Local Models & Self-Evolving Harnesses

Machine Learning Singapore #MSLG

- Martin Andrews = Harnesses @ mdda.net

25-June-2026

About Me

- Machine Intelligence / Startups / Finance
  - Moved from NYC to Singapore in Sep-2013
- 2014 = 'fun' :
  - Machine Learning, Deep Learning, NLP
  - Robots, drones
- Since 2015 = 'serious' :: NLP + deep learning
  - Including Papers...
  - & GDE ML; ML-Singapore co-organiser...
  - & Red Dragon AI...

MLSG
MACHINE LEARNING
SINGAPORE

Outline

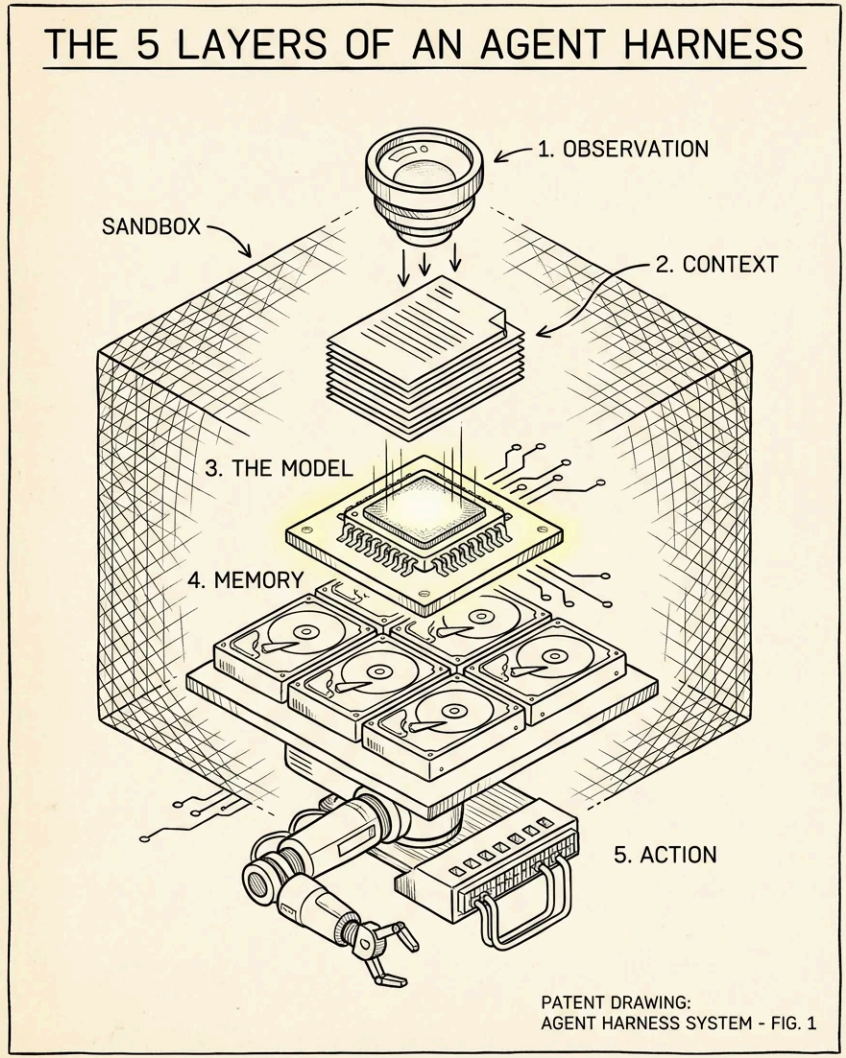
- Basic Harnesses
- Self-Modifying Harnesses
- Token Sourcing Risks
  - ... and mitigation
- Wrap-up & QR-code

MLSG
MACHINE LEARNING
SINGAPORE

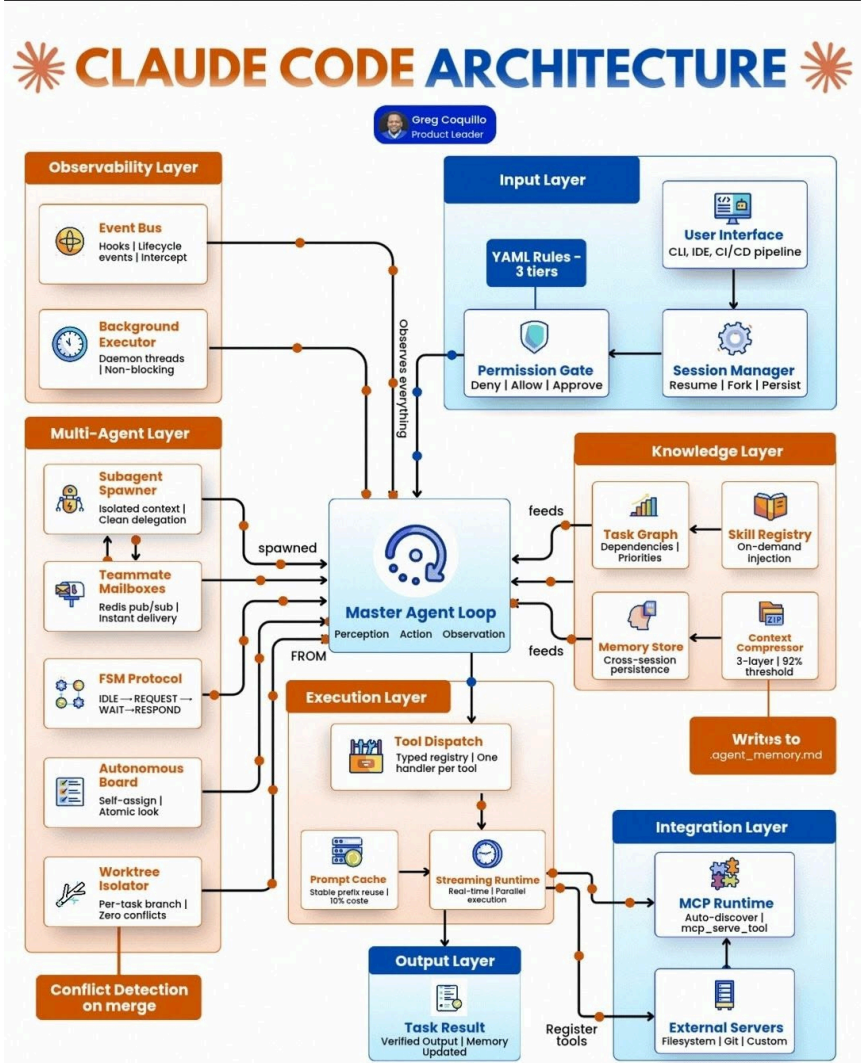
Basic Harnesses

What is a Harness?

