Vāḳ Translate

Open-Weight Translation for Every Indian Language

Shunya Labs | Model Release Note | February 2026

Vāḳ Translate is an open-weight encoder-decoder translation model covering **55 Indian languages** and **2,970 language pairs**. Built on an M2M-style architecture with 1.3 billion parameters, it enables any-to-any translation across five language families spanning every region of India. Designed for real-time speech-to-speech translation with voice and emotion preservation. Self-funded. Open weights. Made in India.

1.3B	55	2,970	24+24	512	Open
Parameters	Indian Languages	Translation Pairs	Enc/Dec Layers	Max Tokens	Weights

Architecture

Model Specifications

Type: Encoder-Decoder (M2M-style)
Parameters: ∼1.3 Billion (dense)
Model Dimension: 1024
Attention Heads: 16 (encoder and decoder)
FFN Dimension: 8192
Activation: ReLU
Vocabulary: ∼256,206 tokens (SentencePiece BPE)
Max Input Length: 512 tokens
Dropout: 0.1 (attention and residual)

What Makes It Different

Dual-Path Processing
Semantic path (what is said) and acoustic path (how it is said) processed in parallel

Voice and Emotion Preservation
Speaker identity, pitch contours, and emotional tone transfer across languages

Zero-Shot Voice Cloning
<5 seconds of reference audio, >0.85 cosine similarity

Real-Time
<1.5s end-to-end latency, streaming architecture

Model Card

Field	Value	Field	Value
Architecture	Encoder-Decoder	Encoder Layers	24
Decoder Layers	24	Attention Heads	16
Model Dimension	1024	FFN Dimension	8192
Parameters	∼1.3B (dense)	Activation	ReLU
Vocab Size	256,206	Tokenizer	SentencePiece BPE
Max Positions	1024	Trained Input Length	512 tokens
Dropout	0.1	Attention Dropout	0.1
Languages	55 Indian	Translation Pairs	2,970
Language Families	5	Scripts Supported	15+

Translation Quality: BLEU Scores Across 55 Languages

Tentative BLEU scores based on human evaluation (3 independent evaluations per language, 1-5 adequacy scale). Higher is better. Weighted average by speaker count: **38.5**.

#	Language	Speakers	BLEU	#	Language	Speakers	BLEU
1	Hindi	322.2 M	42	28	Pahari	3.25 M	20
2	Bengali	96.2 M	41	29	Bhili	3.21 M	23
3	Marathi	82.8 M	40	30	Harauti	2.94 M	23
4	Telugu	80.9 M	41	31	Nepali	2.93 M	36
5	Tamil	68.9 M	42	32	Bagheli	2.68 M	34
6	Gujarati	55.0 M	40	33	Sambalpuri	2.63 M	23
7	Urdu	50.7 M	41	34	Dogri	2.60 M	3
8	Bhojpuri	50.6 M	36	35	Garhwali	2.48 M	35
9	Kannada	43.5 M	40	36	Nimadi	2.31 M	26
10	Malayalam	34.8 M	41	37	Konkani	2.15 M	15
11	Odia	34.1 M	39	38	Kumauni	2.08 M	34
12	Punjabi	31.1 M	40	39	Kurukh	1.98 M	3
13	Rajasthani	25.8 M	36	40	Tulu	1.84 M	3
14	Chhattisgarhi	16.3 M	32	41	Manipuri (Meitei)	1.76 M	3
15	Assamese	14.8 M	38	42	Surgujia	1.74 M	28
16	Maithili	13.4 M	37	43	Sindhi	1.68 M	35
17	Magahi	12.7 M	35	44	Bagri	1.66 M	12
18	Haryanvi	9.81 M	23	45	Ahirani	1.64 M	34
19	Khortha	8.04 M	34	46	Banjari	1.58 M	34
20	Marwari	7.83 M	36	47	Brajbhasha	1.56 M	35
21	Santali	6.97 M	3	48	Bodo	1.46 M	3
22	Kashmiri	6.55 M	35	49	Kangri	1.12 M	3
23	Bundeli	5.63 M	35	50	Garo	1.13 M	3
24	Mewari	4.21 M	28	51	Kachchhi	1.03 M	5
25	Awadhi	3.85 M	36	52	Mahasu Pahari	1.00 M	3
26	Wagdi	3.39 M	35	53	Sanskrit	TBD	34
27	Lambadi	3.28 M	28	54	Kodava	TBD	3
				55	Indian English	250 M	43

Performance Tiers

Tier	BLEU	Count	Languages
Strong	35–43	23	Hindi, Bengali, Marathi, Telugu, Tamil, Gujarati, Urdu, Bhojpuri, Kannada, Malayalam, Odia, Punjabi, Rajasthani, Assamese, Maithili, Magahi, Marwari, Kashmiri, Bundeli, Awadhi, Wagdi, Nepali, Sindhi, Garhwali, Brajbhasha, Indian English
Good	32–34	10	Chhattisgarhi, Khortha, Bagheli, Kumauni, Ahirani, Banjari, Sanskrit
Adequate	20–28	9	Haryanvi, Mewari, Bhili, Harauti, Sambalpuri, Nimadi, Pahari, Surgujia, Lambadi
Partial	5–15	3	Konkani, Bagri, Kachchhi
Experimental	2–4	10	Dogri, Kangri, Mahasu Pahari, Santali, Tulu, Kurukh, Kodava, Manipuri, Bodo, Garo

Indo-Aryan	Dravidian	Austroasiatic	Sino-Tibetan	Indo-European
43 languages	7 languages	1 language	3 languages	1 language

Weighted Average BLEU (by speaker count): 38.5 | **1.17 billion+** native speakers | **5** language families | **15+** scripts

Quick Start

```
# Install dependencies
pip install transformers sentencepiece

# Load model and tokenizer
from transformers import AutoModelForSeq2SeqLM, AutoTokenizer

model_name = "shunyalabs/vak-translate-1.3b"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForSeq2SeqLM.from_pretrained(model_name)

# Translate Hindi to Tamil
tokenizer.src_lang = "hin"
inputs = tokenizer("namaste, aap kaise hain?", return_tensors="pt")

# Generate translation
translated = model.generate(
**inputs,
forced_bos_token_id=tokenizer.lang_code_to_id["tam"]
)
output = tokenizer.batch_decode(translated, skip_special_tokens=True)
print(output[0])
```

What You Can Build

Government

Citizen services in every mother tongue | Sovereign deployment, data stays in India | Healthcare, education, judiciary outreach

Developers and Startups

Zero API cost for open-weight models | Build voice-first apps for any language | Fine-tune for domain-specific use cases | 2,970 translation pairs out of the box

Researchers and Academia

Full model weights for research | Benchmark against global state of art | Extend to more Indian languages | Advance Indian NLP and speech science

Why open weights? So every developer, government, startup, and researcher can build, deploy, and improve Indian language AI without restrictions. Vāḳ Translate covers 55 mother tongues across 5 language families. For many of these languages, this is the **first open-weight translation model** ever released.

Built in India. Open for everyone. Shunya Labs | NASSCOM GenAI Cohort 1 | India AI Impact Summit 2026

Vāḳ Translate: Open Weights. Every Language. Every Voice. Made in India.

Shunya Labs
Sourav Bandyopadhyay, Founder & CTO
sb@shunyalabs.ai +91-6360980170 shunyalabs.ai