Learning from Decision Models

Machine Learning Singapore #MLSG

- Martin Andrews = LearningDecisionModels @ mdda.net

26-November-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

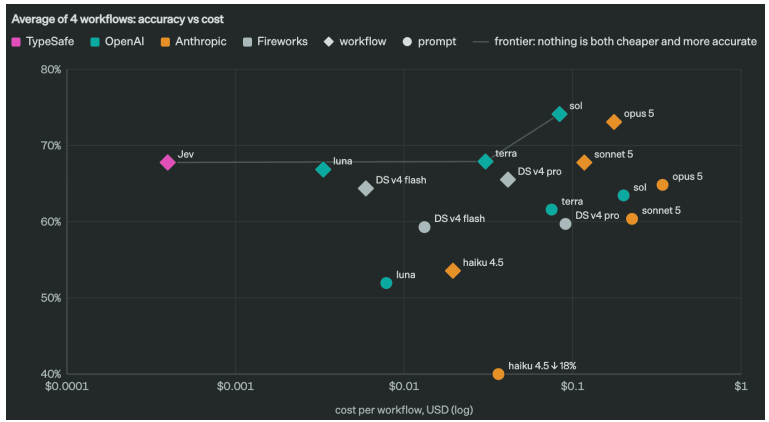
- Decision Model basics
- Why go local?
  - how it works
  - demo
- In the news...
- Wrap-up & QR-code

MLSG
MACHINE LEARNING
SINGAPORE

What is a Decision Model?

TypeSafe AI : Introducing System One Models & Jev





- Introducing System One Models & Jev - Diogo Almeida / TypeSafe AI (15-Sep-2026)
  - Idea : *make judgements inside software*
    - YouTube Interview : "We Could Have Had AGI Last Year"
  - Many Enterprise tasks don't need chat
    - need fast, typed judgements on state (`alert` , `email` , `PR` , `listing`)
    - Decision models : Zero structured-output parse errors, at ~100× lower workflow cost

MLSG
MACHINE LEARNING
SINGAPORE

A New Class of Model : "System 1" Decisions

- Different from an Autoregressive LLM:
  - *"Your LLM knows the answer before it writes a single token - so we stopped making it write"*
  - Autoregressive `model.generate()` bottleneck:
    - decodes token-by-token
      - allocates KV cache across steps
    - brittle JSON / regex string parsing
    - *uncalibrated* confidence when sampling text
  - System 1 Readout:
    - single forward pass
    - reads exact logits at decision probe positions
    - scores **many questions in parallel** on one shared context

- What is a **Decision**?
  - Given a state + schema of questions
    → typed, calibrated outputs:
    - `noul` : Yes/No probability (`has_spill: 0.95`)
    - `choice` : distribution over options `(A, B, C, D)`
    - `score` : bounded values (`severity: 1.19 / 2.00`)

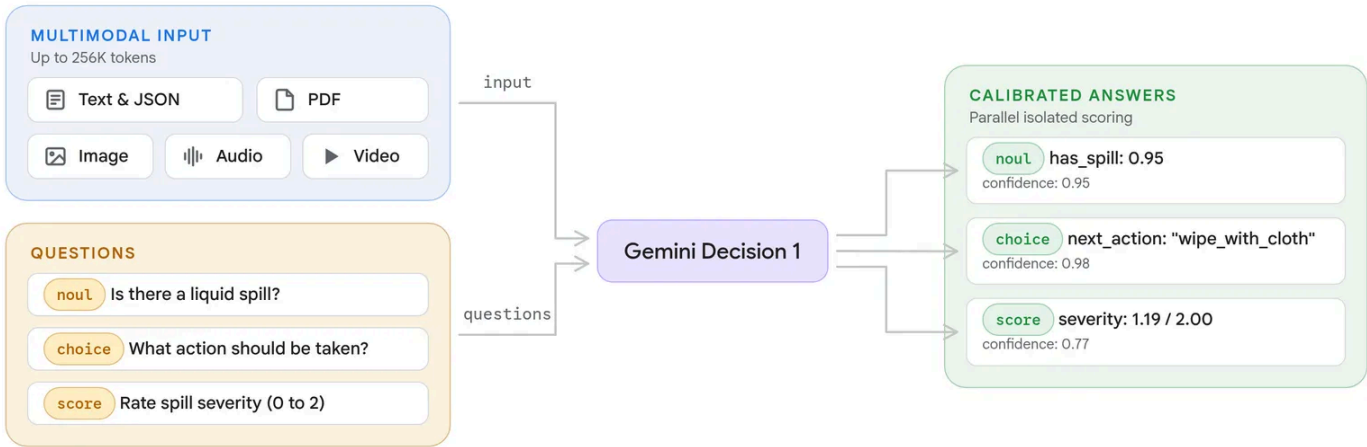
- Everyone is suddenly building these!
  - Turing Post : 15 System-1 Models Beyond Jev (Oct-2026)
    - routers, guardrails, triage, e-commerce catalog audits
    - HF `transformers serve /v1/systemone` (`@_marcsun`)

MLSG
MACHINE LEARNING
SINGAPORE

Gemini Decision 1

MLSG
MACHINE LEARNING
SINGAPORE

Google Validates the Space : Gemini Decisions API



- Gemini Decisions API: Fast, calibrated multimodal decisions - Google (11-Nov-2026)
  - Recently announced :: `Gemini Decision 1`
    - native multimodal input (`Text` , `JSON` , `PDF` , `Image` , `Audio` , `Video` up to `256K` tokens)
    - evaluates up to 1,000 isolated questions in a single pass (`noul` , `choice` , `score`)
    - zero output tokens → calibrated probabilities & confidence scores