- Harnesses provide:
  - software infrastructure
  - runtime environment
  - engineering layer
- Wraps around a raw LLM to make it:
  - functional, reliable, and usable
- i.e. : Everything except the model itself



MLSG
MACHINE LEARNING
SINGAPORE

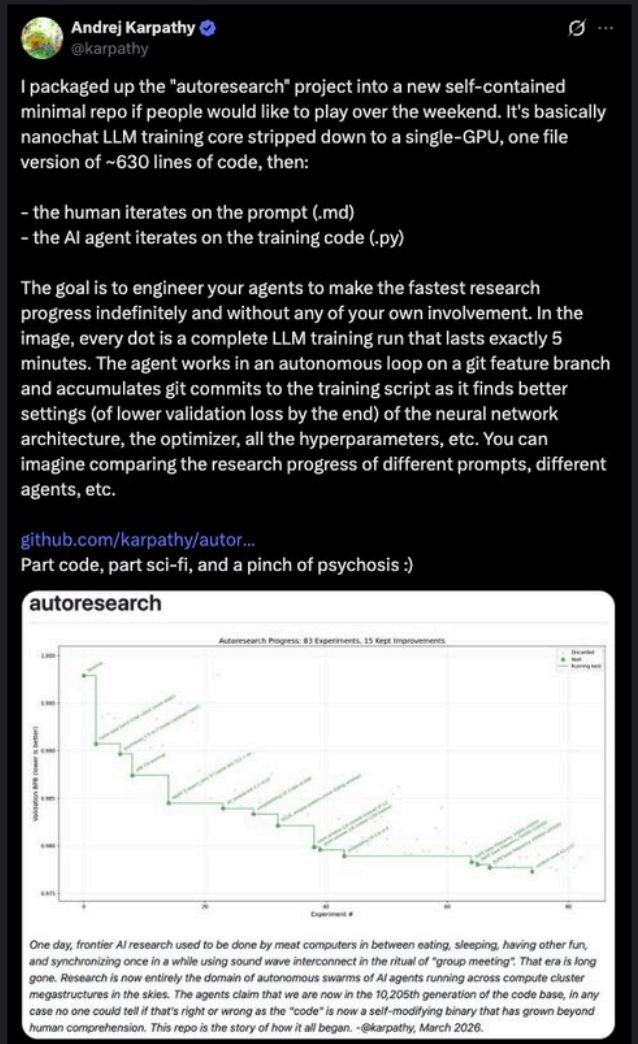


Claude Code

- Coding agent system
- Can launch subagents
  - write tests
  - loop until 'satisfied'
- Can also author new SKILL.md files
  - but this is an extra request

AutoResearch

- Started last year
  - Darwin Gödel Machine (May-2025)
    - Sakana AI
  - AlphaEvolve (June-2025)
    - DeepMind : A coding agent for scientific and algorithmic discovery
  - GPU Kernel Scientist (June-2025)
    - AMD GPU kernels...
- Karpathy meme version (March-2026)
  - simple repo deployment FTW
  - slick git branching for experiments
    - target : Karpathy's own `nanochat` experiments



MLSG
MACHINE LEARNING
SINGAPORE

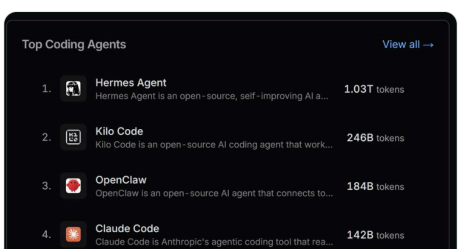
OpenClaw

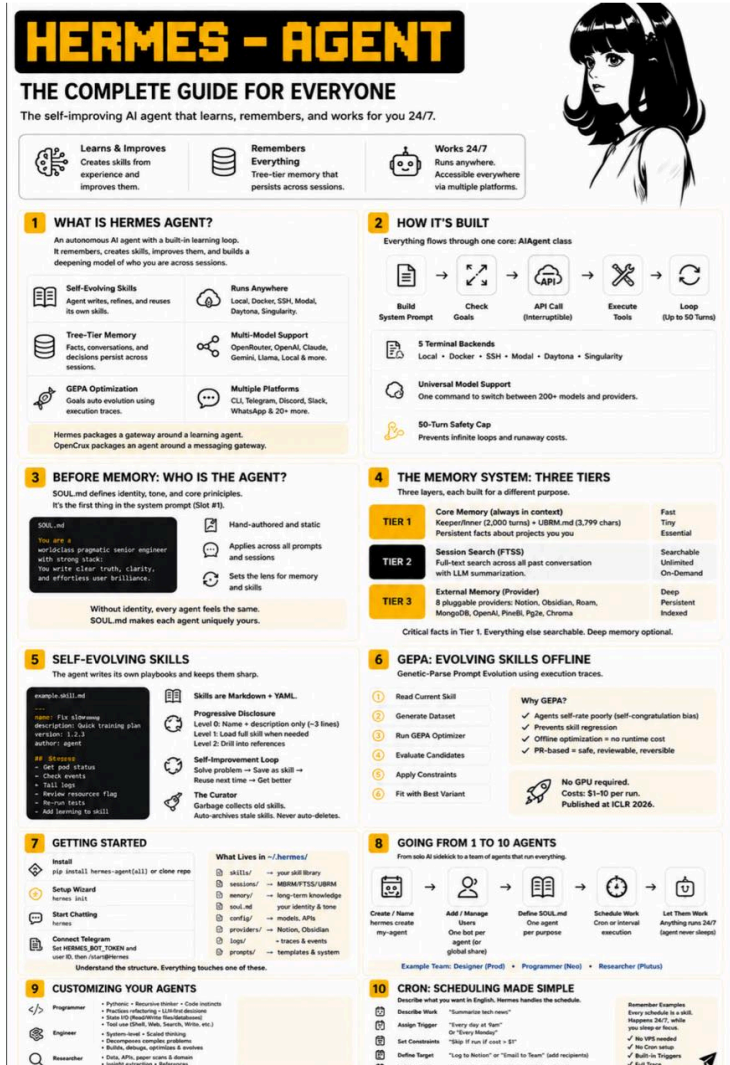
- Personal Agent loop = Inspirational
  - fka ClawdBot ~Jan-2026
    - ClawCon = 4-Feb-2026 in SF
    - MLSG night = 26-Feb-2026
    - Peter Steinberger now an OpenAI employee
  - caused a run on the Mac-Mini
    - security issues for the unwary
    - ... says Cyber Security Agency of Singapore
- Still lots of hype around the project
  - there's even an event on tonight!
    - "No Technical Background Required"

