Title: Ernie Bot

URL: https://en.wikipedia.org/wiki/Ernie_Bot

PageID: 73311298

Categories: Category:2023 in artificial intelligence, Category:Baidu, Category:Chatbots, Category:Generative pre-trained transformers, Category:Large language models, Category:Open-source artificial intelligence, Category:Software using the Apache license

---

Ernie Bot (Chinese : ■■■■ , Pinyin : wénx■n y■yán), full name Enhanced Representation through Knowledge Integration , [1] is an artificial intelligence chatbot developed by the Chinese technology company Baidu . It is built on the company's ERNIE series of large language models , which have been in development since 2019. The service was first launched for invited testing on March 16, 2023, [2] and was released to the general public on August 31, 2023, after receiving approval from Chinese regulators. [3]

Since its public launch, Ernie Bot has undergone several updates, with newer versions like ERNIE 4.0 and 4.5 released to improve its capabilities. The service has seen rapid user adoption, reportedly reaching over 200 million users by April 2024. [4] It has been integrated into various products, notably powering AI features for the Chinese release of Samsung 's Galaxy S24 smartphones. [5]

As a product operating in China, Ernie Bot is subject to the country's censorship regulations . It has been observed to refuse answers to politically sensitive questions, such as those regarding Xi Jinping , the 1989 Tiananmen Square protests and massacre , and other topics deemed taboo by the government. [6] [7]

History

Ernie Bot was initially released for invited testing on March 16, 2023. [8] [9] The live release demo was reported to have been prerecorded, which caused Baidu's stock to drop 10 percent on the day of the launch. [10] The company's stock gained 14 percent the following day after analysts from Citigroup and Bank of America tested Ernie Bot and gave it positive preliminary reviews. [11]

On August 31, 2023, Ernie Bot was released to the public after receiving approval from Chinese regulatory authorities. [12] By December 2023, Baidu announced the service had surpassed 100 million users. [13]

In January 2024, Hong Kong newspaper South China Morning Post reported that a university research lab linked to the People's Liberation Army (PLA) had tested Ernie Bot for military response scenarios. Baidu denied the allegations, stating it had no connection with the academic paper. [14] That same month, Ernie was integrated into Samsung 's Galaxy S24 lineup for its launch in China. [15] [16]

The user base reportedly grew to 200 million by April 2024 and 300 million by June 2024. [17] [18] In September 2024, Baidu changed the chatbot's Chinese name from "Wenxin Yiyan" (■■■■) to "Wenxiaoyan" (■■■) to position it as a search assistant. [19] [20]

On March 16, 2025, Baidu announced version 4.5 and the reasoning model ERNIE X1. [21] The following month, at the Create2025 Baidu AI Developer Conference, the company released the Wenxin 4.5 Turbo and Wenxin X1 Turbo models, designed to be faster and less expensive to operate. [22]

Development

Ernie Bot is based on Baidu's ERNIE (Enhanced Representation through Knowledge Integration) series of foundation models. The general training process begins with pre-training on large datasets, followed by refinement using techniques like supervised fine-tuning, reinforcement learning with human feedback , and prompt engineering . [23]

Foundation models

Ernie 3.0

The model powering the initial launch of Ernie Bot.

It was trained with 10 billion parameters on a 4-terabyte corpus consisting of plain text and a large-scale knowledge graph. [24]

Ernie 3.5

Released in June 2023. At the time of release, its performance was reported as "slightly inferior" to OpenAI's GPT-4 . [25]

Ernie 4.0

Unveiled in October 2023 and released to paying subscribers in November.

According to Baidu, this version featured improved performance over its predecessor, with information updated to April 2023. [26]

Ernie X1

Announced in March 2025, with Ernie X1 positioned as a specialized reasoning model.

Baidu stated that performance improvements were achieved through new technologies such as "FlashMask" dynamic attention masking and a heterogeneous multimodal mixture-of-experts architecture. [21]

Turbo Models

In June 2024, Baidu announced Ernie 4.0 Turbo. In April 2025, Ernie 4.5 Turbo and X1 Turbo were released.

