ZUP

SLM + Ollama

Rodando uma IA direto do seu computador

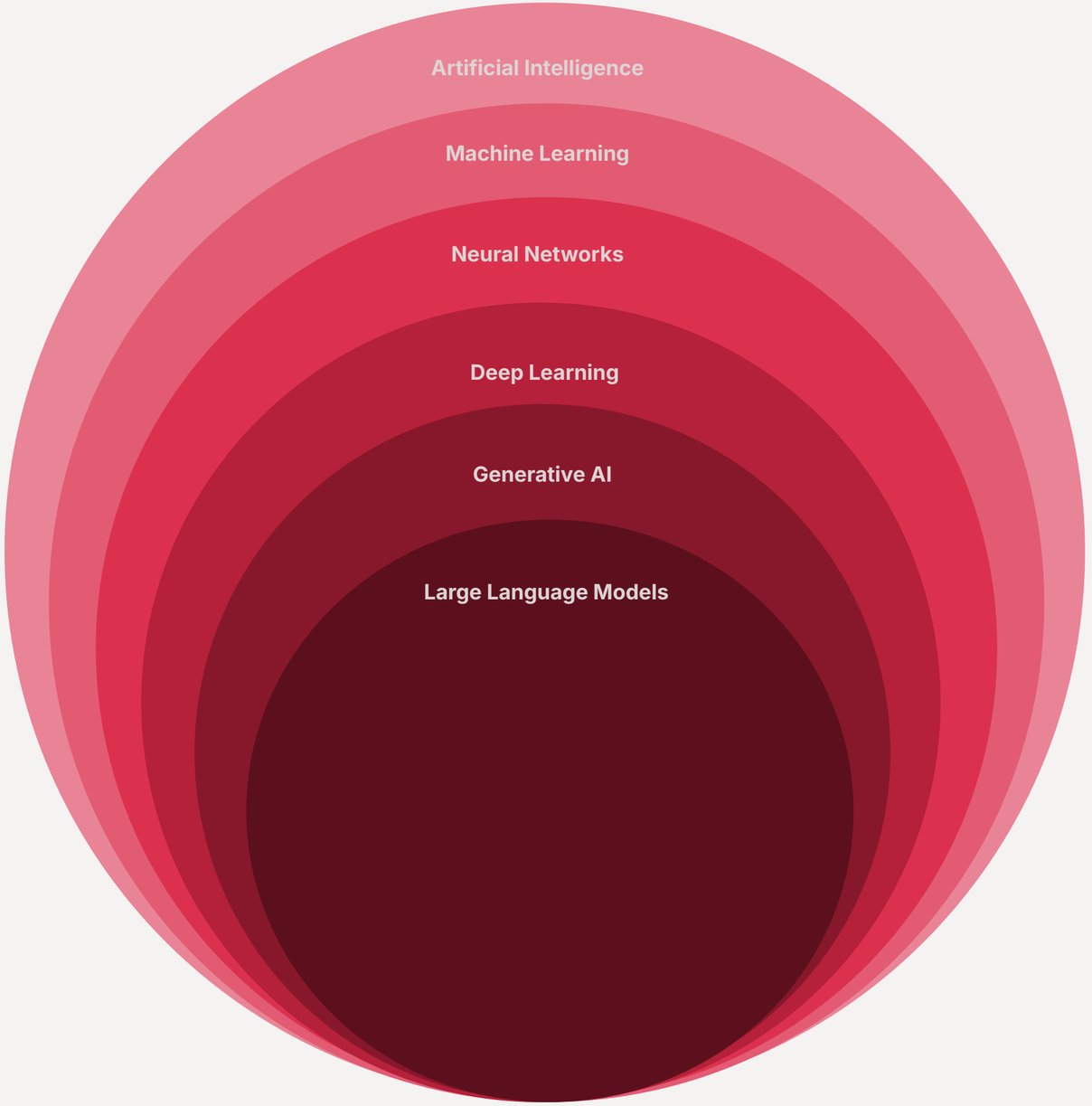
Tio .faso

Popularmente conhecido como “Fábio Sousa”