MLSG
MACHINE LEARNING
SINGAPORE

`Gemini Decision 1` : Benchmarks & Calibration

- SOTA accuracy & calibration:
  - 92.2% on `JevBench`
    - `90.0%` TypeSafe Jev 1.14
    - `90.9%` OpenAI GPT-6 Luna
  - `2×` better calibrated than Jev 1.14
  - native audio + video
    - Jev / Luna are N/A

	Benchmark		Gemini Decision 1 Decisions API	TypeSafe Jev 1.14	GPT-6 Luna OpenAI Decisions API	Gemini 3.5 Flash-Lite Interactions API (minimal)
Key benchmarks	JevBench	Public split	**92.2%**	90.0%	90.9%	86.6%
	Amazon ESCI	Macro-F1	**58.7%**	56.7%	54.7%	47.2%
	New Yorker Caption	Text-only	**79.7%**	72.0%	68.4%	73.7%
	Mind2Web Multimodal		86.2%	**86.8%**	81.8%	78.0%
	PubMedQA		**79.8%**	78.2%	74.4%	77.6%
	Global-MMLU-Lite	Non-English	**86.0%**	85.0%	79.6%	78.8%
	When2Call		79.9%	**82.0%**	70.4%	60.9%
	MMAU		**71.8%**	—	—	67.7%
	Video-MME	No subtitles	**69.1%**	—	—	63.3%
Calibration	Top-1 ECE	7 shared evals, lower is better	**0.0250**	0.0494	0.0721	0.2104
Cost	Total price	7 shared evals	$0.88	**$0.68**	$0.93	$4.14

Refusals, safety blocks, and requests a model could not accept are scored as incorrect; every score uses the full benchmark as the denominator. Scores are Top-1 Accuracy unless otherwise noted.

MLSG
MACHINE LEARNING
SINGAPORE

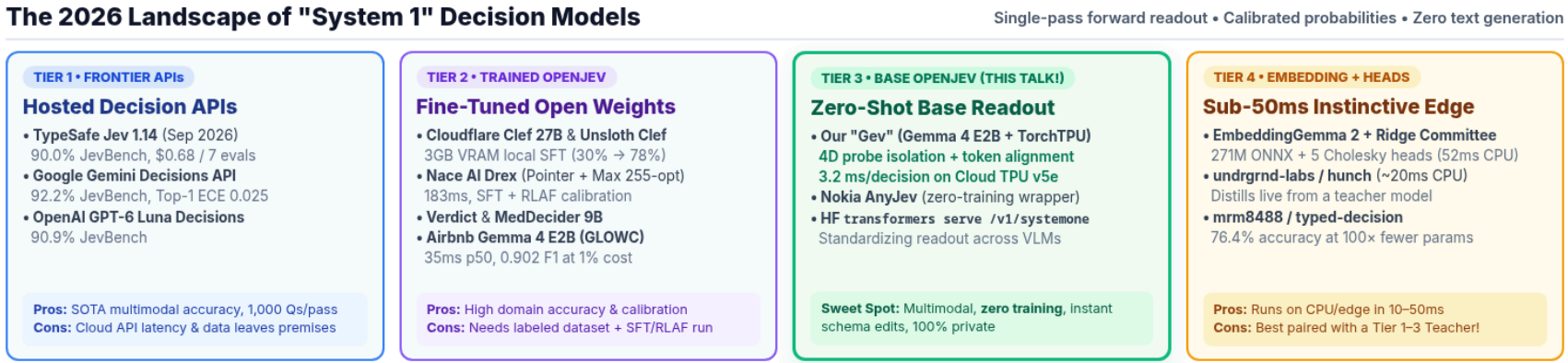
Code example : Calling Gemini Decision 1

```python
from google import genai                                    # 1
client = genai.Client(http_options={"api_version": "v1beta"})

res = client.decisions.create(                              # 2
    model="gemini-decision-1",
    input="The blue mug arrived cracked.",                  # 3
    questions={
        "damaged": {                                        # 4
            "type": "noul",
            "instructions": "Arrived damaged?"},
        "queue": {
            "type": "choice",
            "instructions": "Which queue?",
            "criteria": {"billing": "…", "returns": "…"}},
        "severity": {
            "type": "score",
            "instructions": "How severe?",
            "criteria": ["minor", "moderate", "critical"]},
    },
)
p = res.answers["damaged"].noul                             # 5
q = res.answers["queue"].choice
```

- Gemini Decisions API (`google-genai ⩾ 3.2`)
  - *almost drop-in* vs TypeSafe Jev (`typesafe-sdk`)
- Only **5 small changes** vs 'real' Jev:
  - Client : `genai.Client(v1beta)`
    - Jev : `TypeSafeClient()`
  - Call : `client.decisions.create( … )`
    - Jev : `client.system_one(model="jev-latest", … )`
  - Context : `input=` *(text, image, audio, video, PDF)*
    - Jev : `state=` *(text / JSON)*
  - Questions : dicts with `"type": "noul"|"choice"|"score"`
    - Jev : `Noul( … )` , `Choice( … )` , `Score( … )`
  - Outputs : unified `res.answers[k]`
    - Jev : split `res.nouls[k]` , `.choices[k]` , `.scores[k]`