Hermes

- New entrant from Nous Research
  - Project Page, and MIT licensed code repo
  - Pronunciation:
    - "noose" & "her-mees", not "noo" & "ermez"

- Key Features
  - a 'better' OpenClaw
    - tasteful software choices
    - batteries included (but disabled by default)
  - self-modification
    - automatically reviews what went well
    - and updates itself for future you



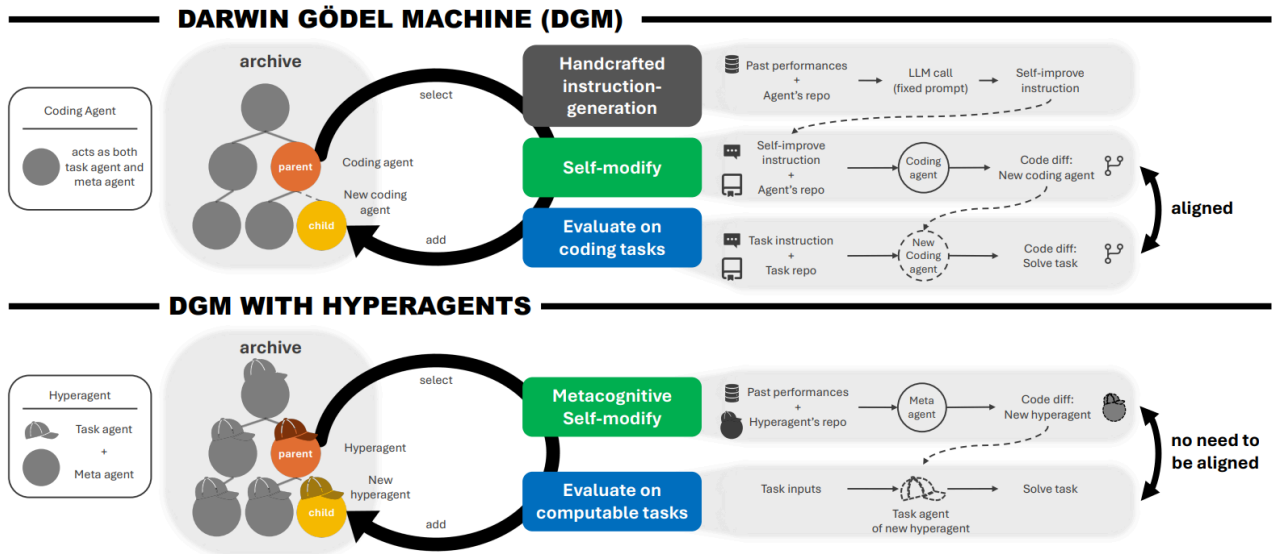


MLSG
MACHINE LEARNING
SINGAPORE

Self-Modifying Harnesses

Darwin Gödel Machine - Hyperagents

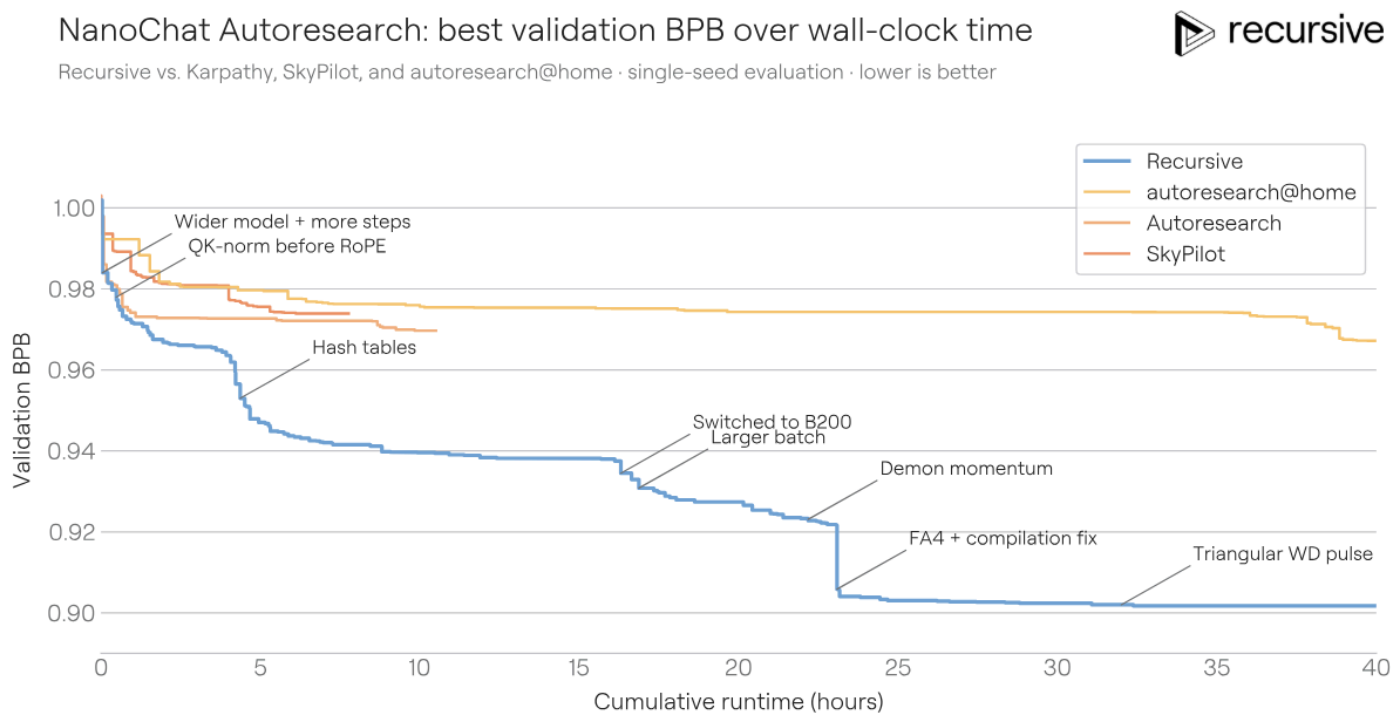
- Hyperagents - Zhang *et al* (03-2026)
  - Jenny Zhang @ Meta (prev MLSG speaker)
    - with Jeff Clune
  - Facebook Research Code Repo
- Expands "AutoResearch" in a Meta direction
- Evolves an agent to do the task
  - *and* the update harness too
    - i.e. can specialise harness to task too
    - hence the 'Hyper level'
  - Can choose to add:
    - memory; performance tracking; etc ...



