June 10, 2025

Unlocking Open Source

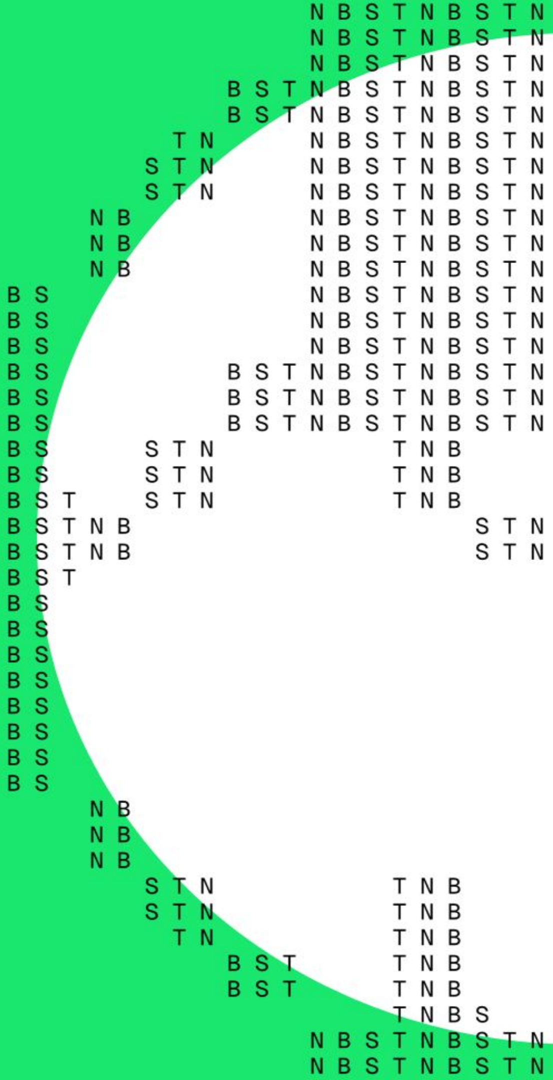
Het Trivedi
Forward Deployed Engineer

Philip Kiely
Developer Advocate

Agenda

- Why switch to open source?
- How to switch to open source
- Options for adopting open models
- Example: rebuilding an agent
- Testing and iteration
- Q&A

Why switch to open source?



Where do AI engineers start?

- Closed models on OpenAI/Anthropic/Google
- Startups: public APIs
- Enterprises: provisioned throughput Azure/Bedrock/Vertex

By default, AI engineers don’t want to change from this known setup

But now we have to

2023: “Toying around”

2024: Production with closed models

2025: Cracks in this approach

Where do people think cracks are?

- Vendor lock-in
- Ballooning cost
- Compliance
- Privacy
- Security

If these aren’t the cracks, what are?

Why build on open source?

- **Quality** (task-specific)
- **Latency** (for real-time use cases)
- **Economics** (at scale)
- **Differentiation** (control of destiny)

Quality

- Frontier open source has closed the gap
- Task-specific quality is differentiator for product
- Example: healthcare document processing

Latency

- Use cases are increasingly latency sensitive
- Endpoints are optimized for system throughput
- Example: small businesses answer every phone call

Economics

- Unit economics matter at scale
- Price taker → price maker
- OSS: 60-90% cost savings for same quality

Destiny

- Every company is now an AI company
- Don’t outsource ownership of AI strategy
- AI as differentiated alpha

How to switch to open models

Why → How

- **Before:** Need an OpenAI API key
- **After:** Scale up inference, do so quickly and cost-efficiently

vLLM$_{\text{(et al)}}$ + GPU != Production

Challenges

Performance

- Latency (guaranteed p99 TTFT)
- Throughput

Infrastructure

- 99.99% availability despite GPU failures
- Fast scale-ups for traffic bursts

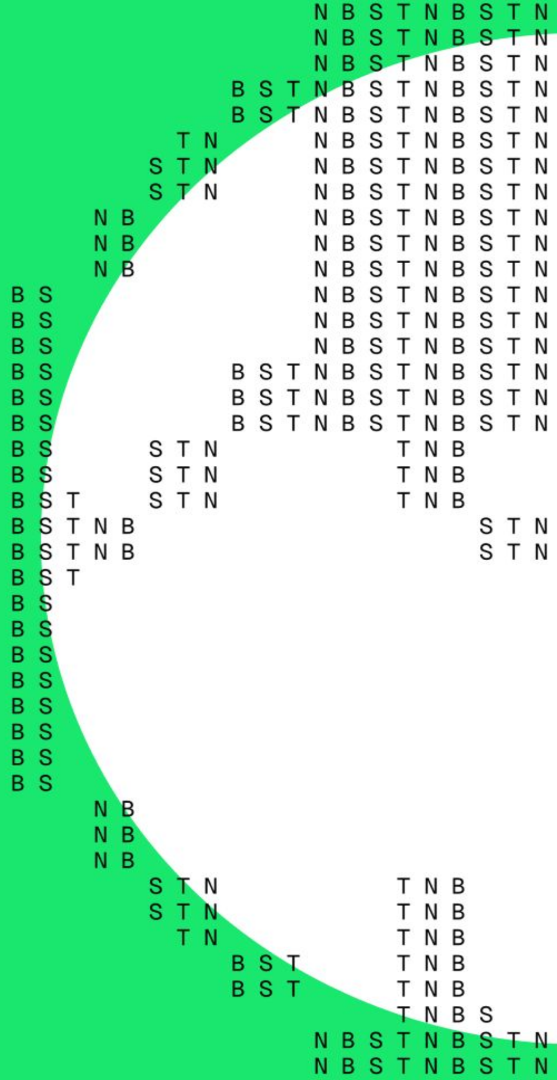
Evals

- Is the model good for your product?
- Does every task need the smartest model?