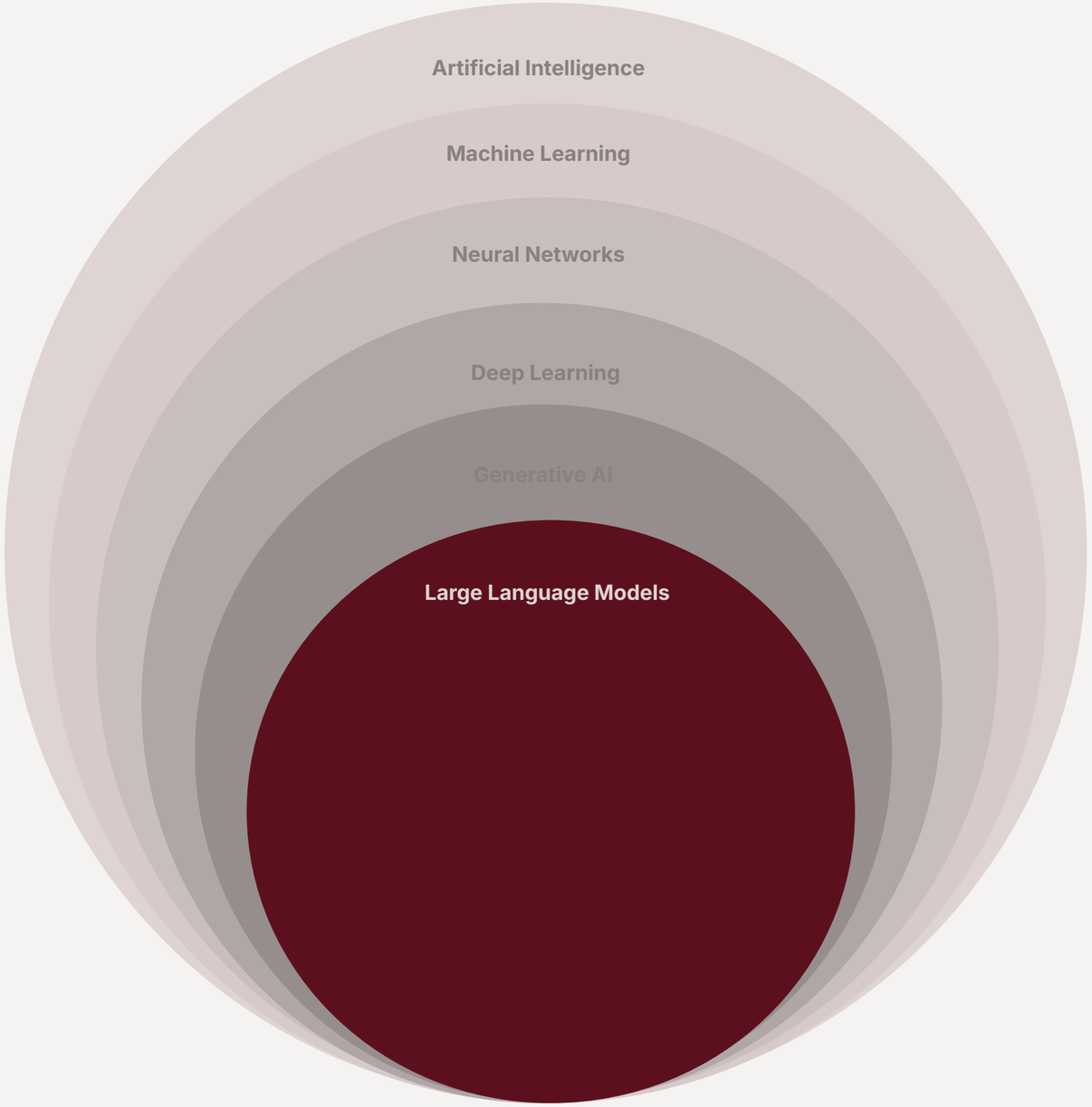
Backend dev e facilitador de AI

SLM: Small Language Models

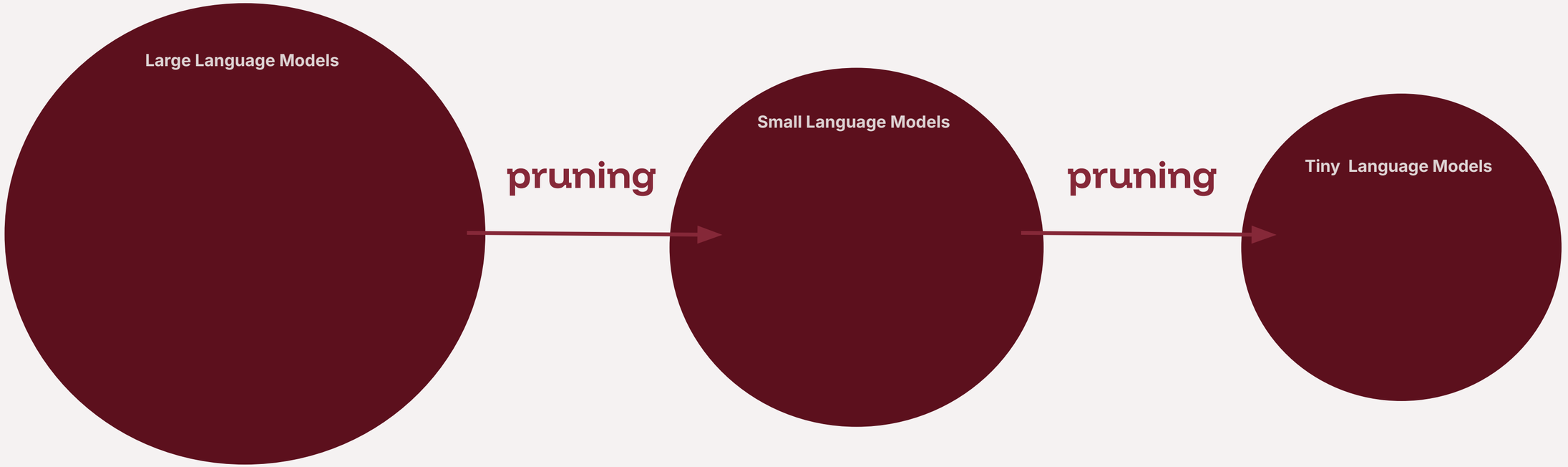
SLM: Hierarquia da Inteligência Artificial



Paradigma dos Grandes Modelos



Pruning (poda)



SLM: Vantagens

Menos parâmetros

+

Menor gasto de memória

=

Roda em máquinas mais simples

+

Mantém quase mesma acurácia

Modelos de AI: Comparativo

Categorias	Número de Parâmetros	Exemplos	Onda roda bem?
LLM	70B – 671B+	GPT-4, LLaMA-70B, DeepSeek R1 full	**Data centers com GPUs A100/H100, clusters caros**
SLM	~1B – 30B	DistilBERT, Mistral-7B, Phi-3-mini, DeepSeek distill	**PCs high-end (ex.: RTX 4090), servidores médios, alguns notebooks com GPU dedicada**
Tiny LM	< 1B (milhões de parâmetros)	TinyBERT (~14M), MobileBERT (~25M), NanoGPT (~10-100M)	**Smartphones, notebooks leves, placas embarcadas (ESP32, Raspberry Pi, ARM Cortex-M)**

Ollama

Ollama

O que é o Ollama?

O Ollama é uma ferramenta de código aberto que **gerencia** e **executa modelos de inteligência artificial localmente**, mantendo nas mãos do usuário o controle e a privacidade dos dados.

Por rodar localmente o **usuário mantém controle total dos dados** transmitidos e compartilhados com o modelo, evitando possíveis riscos de segurança ao enviar informações para a nuvem.

www.ollama.com

ZUP

Sua aliada tech
do agora ao futuro.