Tier List of System-1 Decision Models



- Four distinct tiers have emerged in just 8 weeks:
  - Tier 1 (Hosted Frontier APIs) : TypeSafe Jev, Google Gemini Decisions, OpenAI Luna
  - Tier 2 (Trained OpenJev) : Drex, Verdict, `Cloudflare Clef 27B`, Unsloth Clef-head, Airbnb Gemma 4 E2B
  - Tier 3 (Base OpenJev = *Our Gev*) : zero-shot single-pass readout (`Gemma 4 E2B` , Nokia AnyJev)
  - Tier 4 (Embedding + Heads) : `EmbeddingGemma 2` + online ridge heads / hunch

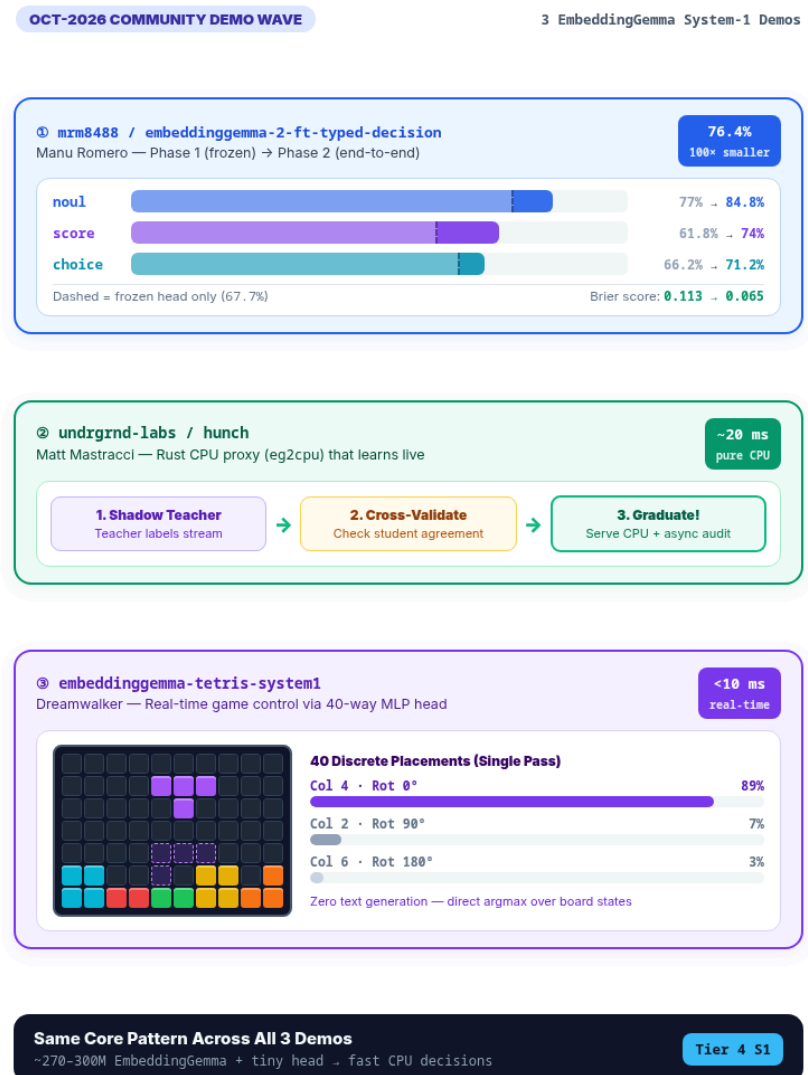
MLSG
MACHINE LEARNING
SINGAPORE

Fast Local Decisions

sub-50ms On-Device Decisions

Oct-2026 Demo Wave

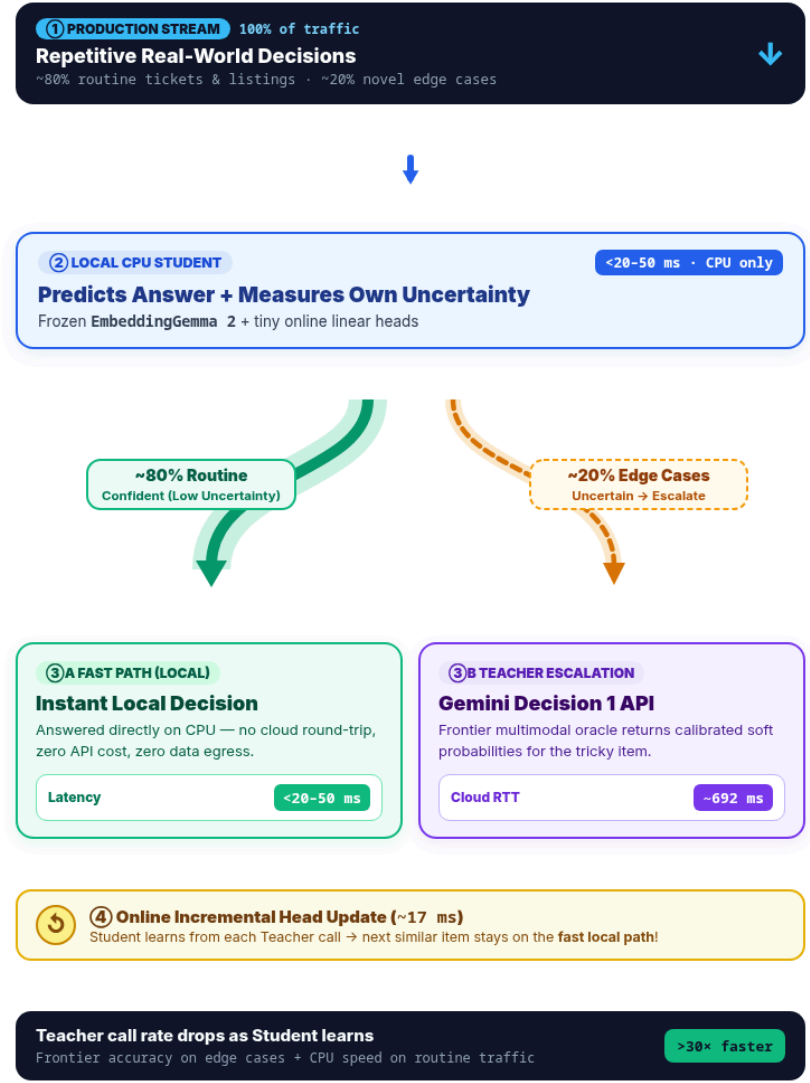
- `mrm8488/embeddinggemma-2-ft-typed-decision` - Manu Romero
  - state + question schema → `choice · score · noul`
  - `76.4%` on `typed-decisions`
    - `~6 pts` behind a 27B LLM with `100×` fewer parameters
- `undrgrnd-labs/hunch` - Matt Mastracci
  - Jev-like proxy in `~20ms` on CPU (no GPU)
    - learns live from a teacher model!
- Sub-10ms "System 1" Tetris Engine - Dreamwalker
  - real-time game decisions from embeddings alone



