methodology, performance benchmarks, and inference optimizations, visit our blog .

Model Specifications

Key Considerations Model IDs: sarvam-105b , sarvam-105b-conversations Total Parameters: 105B+ with MoE architecture and 128 sparse experts Attention: Multi-head Latent Attention (MLA) Pre-training Data: 12T tokens Temperature range: 0 to 2 Top-p range: 0 to 1 Supports streaming and non-streaming responses OpenAI-compatible chat completions format License: Apache 2.0

Sarvam-105B is our flagship chat model, delivering the highest quality outputs for complex reasoning and agentic tasks. Use model ID sarvam-105b-conversations when you need a conversational variant tuned for real-time dialogue and voice agents. See Conversational variant . Sarvam-M (24B) has been deprecated and is no longer available through the API.

Conversational variant

Model ID sarvam-105b-conversations is a post-trained variant of Sarvam-105B for real-time conversational workloads : voice agents, chatbots, and multi-turn dialogue where natural, colloquial Indic responses matter more than deep reasoning traces.

sarvam-105b sarvam-105b-conversations Best for Complex reasoning, coding, long-context analysis, agentic tool use Real-time dialogue, voice agents, customer-facing chat Endpoint POST /v1/chat/completions or POST /v2/chat/completions POST /v1/chat/completions only Context window 128K tokens 32K tokens Pricing Sarvam 105B Sarvam 105B Chat , same rates

Both model IDs share the same OpenAI-compatible request schema and Indian language coverage, but sarvam-105b-conversations has a smaller 32K context window (vs. 128K for sarvam-105b). Pass the model ID in the model field, everything else stays the same.

Python JavaScript cURL

```
from sarvamai import SarvamAI
```

```
client = SarvamAI ( api_subscription_key = " YOUR_SARVAM_API_KEY " )
```

```
response = client . chat . completions ( model = " sarvam-105b-conversations " , messages = [ { " role " : " user " ,  
" content " : " who are you? " } ] , temperature = 1 , top_p = 1 , max_tokens = 100 , )
```

```
print ( response . choices [ 0 ] . message . content )
```

Key Capabilities

Basic Chat Completion Multi-turn Conversation Streaming

Simple, one-turn interaction where the user asks a question and the model replies with the highest quality response leveraging its 105B parameter knowledge.

Python JavaScript cURL

```
from sarvamai import SarvamAI
```

```
client = SarvamAI ( api_subscription_key = " YOUR_SARVAM_API_KEY " , )
```

```
response = client . chat . completions ( model = " sarvam-105b " , messages = [ { " role " : " user " , " content " : "  
Explain the economic impact of GST implementation in India. " } ] , temperature = 0.5 , top_p = 1 , max_tokens =  
2000 , )
```

```
print ( response . choices [ 0 ] . message . content )
```

Limits

Limit Value Context window 128K tokens (sarvam-105b) / 32K tokens (sarvam-105b-conversations)
max_tokens Starter 4096 / Pro 16384 / Business 128000 temperature 0–2 (default 0.5 when reasoning is enabled, the default, and 0.2 when reasoning is disabled) top_p 0–1 n (completions per request) 1–128 frequency_penalty / presence_penalty -2 to 2 stop Up to 4 sequences Rate limits See Rate Limits

Known Limitations

Limitation Detail Workaround Thinking mode is on by default Reasoning (reasoning_effort , default low) is enabled by default, and reasoning tokens count toward completion tokens, a small max_tokens can be consumed

entirely by reasoning Increase max_tokens to leave room for the visible answer, or disable reasoning with reasoning_effort=None

Next Steps

Developer quickstart Learn how to integrate chat completion into your application.

API Reference Complete API documentation for chat completion endpoints.

DesktopMenu) -->