MLSG MACHINE LEARNING SINGAPORE

DGM-HyperAgents ++

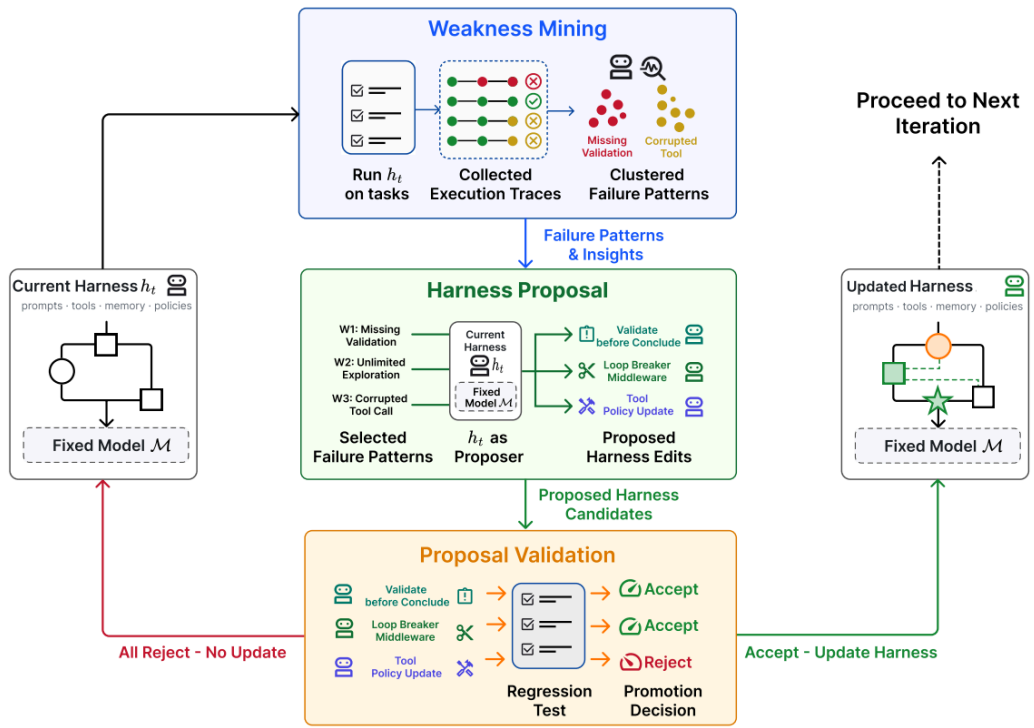
- Jenny Zhang (and others) now at Recursive Superintelligence, Inc.
  - RSI raised $650 million Series @ $4.65 billion post-money



MLSG
MACHINE LEARNING
SINGAPORE

Meta-Harness / Self-Harness

- Meta-Harness: End-to-End Optimization of Model Harnesses - Lee *et al* (03-2026)
  - with Omar Khattab (DSPy + Recursive Language Models)
    - and Chelsea Finn (Stanford)
- Self-Harness: Harnesses That Improve Themselves - Zhang *et al* (06-2026)
  - a framework that lets AI agents rewrite their own rules
    - boosting performance up to 60%