MLSG
MACHINE LEARNING
SINGAPORE

Can We Train a CPU Student from `Gemini Decision 1`?

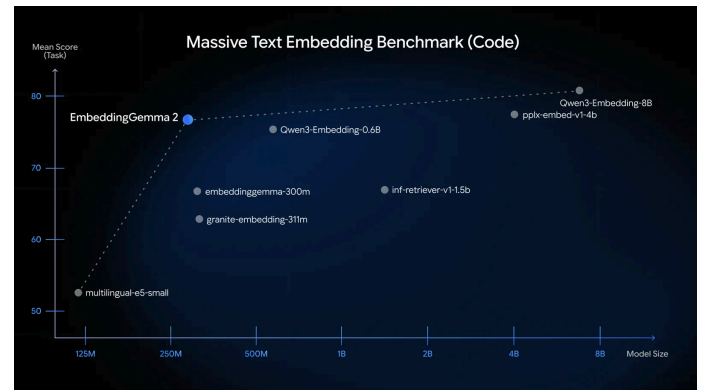
- Most production streams are highly repetitive
  - `~80%` routine tickets / listings, `~20%` novel edge cases
- Calling `Gemini Decision 1`
  - ... requires a cloud round-trip (`~692 ms`)
- What if a tiny **CPU-only student**:
  - answers routine queries locally in `<20–50 ms`
  - measures its own uncertainty
  - queries `Gemini Decision 1` Teacher
    - *only when uncertain*
  - updates its decision heads:
    - online in `~17 ms` *incrementally*



