Title: 1.58-bit large language model

URL: https://en.wikipedia.org/wiki/1.58-bit_large_language_model

PageID: 79773289

Categories: Category:Large language models

A 1.58-bit large language model (also known as a ternary LLM) is a type of large language model (LLM) designed to be computationally efficient. It achieves this by using weights that are restricted to only three values: -1, 0, and +1. This restriction significantly reduces the model's memory footprint and allows for faster processing, as computationally expensive multiplication operations can be replaced with lower-cost additions. This contrasts with traditional models that use 16-bit floating-point numbers (FP16 or BF16) for their weights.

Studies have shown that for models up to several billion parameters, the performance of 1.58-bit LLMs on various tasks is comparable to their full-precision counterparts. [1] [2] This approach could enable powerful AI to run on less specialized and lower-power hardware. [3]

The name "1.58-bit" comes from the fact that a system with three states contains log 2 ■ 3 ≈ 1.58 {\displaystyle \log _{2}3\approx 1.58} bits of information . These models are sometimes also referred to as 1-bit LLMs in research papers, although this term can also refer to true binary models (with weights of -1 and +1). [1] [4]

BitNet

In 2024, Ma et al., researchers at Microsoft , declared that their 1.58-bit model, BitNet b1.58 is comparable in performance to the 16-bit Llama 2 and opens the era of 1-bit LLM. [5] BitNet creators did not use the post-training quantization of weights but instead relied on the new BitLinear transform that replaced the nn.Linear layer of the traditional transformer design. [6]

In 2025, Microsoft researchers had released an open-weights and open inference code model BitNet b1.58 2B4T demonstrating performance competitive with the full precision models at 2B parameters and 4T training tokens. [7]

Post-training quantization

BitNet derives its performance from being trained natively in 1.58 bit instead of being quantized from a full-precision model after training. Still, training is an expensive process and it would be desirable to be able to somehow convert an existing model to 1.58 bits. In 2024, HuggingFace reported a way to gradually ramp up the 1.58-bit quantization in fine-tuning an existing model down to 1.58 bits. [8]

Critique

Some researchers [9] point out that the scaling laws [10] of large language models favor the low-bit weights only in case of undertrained models. As the number of training tokens increases, the deficiencies of low-bit quantization surface.

References

Sources

Ma, Shuming; Wang, Hongyu; Ma, Lingxiao; Wang, Lei; Wang, Wenhui; Huang, Shaohan; Dong, Li; Wang, Ruiping; Xue, Jilong; Wei, Furu (2024-02-27). "The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits". arXiv : 2402.17764 [cs.CL].

Ma, Shuming; Wang, Hongyu; Huang, Shaohan; Zhang, Xingxing; Hu, Ying; Song, Ting; Xia, Yan; Wei, Furu (2025). "BitNet b1.58 2B4T Technical Report". arXiv : 2504.12285 [cs.CL].

Friha, Othmane; Amine Ferrag, Mohamed; Kantarci, Burak; Cakmak, Burak; Ozgun, Arda; Ghoualmi-Zine, Nassira (2024). "LLM-Based Edge Intelligence: A Comprehensive Survey on

Architectures, Applications, Security and Trustworthiness" . IEEE Open Journal of the Communications Society . 5 : 5799– 5856. Bibcode : 2024IOJCS...5.5799F . doi : 10.1109/OJCOMS.2024.3456549 . ISSN 2644-125X .

Hutson, Matthew (2024-05-30). "1-bit LLMs Could Solve AI's Energy Demands" . IEEE Spectrum . Retrieved 2025-04-22 .

Huyen, Chip (2024-12-04). AI Engineering . "O'Reilly Media, Inc.". ISBN 978-1-0981-6627-4 . Retrieved 2025-04-22 .

Kumar, Tanishq; Ankner, Zachary; Spector, Benjamin F.; Bordelon, Blake; Muennighoff, Niklas; Paul, Mansheej; Pehlevan, Cengiz; Ré, Christopher; Raghunathan, Aditi (2024). "Scaling Laws for Precision". arXiv : 2411.04330 [cs.LG].

Morales, Jowi (2025-04-17). "Microsoft researchers build 1-bit AI LLM with 2B parameters" . Tom's Hardware . Retrieved 2025-04-21 .

Ouyang, Xu; Ge, Tao; Hartvigsen, Thomas; Zhang, Zhisong; Mi, Haitao; Yu, Dong (2024). "Low-Bit Quantization Favors Undertrained LLMS: Scaling Laws for Quantized LLMS with 100T Training Tokens". arXiv : 2411.17691 [cs.LG].

Wang, Hongyu; Ma, Shuming; Dong, Li; Huang, Shaohan; Wang, Huaijie; Ma, Lingxiao; Yang, Fan; Wang, Ruiping; Wu, Yi; Wei, Furu (2023). "BitNet: Scaling 1-bit Transformers for Large Language Models". arXiv : 2310.11453 [cs.CL].

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