Mixture-of-Experts Handbook: Hugging Face's Free Guide to Training AI

Kabui, Charles

2026-11-11

Read at ToKnow.ai



Hugging Face's Mixture-of-Experts Handbook is a free, book-length guide for teams with a GPU cluster who want to train a mixture-of-experts model (MoE) with NVIDIA's Megatron-LM software. An MoE works like a team of specialists: a small router sends each token (a word or word piece) to only a few experts, so the model holds far more parameters (learned numbers) than it uses on any one token. The handbook's starter model has 64 experts, picks

8 per token, and holds 6.86 billion parameters while using 1.22 billion at a time. Authors Nouamane Tazi, Leandro von Werra and Thomas Wolf write that technical reports on open models rarely show what a team tried and dropped. They count 18 of the 20 top-scoring open-weights (downloadable) models on the Artificial Analysis ranking as MoEs.

Each recommendation comes with a measurement from the authors' runs on Amazon Web Services machines with eight H100 chips each. A new router splits work unevenly, so every GPU waits for the one holding the busiest experts. Forcing an even split ran the same model 1.5 times faster. Two runs that quietly fell back to a slow network connection ran about 25 times slower than a run on the fast one, and the guide shows the log line that gives it away. Readers without a cluster can rent an eight-GPU machine on Hugging Face Jobs, billed by the minute. Hugging Face says the experiments took about 800,000 GPU hours, roughly 91 years on one GPU.

Read More: Mixture-of-Kittens: Cursor's software that cuts the time mixture-of-experts training spends moving tokens between GPUs

Sources:

- The Ultimate MoE Training Handbook (Hugging Face)
- Hugging Face publishes 70,000-word MoE training handbook (AI Weekly)
- Hugging Face Releases "The Ultimate MoE Training Handbook" (explainx.ai)
- Comparison of Open Source Models (Artificial Analysis)
- Mixture of Experts, Megatron Core documentation (NVIDIA)

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** */terms-of-service*