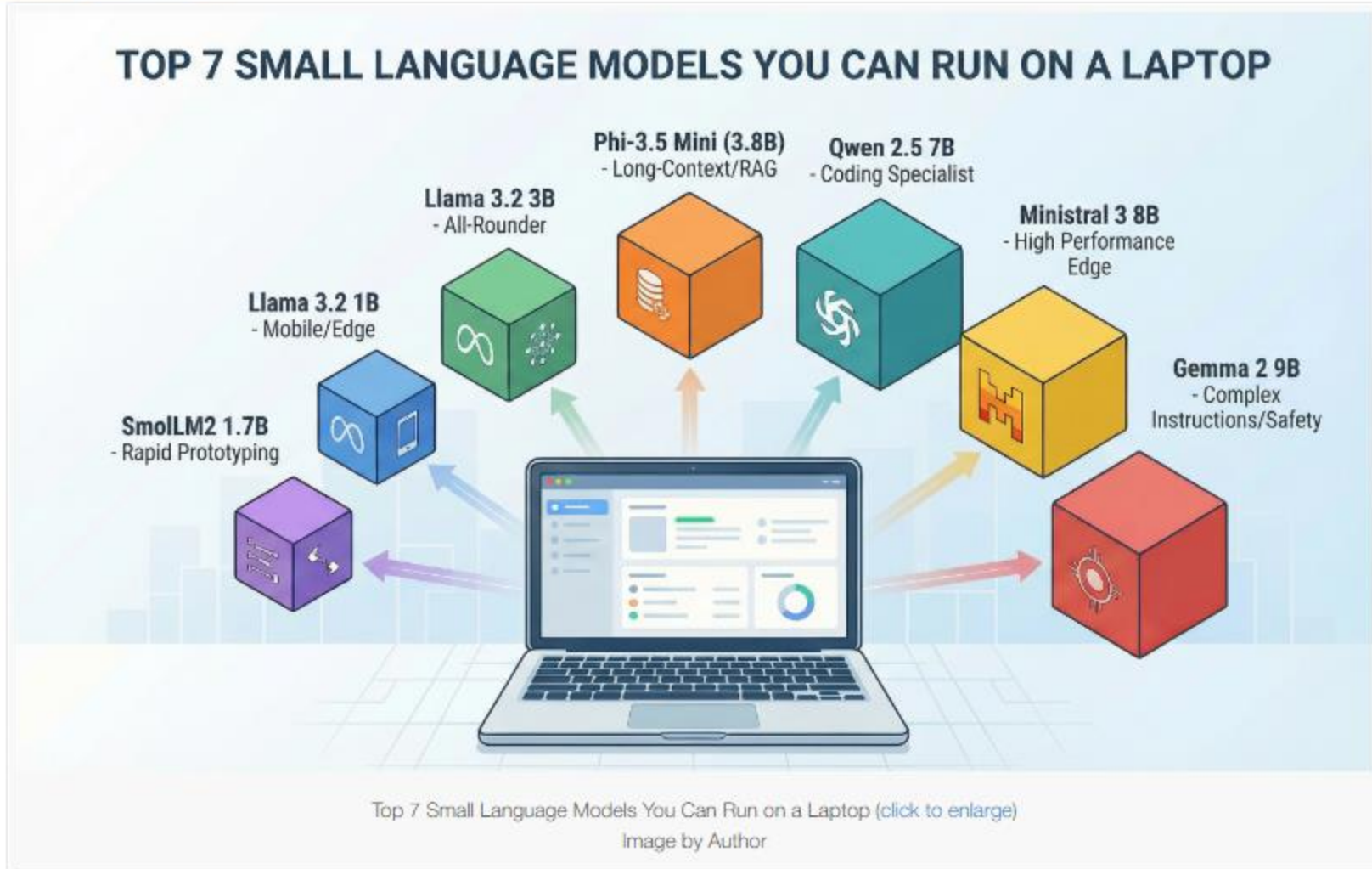
LLM open-source locali

Top 7 Small Language Models You Can Run on a Laptop

by **Vinod Chugani** on February 17, 2026 in **Language Models** 4

Share Share Post



Top 7 Small Language Models You Can Run on a Laptop (click to enlarge)
Image by Author

Introduction

Powerful AI now runs on consumer hardware. The models covered here work on standard laptops and deliver production-grade results for specialized tasks. You'll need to accept license terms and authenticate for some downloads (especially Llama and Gemma), but once you have the weights, everything runs locally.