Product

- Do you need to change anything in code?
- Do you need to change prompts?

Options for open source



Model APIs

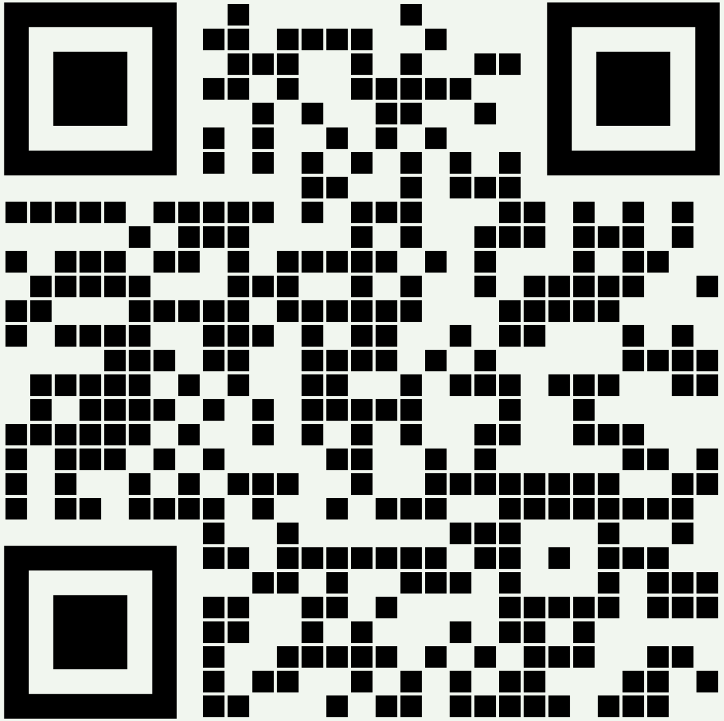
- Low to moderate volume
- Done-for-you performance
- Zero overhead
- Secure infrastructure
- Access frontier open models

Dedicated

- High volume
- Tight SLAs (e.g. p99 TTFT)
- Cost at scale
- Deploy in your VPC
- Custom models / fine-tunes / models we don’t support by API (4 → 1,000,000)

You already know AI engineering. We’ll prove you already know how to build on open source.

Example:
Build an agent



Testing and iteration

Performance benchmarks

Time to first token (TTFT)	Tokens per second (TPS)
DeepSeek V3: 300 ms	DeepSeek V3: 40

- TTFT + TPS = responsive agent
- 1 user action → 50 inference requests
- Show users intermediate steps for agents

Quality evals

- Head-to-head eval strategy
    - White belt: standard benchmarks
    - Blue belt: LGTM (vibe check)
    - Brown belt: consistency across runs
    - Black belt: product-specific evals
- Clear outcome-based agent evals

What to build from here

- Black belt evals (Patronus, Braintrust)
- Task-specific fine-tuning (Oxen)
- Multi-model, multi-step agents (Baseten chains)
- Integrations with ecosystem/agent frameworks

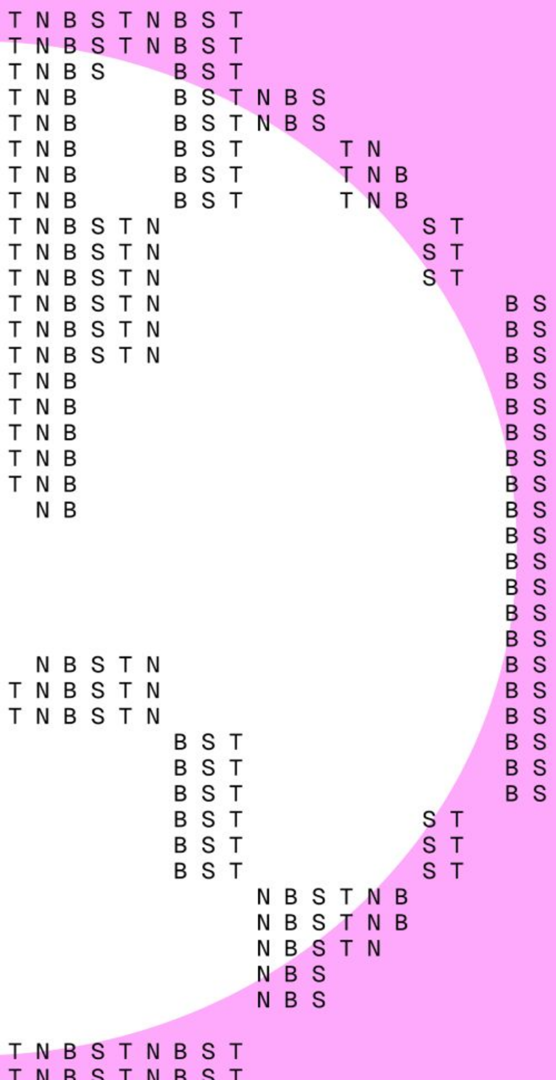
Open model options

Model families (LLMs):

- DeepSeek
- Llama
- Qwen
- Mistral
- Gemma

Model modalities:

- Vision
- ASR (transcription)
- Speech synthesis
- Embedding
- Image/video



Questions?

Thank you

x.com/basetenco

linkedin.com/company/baseten

x.com/het_trivedi05

linkedin.com/in/het-trivedi05

x.com/philip_kiely

linkedin.com/in/philipkiely