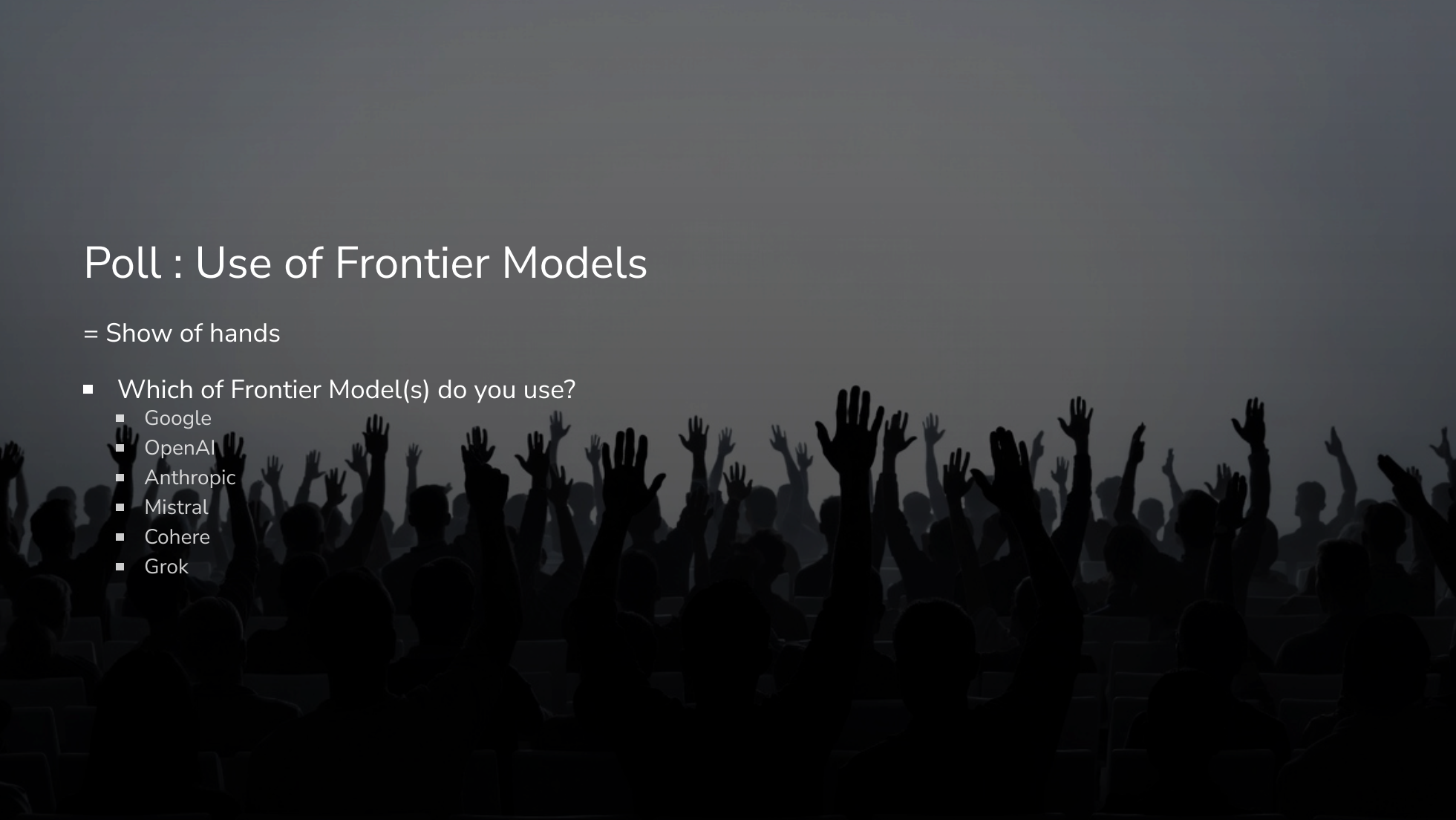
MLSG
MACHINE LEARNING
SINGAPORE

Token Sourcing Risks

Poll : Use of Frontier Models

= Show of hands

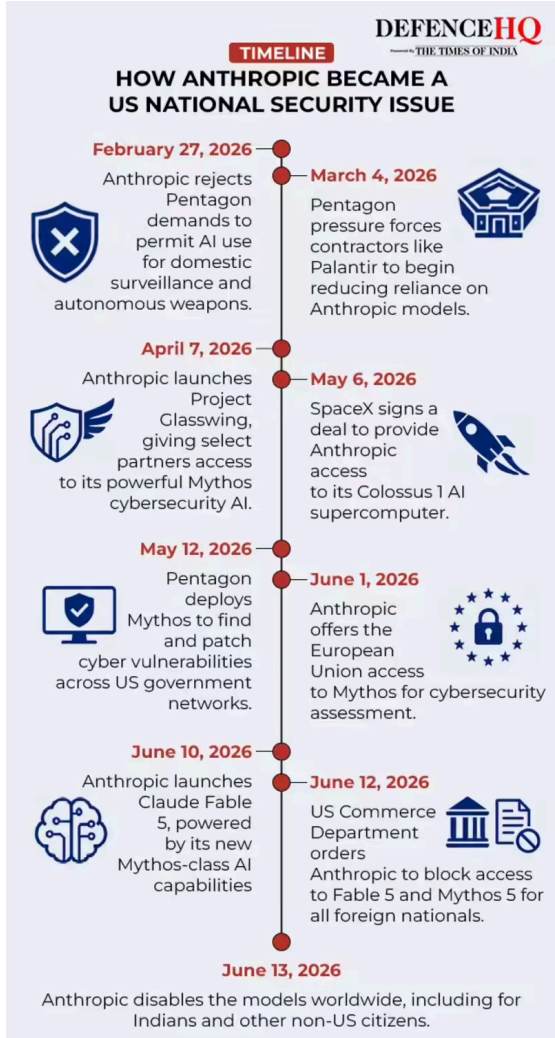
- Which of Frontier Model(s) do you use?
  - Google
  - OpenAI
  - Anthropic
  - Mistral
  - Cohere
  - Grok



The Mythos/Fable saga

- Walk through Anthropic's timeline:
  - defense department
    - OpenAI takes up 'lawful use' position
  - Mythos model finds security issues
    - alarmingly many...
  - Fable released with guardrails
    - not just refusal!
    - AI researchers now feel the pinch
  - Amazon finds 'jailbreak'
    - enough to give leverage to US Government
  - Now blocked for everyone!

- PS: Don't assume OpenAI is safe choice ...