This guide covers seven practical small language models, ranked by use case fit rather than benchmark scores. Each has proven itself in real deployments, and all can run on hardware you likely already own.

Note: Small models ship frequent revisions (new weights, new context limits, new tags). This article focuses on which model *family* to choose; check the official model card/**Ollama** page for the current variant, license terms, and context configuration before deploying.

LLM open-source (open-weight) scaricabili in locale

febbraio 2026

GPT OSS 20b & 120b

link openAI

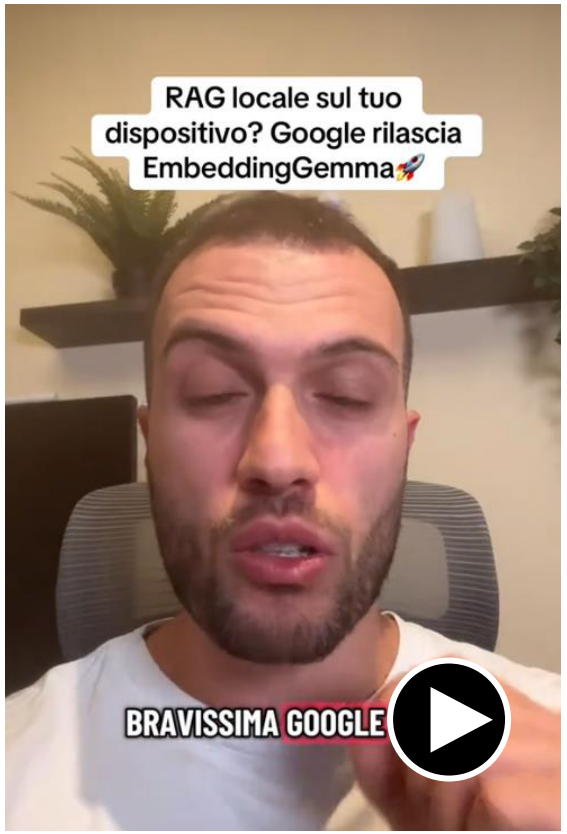
Released alongside GPT-5, these are open-weight, decoder-only transformer models (20B and 120B parameters) available on Hugging Face that can be run locally.

link Hugging Face 20B

link Hugging Face 120B

Introducing EmbeddingGemma

Embedding Gemma



Ollama

ollama

LM Studio

LM Studio

LM Studio / Ollama: You can use these tools to run open-weights models (like Llama 3, Mistral, or gpt-oss) locally on your own hardware to avoid API costs and internet dependency.

Confronto diretto: Ollama vs LM Studio

Aspetto	LM Studio	Ollama
Interfaccia grafica	✔	✘ (CLI)
Semplicità per non tecnici	Molto alta	Media
Automazione script	Media	Alta
Deployment server	Locale desktop	Più adatto a server
Controllo avanzato	Buono	Molto buono

In sintesi:

- LM Studio è ottimo per **ricerca, testing, prototipazione**
- Ollama è più naturale per **integrazione Dev/automazione**

🌐 Modelli generali open-source da scaricare

⭐ Modelli recenti e “forti”

1. **GPT-OSS (OpenAI)** — modelli open-weight rilasciati da OpenAI (es. *gpt-oss-20B*, *gpt-oss-120B*) scaricabili e utilizzabili localmente. Business Insider
2. **LLaMA 4 / LLaMA 3 (Meta)** — famiglia di modelli decoder-only con ottimo bilanciamento tra prestazioni e costi di esecuzione. Reuters
3. **DeepSeek R1 / DeepSeek V3.2** — modelli Mixture-of-Experts (MoE) con focus su reasoning avanzato e performance competitive. bentoml.com
4. **Qwen 3 / Qwen 2.5** — modelli generali e multilingue con dimensioni varie e forte supporto linguistico. Hugging Face
5. **Gemma 3** — modelli con contesti lunghi e buone performance multimodali. LM Studio

---

Differenza fondamentale (spesso confusa)

Termine	Significato reale
Open-source classico	codice + training data + pipeline
Open-weights	pesi disponibili, training non replicabile
Closed	solo API

The gpt-oss models are Apache 2.0 open-source license, similar to Qwen3, which is great. This means that the models can be distilled into other models or used in commercial products without restriction.