MLSG
MACHINE LEARNING
SINGAPORE

`EmbeddingGemma 2` : An Instinctive Backbone



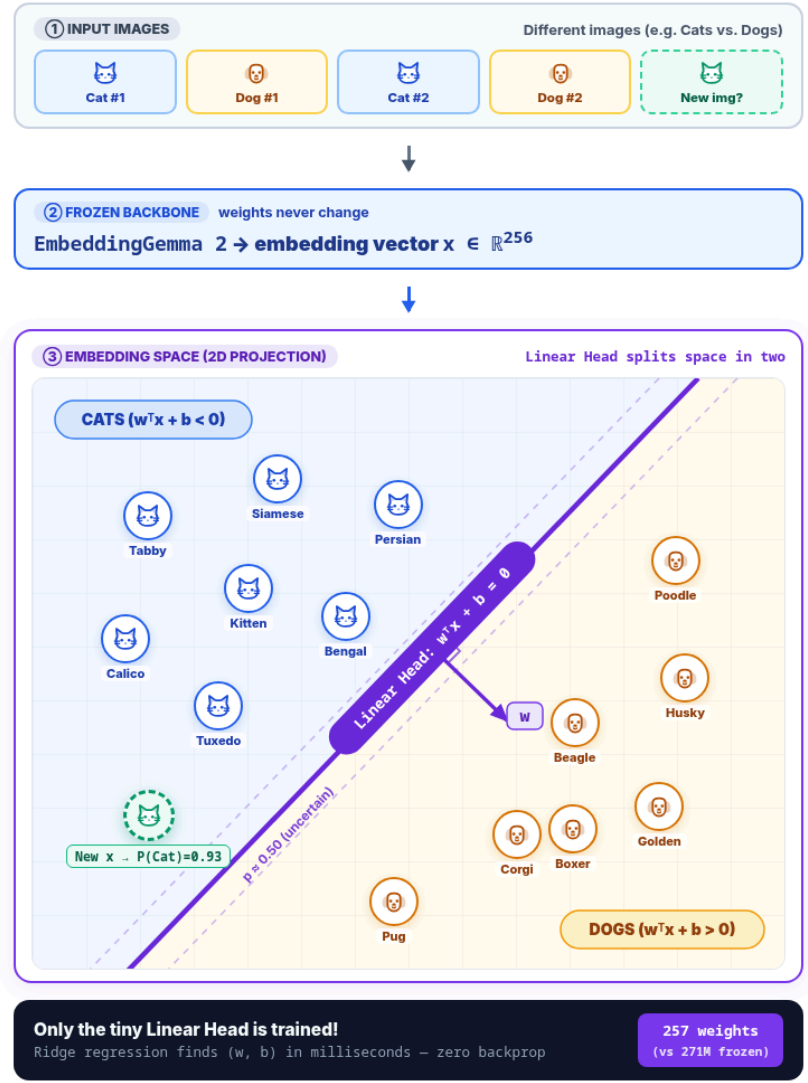


- EmbeddingGemma 2: Best-in-class open multimodal embeddings - Google
  - `271M` parameters for text-only version (`onnx-community/embeddinggemma-2-ONNX`)
    - `fp32` (`1,034 MB`)
    - `q8` (`300 MB` , `<0.7°` angle shift, $\cos\theta = 0.99993$)
    - `q4` (`166 MB` , `53.9 ms` on Skylake AVX2!)
  - Matryoshka Representation Learning (MRL) :
    - can truncate 768d embeddings to 256d, and it still (mostly) works!

MLSG
MACHINE LEARNING
SINGAPORE

Training an EmbeddingGemma2 head

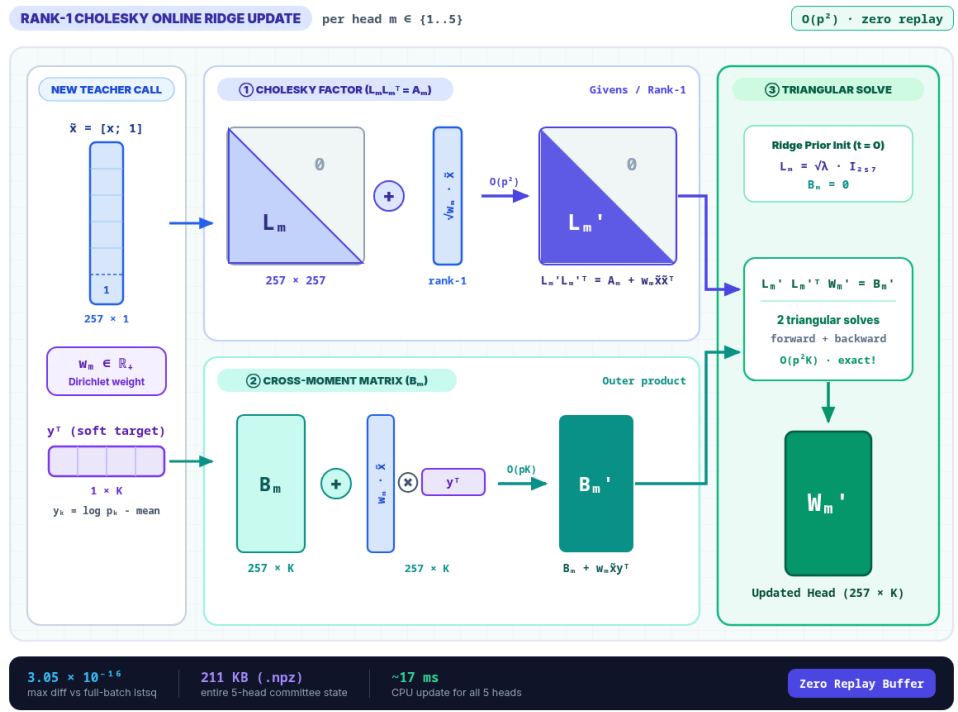
- We can embed different things
  - in the same embedding space
    - documents, images, audio ...
- Once we have embeddings for several classes
  - we can quickly find a decision boundary
    - that separates them *optimally*
- We can also train this 'head' dynamically
  - updating it as new data arrives
    - without a full retraining



MLSG
MACHINE LEARNING
SINGAPORE

Incremental Online Updates

- Why not update using SGD?
  - forgets earlier items without a replay buffer
  - sensitive to step size / learning rate
- Exact online Ridge : Rank-1 Cholesky
  - regresses onto Teacher's centered log-probs
    - preserves soft Teacher uncertainty!
  - updates sufficient stats (L_m, B_m) directly
    - matches full-batch `lstsq` (`3e-16` error)
    - `211 KB` `.npz` state, zero replay buffer
- Fast & quantization-resilient
  - $\mathcal{O}((d = 256 + 1)^2)$ = `~4 ms` per head
  - absorbs `q4` ONNX embedding rotations natively

