Why use open source models?
1. They’re pretty good at important tasks

   e.g. summarization, document extraction, CoT for basic reasoning, etc.
2. More control thanks to rapidly growing ecosystem of tools

Inference Servers: vLLM, TRT, Llama.cpp, etc.

Structured Output: Outlines, Guidance, JSONFormer

Local models: Ollama, Llama.cpp

Custom logic over primitives: KV caching, batching (speculative sampling!!), SGLang, DSPy, fun w/ logprobs
3. Different cost structure than pay per token

Tokens are expensive! What happens when you mess up…

Instead, pay per GPU. Amortize token cost over hardware.

Significantly cheaper at high enough load.
4. It’s just more fun

State space of possible projects is larger and more interesting

Follow the craziness with models, data, and training schemes
Quick aside on GPUs
2 things: GPU VRAM and memory bandwidth
1 main thing: GPU VRAM
2 * (num_billion_params) GB of VRAM
H100 > A100 > A10 > T4
Why not use open source models?
1. Closed source has better reasoning

Less time spent worrying about prompting or if a model is right for the job
2. Closed source *can* be cheaper

Depends on user traffic and load

You may pay for idle time or cold starts with open source
3. Open source can be a pain to host

Set up a cluster with load balancing
How do you optimize GPU usage?
Prompts aren’t necessarily interoperable
Baseten makes open source easy.