Open-weight vs. open-source LLMs. This distinction has been debated for years, but it is worth clarifying to avoid confusion about this release and its artifacts. Some model developers release only the model weights and inference code (for example, Llama, Gemma, gpt-oss), while others (for example, OLMo) release everything including training code, datasets, and weights as true open source.

By that stricter definition, gpt-oss is an *open-weight* model (just like Qwen3) because it includes the weights and inference code but not the training code or datasets. However, the terminology is used inconsistently across the industry.

I assume the "oss" in "gpt-oss" stands for *open source software*; however, I am positively surprised that OpenAI itself clearly describes gpt-oss as an open-weight model in their official announcement article.

link all'articolo di Sebastian Raschka

AHEAD OF AI

Sebastian Raschka, PhD

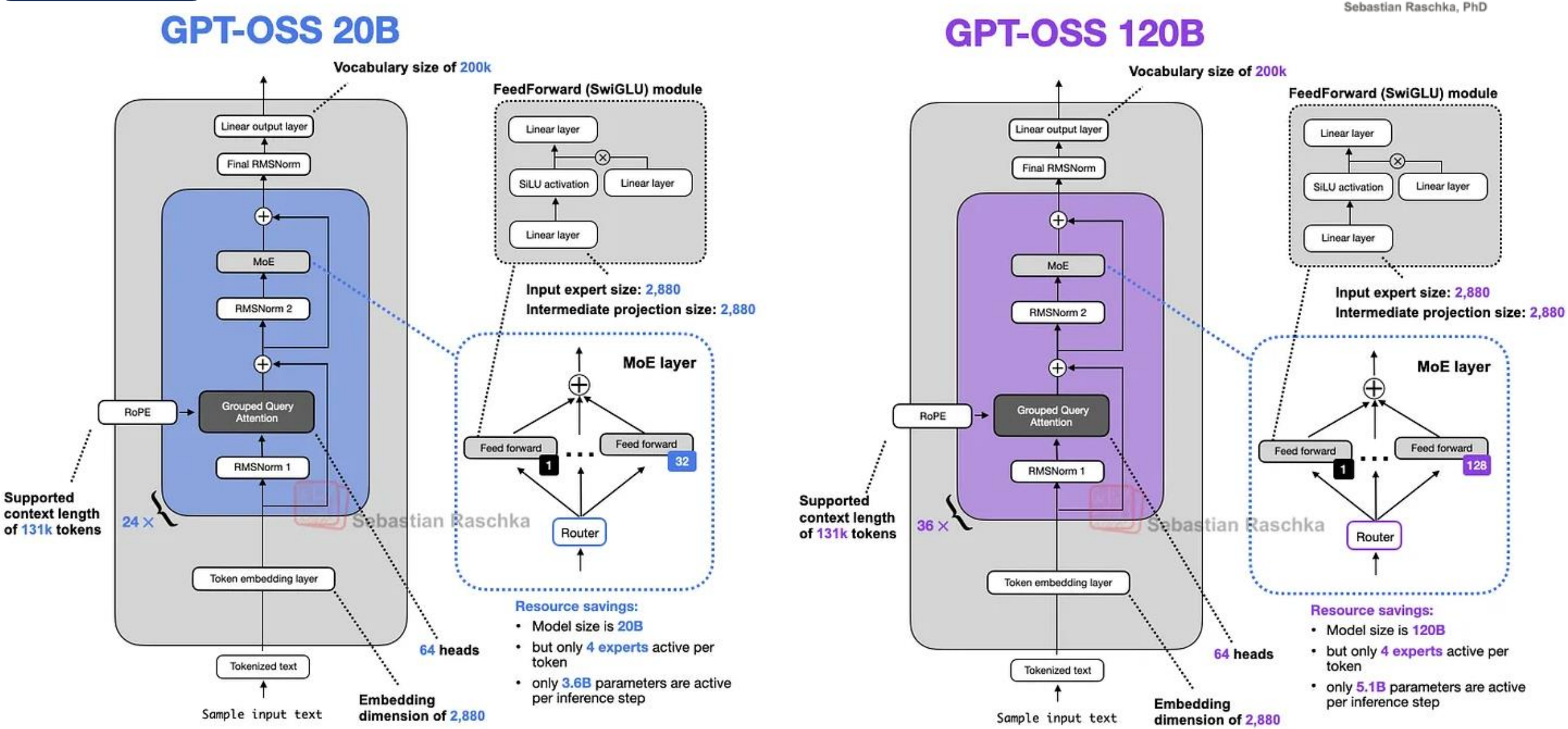


Figure 1: The two gpt-oss models side by side.

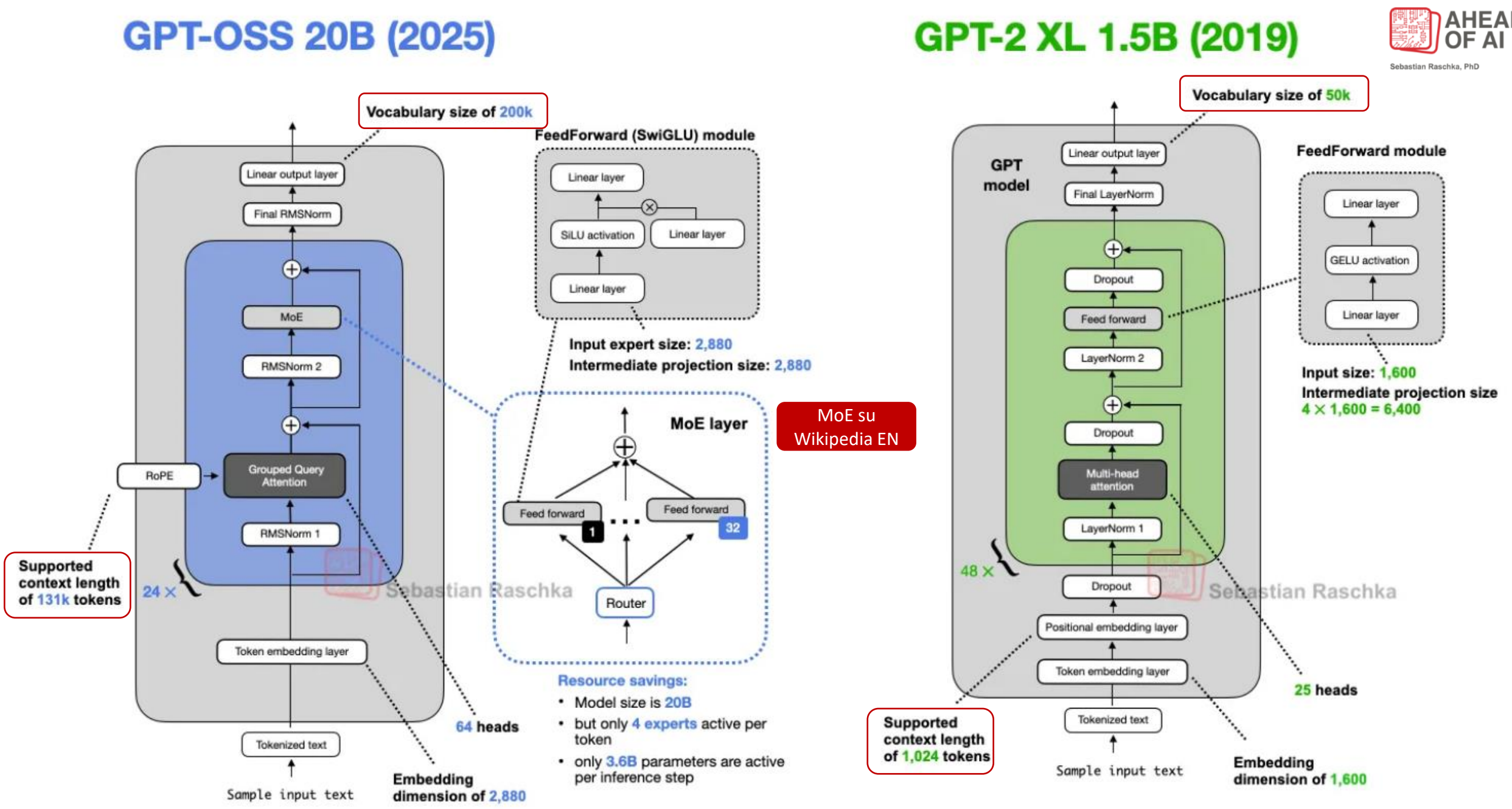


Figure 2: A side-by-side comparison between gpt-oss-20b and GPT-2 XL 1.5B.

