[PPT-07] THE SMALL MODELS

Which model?

- Size you can manage: 2B parameters with 8 GB RAM, 7B with 16 GB RAM.
- Providers: Hugging Face, Ollama.
- Models: Llama (Meta), Gemma (Google), Mistral, DeepSeek, Qwen (Alibaba).

Which interface?

- Chat interface: LM Studio, Hugging Face,Ollama.
- API interface: Groq, together.ai, Hyperbolic, SambaNova.