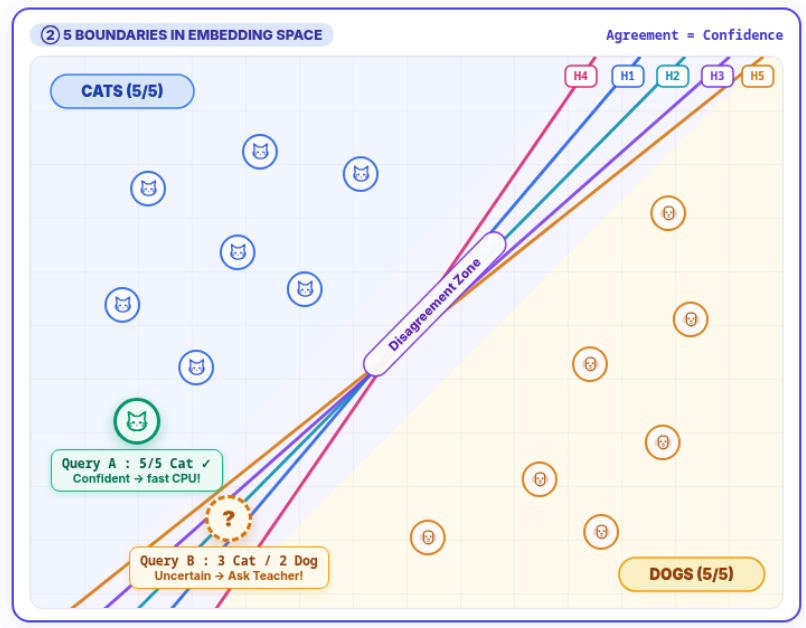


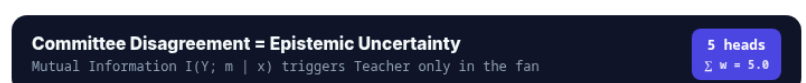
MLSG MACHINE LEARNING SINGAPORE

Training 5 different heads *on the same task*

- One head always gives an answer
  - ... even when it's just guessing
    - hard to know when to ask the Teacher
- Instead : train 5 different versions of the head
  - each new piece of training data
    - gets random weights :: $w \sim 5 \cdot \mathrm{Dirichlet}(1, 1, 1, 1, 1)$
  - smooth Bayesian bootstrap : always sums to `5.0`
    - zero wasted updates (every item teaches *someone*)
  - produces 5 slightly different decision boundaries
- Disagreement = *epistemic uncertainty*
  - all 5 agree ⟶ confident (answer locally on CPU)
  - heads split ⟶ uncertain (ask Teacher & update!)



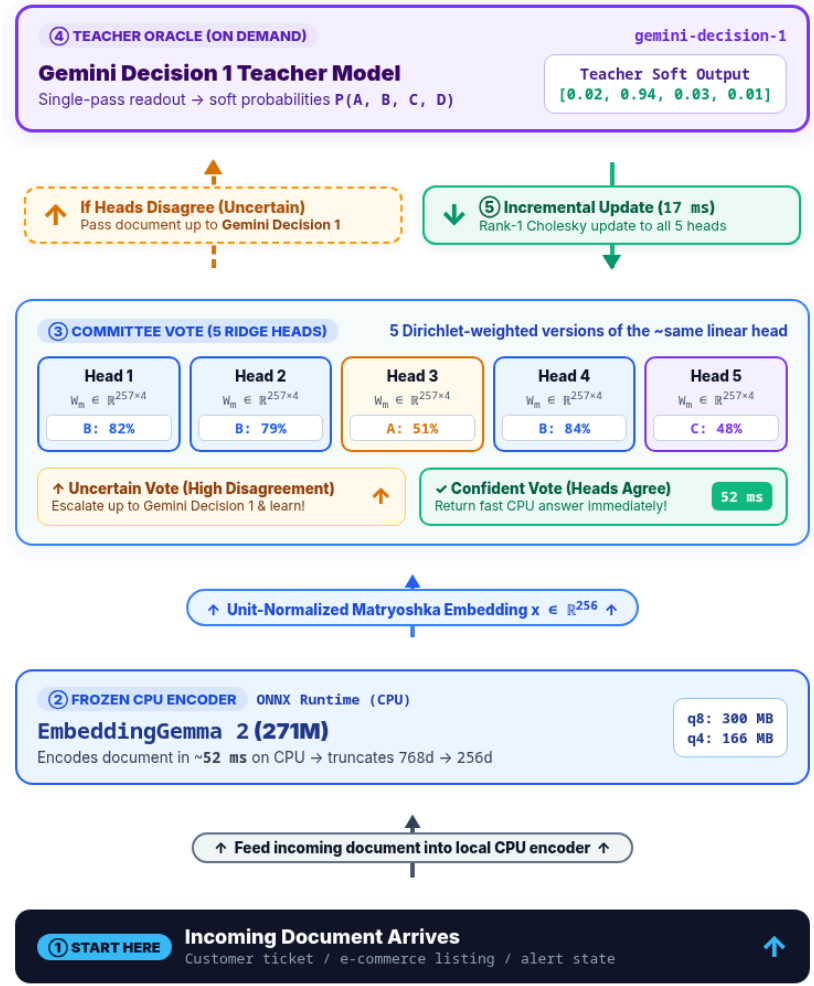




MLSG
MACHINE LEARNING
SINGAPORE

Incremental Online Ridge Committee

- Document → `EmbeddingGemma 2`
  - frozen CPU encoder outputs a `256d` vector x in ~52ms
- 5-Head Ridge Committee Votes
  - 5 Dirichlet-weighted versions of the ~same linear head
  - if heads agree (low Mutual Information $\mathcal{I}$)
    - return fast CPU decision immediately
  - if heads disagree (uncertain)
    - pass document up to the `Gev` Teacher
- Incremental Online Update (takes ~17ms)
  - aligns student with `Gev` 's soft probabilities
    - preserves teacher uncertainty!



MLSG
MACHINE LEARNING
SINGAPORE

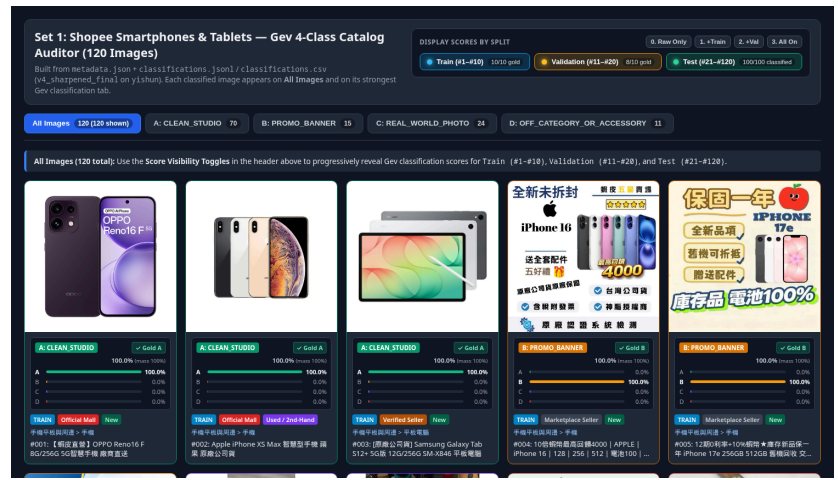
Demo Time

The E-Commerce Problem : Marketplace Catalog Audit

- Wild e-commerce categories are noisy:
  - Official studio renders mixed with
    - flashy discount banners
    - secondhand desk photos &
    - off-category cable spam