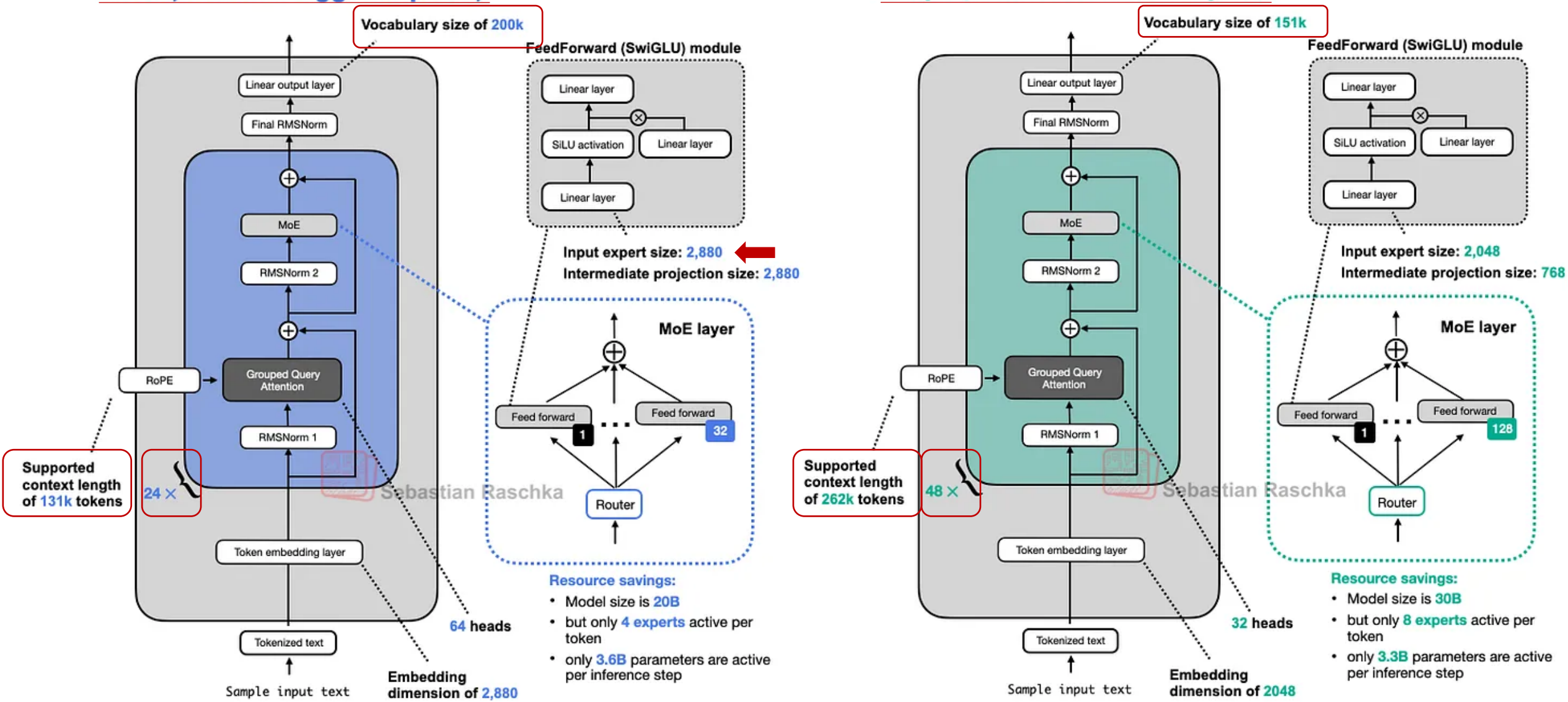


Figure 13: A gpt-oss and Qwen3 model of comparable size side by side.