MLSG
MACHINE LEARNING
SINGAPORE

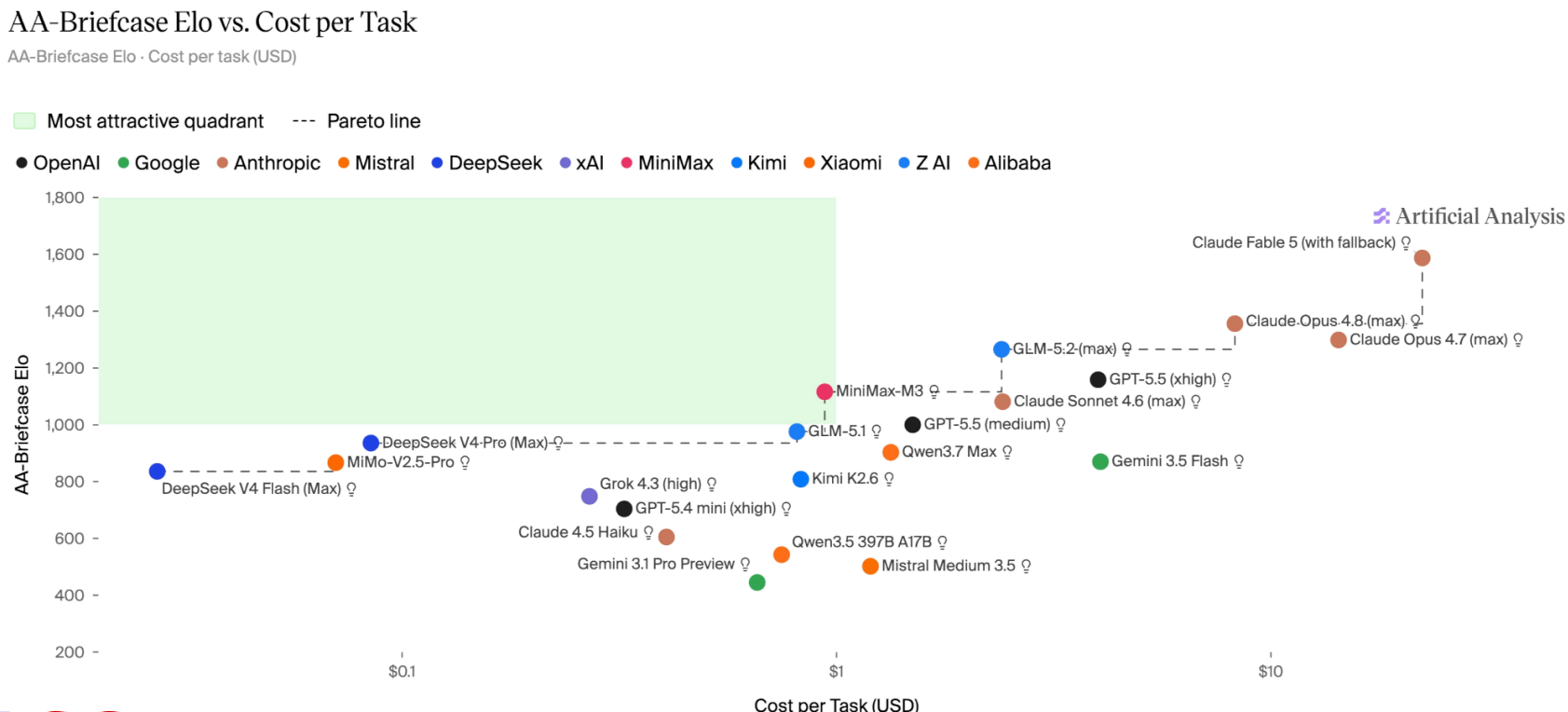
Owning Your AI

Who should be thinking about this?

- Who?
  - Individuals
  - Companies
- Why?
  - cost
  - privacy
    - IP leakage
    - security
  - speed
- Where?
  - cloud (multiple providers?)
  - local (fully costed?)

MLSG
MACHINE LEARNING
SINGAPORE

Model capability / cost frontier

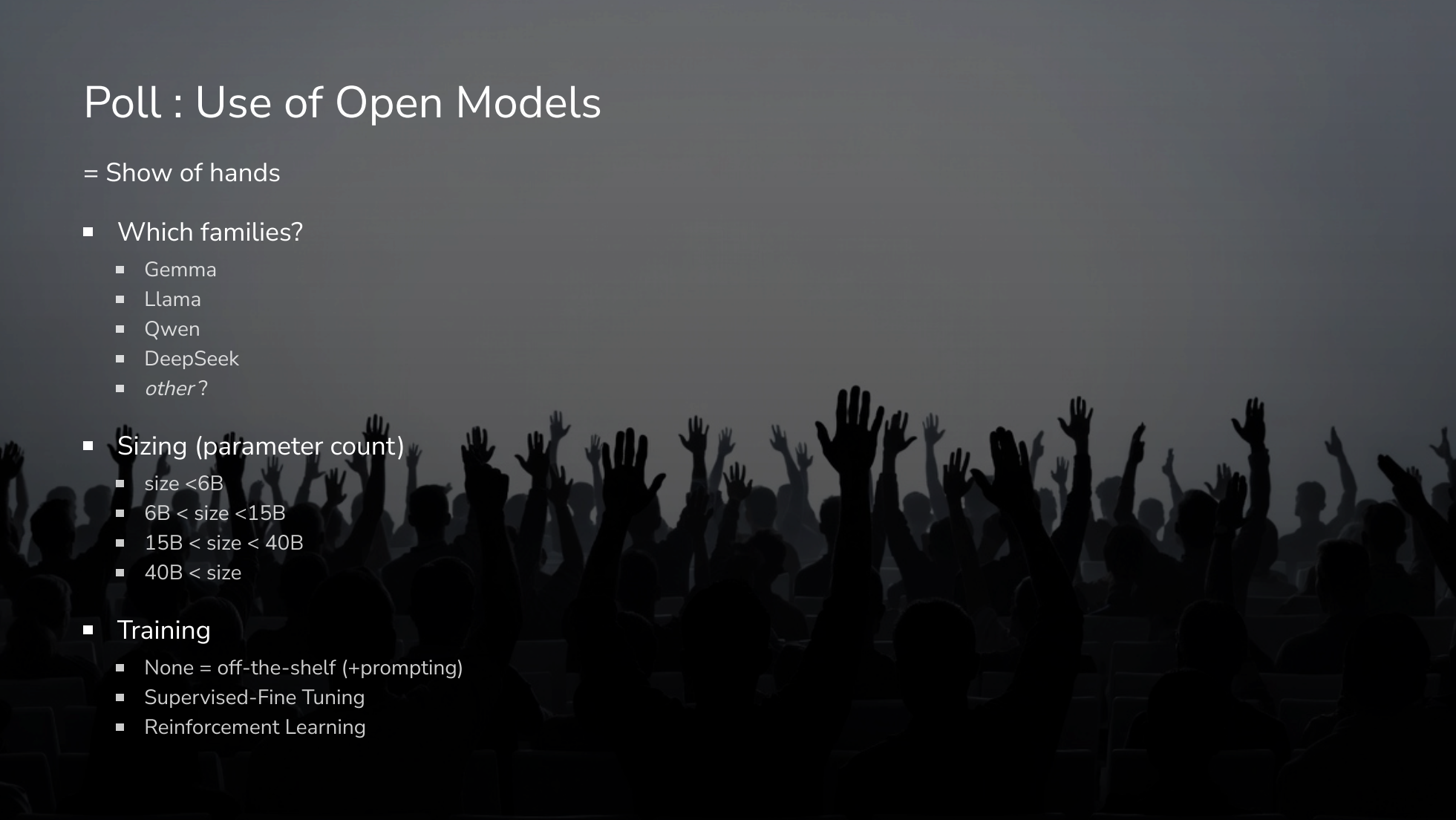


MLSG
MACHINE LEARNING
SINGAPORE

Poll : Use of Open Models

= Show of hands

- Which families?
  - Gemma
  - Llama
  - Qwen
  - DeepSeek
  - *other* ?
- Sizing (parameter count)
  - size <6B
  - 6B < size <15B
  - 15B < size < 40B
  - 40B < size
- Training
  - None = off-the-shelf (+prompting)
  - Supervised-Fine Tuning
  - Reinforcement Learning



Integrating Open Source Models

- Problems with Foundation company lock-in:
  - cost
  - privacy
  - and now : country risk
- Solution
  - Use different models for different tasks
    - eg: Foundation for planning
      plus Open Source models *increasingly*
  - Can do prompt / harness optimisation for both
  - But, with Open Source, you can:
    - train model on your local hardware
      - or in the cloud
    - train smaller models for repetitive tasks
      - for even bigger savings

MLSG
MACHINE LEARNING
SINGAPORE

Who agrees?