These models are optimized for faster response times and lower operational costs. [27] [28]

Service

In its subscription options, the professional plan gives users access to Ernie 4.0 with a payment either for a month or with reduced payment for auto-renewal per month. Meanwhile, Ernie 3.5 is free of charge. [29]

Ernie 4.0, the language model for Ernie bot, has information updated to April 2023. [26]

Censorship

Ernie Bot is subject to the Chinese government's censorship regime . [30] [7] [31]

In public tests with journalists, Ernie Bot refused to answer questions about Xi Jinping , the 1989 Tiananmen Square protests and massacre , the persecution of Uyghurs in China in Xinjiang , and the 2019–2020 Hong Kong protests . [7] [32] [33]

When queried about the origin of SARS-CoV-2 , Ernie Bot stated that it originated among American vape users. [7]

See also

Artificial intelligence industry in China

ChatGPT

Google Gemini

References

External links

Official website

Ernie 4.5 models on Hugging Face

Media related to ERNIE Bot at Wikimedia Commons

v

t

e

Robin Li

Li Mingyuan

Lu Qi

Baidu Search Engine

Baidu News

Baidu MP3

Baidu Images

Baidu Video

Baidu Tieba

Baidu Zhidao

Baidu Maps

Baidu Space

Baidu Baike

Baidu Scholar (Academic Search Engine)

Baidu Hi (Instant Messenger)

Baidu Guoxue

Baidu Toolbar

Baidu Fanyi

Baidu Patents

Baidu Raven

Baidu Youa (Online Shopping)

Baifubao (Mobile Wallet)

Baidu 500

Hao123

Baidu Wenku (File Sharing)

Baidu Yi

Baidu Music

Ernie Bot

Death of Wei Zexi

v

t

e

LMArena

List of chatbots

List of LLMs
character.ai
ChatGPT
Claude
Command
Copilot
DeepSeek
Ernie
Gemini
GLM
Grok
Hunyuan
Kimi
Llama
Mistral
Perplexity
Poe
Qwen
You.com
Category
v
t
e
Autoencoder
Deep learning
Fine-tuning
Foundation model
Generative adversarial network
Generative pre-trained transformer
Large language model
Model Context Protocol
Neural network
Prompt engineering
Reinforcement learning from human feedback
Retrieval-augmented generation
Self-supervised learning
Stochastic parrot
Synthetic data

Top-p sampling
Transformer
Variational autoencoder
Vibe coding
Vision transformer
Waluigi effect
Word embedding
Character.ai
ChatGPT
DeepSeek
Ernie
Gemini
Grok
Copilot
Claude
Gemini
Gemma
GPT 1 2 3 J 4 4o 4.5 4.1 OSS 5
1
2
3
J
4
4o
4.5
4.1
OSS
5
Llama
o1
o3
o4-mini
Qwen
Base44
Claude Code
Cursor
Devstral
GitHub Copilot

Kimi-Dev
Qwen3-Coder
Replit
Xcode
Aurora
Firefly
Flux
GPT Image 1
Ideogram
Imagen
Midjourney
Qwen-Image
Recraft
Seedream
Stable Diffusion
Dream Machine
Hailuo AI
Kling
Midjourney Video
Runway Gen
Seedance
Sora
Veo
Wan
15.ai
Eleven
MiniMax Speech 2.5
WaveNet
Eleven Music
Endel
Lyria
Riffusion
Suno AI
Udio
Agentforce
AutoGLM
AutoGPT
ChatGPT Agent

Devin AI
Manus
OpenAI Codex
Operator
Replit Agent
01.AI
Aleph Alpha
Anthropic
Baichuan
Canva
Cognition AI
Cohere
Contextual AI
DeepSeek
ElevenLabs
Google DeepMind
HeyGen
Hugging Face
Inflection AI
Krikey AI
Kuaishou
Luma Labs
Meta AI
MiniMax
Mistral AI
Moonshot AI
OpenAI
Perplexity AI
Runway
Safe Superintelligence
Salesforce
Scale AI
SoundHound
Stability AI
Synthesia
Thinking Machines Lab
Upstage
xAI

Z.ai

Category