SLM stands for Small Language Model.

What is a Small Language Model (SLM)?

A **Small Language Model** is a type of language model that has a relatively **small number of parameters** (typically **under 10 billion parameters**) compared to Large Language Models (LLMs) like GPT-4, LLaMA 70B, or Claude.

How SLMs Compare to LLMs:

Feature	SLM	LLM
Size	~1B to ~10B parameters	13B to 180B+ parameters
Speed	Fast on CPU or low-end GPU	Requires high-end GPU
Hardware	Can run on laptops, phones	Needs cloud or data center
Use cases	Edge AI, mobile apps, fast inference	Deep reasoning, complex workflows
Training cost	Low	Very high

Why Use an SLM?

- **Low resource usage** – runs on CPUs or mobile devices
- **Faster inference** – ideal for real-time apps
- **Better privacy** – can run locally
- **Cost-effective** – especially for startups or embedded systems

Example Use Cases:

- Offline chatbots on mobile
- Smart assistants on devices (e.g., home automation)
- Fast, on-device summarization
- Lightweight agents in games or embedded systems

Here are the **best open-source SLMs (Small Language Models)** that are **free for commercial use** — meaning you can integrate them into your apps, services, or products without violating licenses:

Model	Size(s)	License	Commercial Use	Best For
Phi-3	3.8B, 7B	MIT	□ Yes	General use, edge apps
Mistral	7B	Apache 2.0	□ Yes	Chat, reasoning, general AI
Gemma	2B, 7B	Gemma License	□ Yes	Clean NLP, multilingual
TinyLlama	1.1B	Apache 2.0	□ Yes	Mobile/IoT
Falcon	7B	Apache 2.0	□ Yes	Multilingual, general purpose

□ Top Open-Source SLMs for Commercial Use

□ 1. Phi-3 (Microsoft)

- **Sizes**: 3.8B (Mini), 7B (Small)
- **License**: **MIT** (fully permissive)
- **Highlights**:
  - Extremely efficient
  - Strong performance at small sizes
  - Ideal for edge and mobile apps

□ GitHub - Phi-3

□ 2. Mistral-7B

- **Size**: 7B
- **License**: **Apache 2.0** (commercial-friendly)
- **Highlights**:
  - Best performance among 7B models
  - Strong reasoning and general-purpose tasks
  - Fast inference with grouped-query attention

□ GitHub - Mistral

□ 3. Gemma (by Google DeepMind)

- **Sizes**: 2B, 7B
- **License**: **Gemma License** (commercial use allowed)
- **Highlights**:
  - Clean, safe dataset
  - Competitive with LLaMA in many benchmarks

□ Gemma Models on Hugging Face

□ 4. TinyLlama (1.1B)

- **Size**: 1.1B
- **License**: **Apache 2.0**
- **Highlights**:
  - Extremely small, under 1GB quantized
  - Good for ultra-low resource environments

□ GitHub - TinyLlama

□ 5. Falcon-7B

- **Size**: 7B
- **License**: **Apache 2.0** (for Falcon 7B)
- **Highlights**:
  - Versatile and multilingual
  - Used in many research and industry deployments

□ Falcon on Hugging Face