- Satya Nadella (Microsoft CEO)
  - published an article on Twitter/X
  - now advocating for an ecosystem approach
- Frontier Tuning: Teaching AI to work the way you do
  - announced by Microsoft on 2-June-2026
  - (no, not Open Source)
- Clearly, there's some tension with OpenAI...
  - and the Microsoft story makes sense
    - ... particularly with Open Source models

Satya Nadella reveals why every company may need its own AI model: the model becomes the new company database.

"To me, a model is like the database market."

"A firm should be able to take the tacit knowledge it has and embed it inside weights in a model that they control."

"When somebody asks me how many models should there be, I'll say as many models as firms in the world."

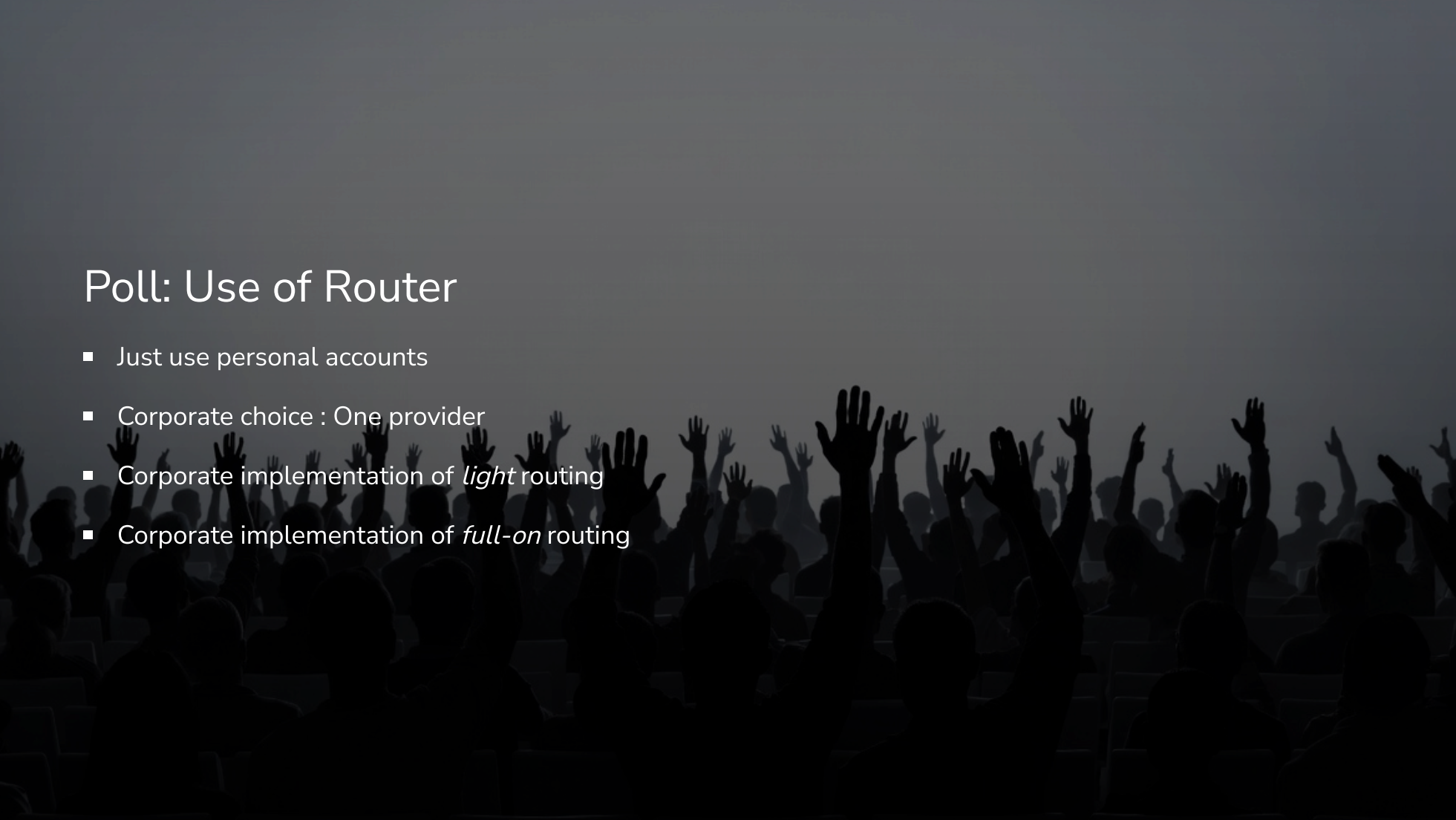
The contrarian part: the value may not sit in one universal frontier model. It sits in each company turning its private operating knowledge into a controlled model.

MLSG
MACHINE LEARNING
SINGAPORE

'Simple' Router approach

- Foundation model choices
  - which model
  - model capability per brand
- Monitor
  - tokens spent / by whom
- Routing
  - multiple sourcing
  - Personally Identifying Information redaction
- More capable *full-on* version
  - understanding caching
    - manage routing more carefully
  - usefulness of answers
    - = training data