- One `DecisionSpec`:
  - `A` (`CLEAN_STUDIO`):
    - clean studio photo / official render
  - `B` (`PROMO_BANNER`):
    - marketing graphic overlaid with slogans, discount badges, star ratings or borders
  - `C` (`REAL_WORLD_PHOTO`):
    - real camera photo of a physical phone/tablet or box (held in hand, on desk)
  - `D` (`OFF_CATEGORY_OR_ACCESSORY`):
    - non-device listing (chargers, cables, cases, etc)

`shopee_phones_120` dataset



MLSG
MACHINE LEARNING
SINGAPORE

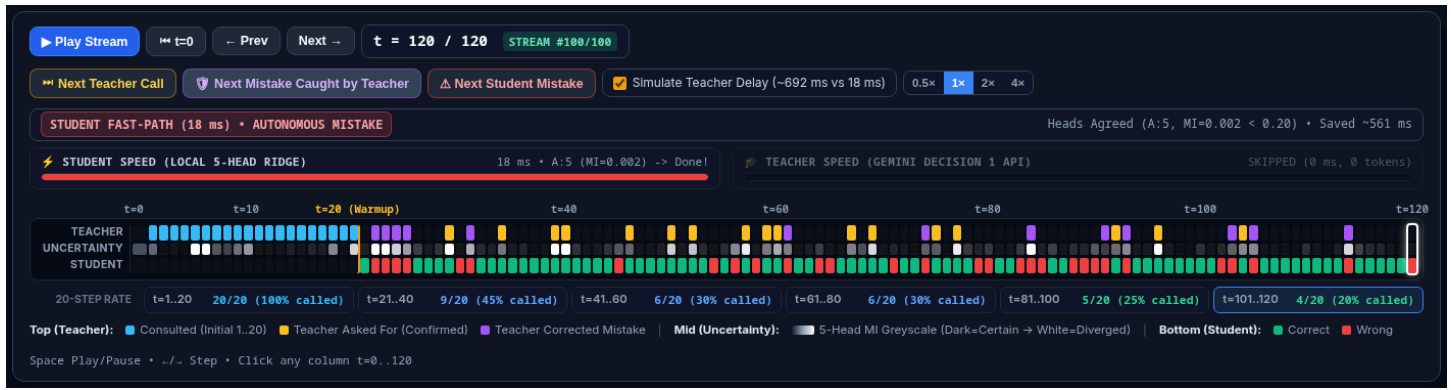
Demo Walkthrough (1/3) : One Arriving Item



- Step `t=22` (`#022` , iPad held in hand) — *Mistake Caught by Divergence!*
  - 5-head pre-vote leans `A` (`3A / 2C`), but `q_c` Mutual Information spikes (`0.312 nats`)
    - escalates to `Gemini Decision 1` (`712 ms`) → corrects to `C` (`REAL_WORLD_PHOTO`)
  - Dirichlet update (`∑w = 5.00`) flips all 5 heads to `C: 5/5 (99%)`

MLSG MACHINE LEARNING SINGAPORE

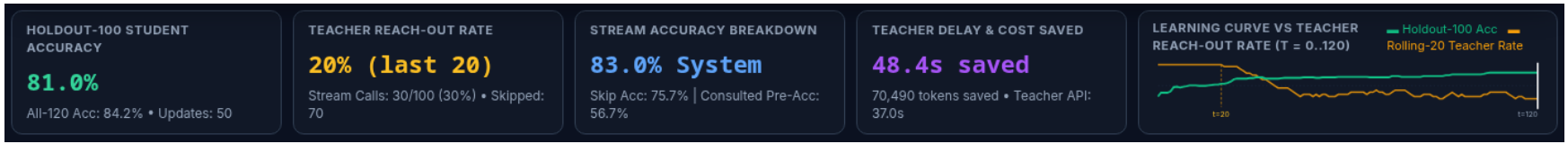
Demo Walkthrough (2/3) : The 120-Step Stream Timeline



- Three stacked tracks across `t = 1..120`:
  - Top (`TEACHER`) : blue warmup (`1..20`), then sparse gold (confirmed) & purple (mistake caught!)
  - Mid (`UNCERTAINTY`) : bright white MI spikes trigger Teacher escalations automatically
  - Bottom (`STUDENT`) : green fast-path hits (`18 ms` CPU) dominate as the committee settles
- 20-step Teacher call rate drops steadily:
  - `100%` (warmup) ⟶ `45%` ⟶ `30%` ⟶ `30%` ⟶ `25%` ⟶ `20%`

MLSG
MACHINE LEARNING
SINGAPORE

Demo Walkthrough (3/3) : Evolution of Learning



- Over 100 streaming items (`20` warmup + `30` selective Teacher calls):
  - Holdout-100 Student Accuracy climbs to `81.0%` (`84.2%` all-120)
  - Teacher Reach-Out Rate falls to `20%` in the final window (`70/100` autonomous CPU skips)
  - System Stream Accuracy hits `83.0%` while saving `48.4 s` & `70,490` Teacher tokens
- Green vs. Orange crossover curve (right):
  - Student accuracy rises as Teacher reliance falls — *without* any offline batch retraining!