MLSG
MACHINE LEARNING
SINGAPORE

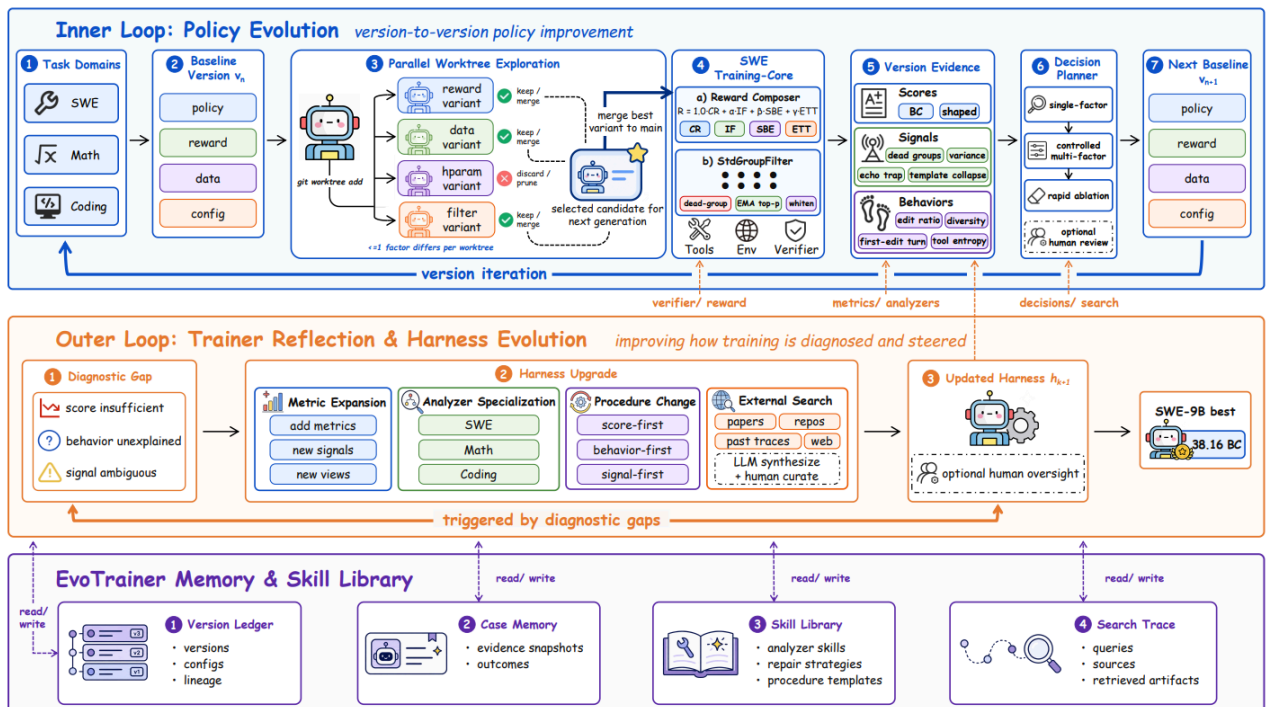


Poll: Use of Router

- Just use personal accounts
- Corporate choice : One provider
- Corporate implementation of *light* routing
- Corporate implementation of *full-on* routing

EvoTrainer

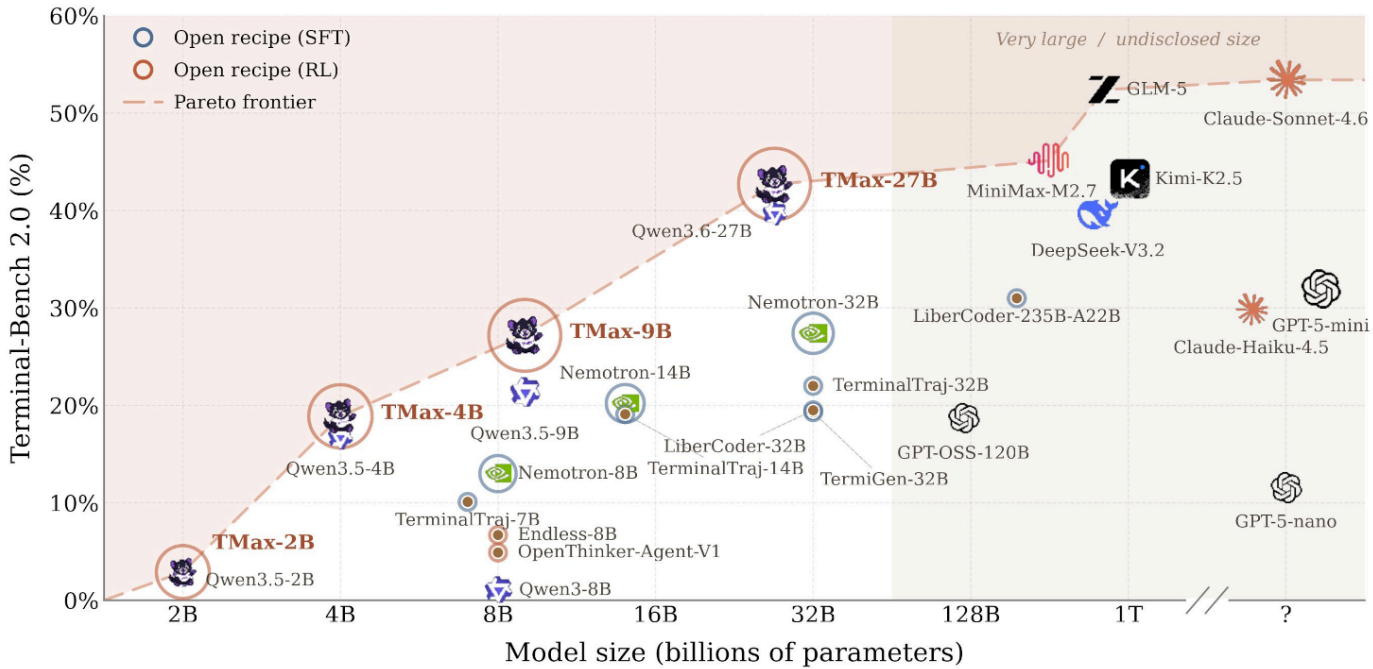
- EvoTrainer: Co-Evolving LLM Policies and Training Harnesses for Autonomous Agentic Reinforcement Learning - Chen *et al* (2026)
- RL over full harness
  - train models holistically
  - based on local outcomes
- Ideas:
  - bootstrap using Foundation Models
  - evaluate "local" and Foundation Models on same task
  - can focus on difficult cases
  - or large model fall-back
- Own your own AI
  - rather than sending IP to SF



MLSG
MACHINE LEARNING
SINGAPORE

How feasible is RL inside the enterprise?

- TMax: A Simple Recipe for Terminal Agents + Apache 2.0 Code



MLSG
MACHINE LEARNING
SINGAPORE

Wrap-Up

- Harnesses can greatly enhance system performance
  - (even more so if you can specialise them)
- Recent geopolitical events should make everyone nervous ...
  - ... about relying on other people's AI
- Companies should also be thinking about Owning Their AI
  - and it's possible to implement this locally

MLSG
MACHINE LEARNING
SINGAPORE

Link to Slides

- https://bit.ly/MLSG_2026-06

MLSG
MACHINE LEARNING
SINGAPORE