MLSG
MACHINE LEARNING
SINGAPORE

Live Demo of Student / Teacher learning process

The Live Stream Dashboard

- **Items arriving in real time**
  - customer tickets or e-commerce listings streaming in
- **Instant Student readout (`<20 ms`)**
  - 5 committee heads vote on decision(s)
  - live metrics:
    - Confidence ($\max_k \bar{p}_k$) &
    - Mutual Information ($\mathcal{I}$)
- **Selective Teacher Retrieval**
  - when $\mathcal{I}(Y; m \mid x) > 0.06$ or $\max \bar{p} < 0.50$:
    - escalates to `Gemini Decision 1` Teacher
    - distills teacher's soft logit distribution incrementally

What happens over 20–100 items

- **Teacher call rate drops rapidly**
  - as the committee locks onto the domain boundaries
- **Average request latency plummets**
  - from `100s of ms` $\longrightarrow$ `<20 ms` on pure CPU
- **Increasing Student Decision accuracy**
- **Dynamic adaptation:**
  - if a new marketplace spam pattern appears, committee disagreement spikes
    - $\longrightarrow$ automatically queries `Gemini Decision 1`
    - $\longrightarrow$ adapts in 2–3 examples
      - without retraining on whole dataset

MLSG
MACHINE LEARNING
SINGAPORE

Demo Economics : Latency & Cost (`120` Images)

- `gemini-decision-1` **on all** `120` **images**
  - `5` questions/image = `600` **decisions**
- Latency : `692 ms` **/ image**
  - `83.0 s` sequential wall time
    - cloud round-trip dominates
- Tokens : `1,007` **input /** `0` **output**
  - `256` image + `751` prompt tokens
    - `120,861` total input tokens
- Total API Cost : `$0.012` *(1.2¢!)*
  - at `$0.10` / 1M input tokens
    - `~0.002¢` per decision

- Wait : if it's only **1.2¢**, why go local?
  - API cost is *already* negligible at small scale
- **Latency & locality** are the real wins:
  - local CPU : `<20 ms` vs `692 ms` (`>30×` faster)
  - runs on-device / offline / zero data egress
    - no rate limits or network jitter
- Hybrid Student ↔ Teacher sweet spot:
  - Teacher labels the tricky `~30%`
  - Student answers the routine `~70%` locally

MLSG
MACHINE LEARNING
SINGAPORE

Comparisons

(these make sense when combined with my DevFest talk...)

`gemini-decision-1` vs DIY 2B Teacher (`Gemma 4 E2B`)

- Higher accuracy on Gold set (`20` images):
  - `4× noul` path:
    - `gemini-decision-1` : `100%` (`20/20`)
    - DIY `Gemma 4 E2B` : `95%` (`19/20`)
  - `4-way choice` :
    - `gemini-decision-1` : `95%` (`19/20`)
    - DIY `Gemma 4 E2B` : `90%` (`18/20`)
- Big win : calibration out of the box
  - `noul` probs cleanly separated:
    - True : `~0.92–0.95`
    - False : `~0.05–0.08`
- No "secondary feature leakage"
  - 2B model over-fires on background clutter
    - charging cables / boxes in corner, promo badges
    - needed hand-tuned logit gates (`85%` → `95%`)
  - `gemini-decision-1` needs *zero* threshold hacks
    - naive boolean rule hits `100%` (`20/20`)
    - agrees with tuned gate on `119/120` (`99.2%`)
- Cleaner Teacher → faster Student
  - sharper soft labels = cleaner linear hyperplanes

MLSG
MACHINE LEARNING
SINGAPORE

Why the Student Learns Faster from `gemini-decision-1`

- **Cleaner Teacher labels = cleaner linear hyperplanes**
  - no false-positive `other` spikes on background accessories / badges
  - `EmbeddingGemma 2` ridge heads don't waste capacity chasing Teacher noise
- **Biggest per-question Student gains** *(vs DIY 2B Teacher)*:
  - `phone` (`q_a`): `90%` vs `76%` (`+14 pts`)
  - `other` (`q_d`): `86%` vs `76%` (`+10 pts`)
- **Bottom line (by step 120):**
  - reaches `84%` holdout accuracy (vs `76%` with 2B Teacher)
  - … while consulting the Teacher on only `30%` of the stream
    - `70%` answered locally on CPU in `<20 ms`

In the news...

Stock market action 1





MLSG
MACHINE LEARNING
SINGAPORE

Geopolitics : U.S. Licensing Rules for Singapore Datacenters

- **U.S. Commerce Dept interim rule (16-Nov-2026)**
  - adds **Singapore** (+ Malaysia, Thailand, Indonesia, Brazil) to export-restriction list
    - aimed at closing remote-access loopholes to U.S. AI hardware
- **What local data centers now face:**
  - **Chip Purchase Licenses** : U.S. license required to buy advanced AI GPUs (Nvidia, etc.)
  - **Remote Compute Cap** : license needed for China / Macau access `> 12M` aggregate TPP
    - TPP = Total Processing Performance — the U.S. Bureau of Industry and Security (BIS) export-control metric for AI chips
    - One Nvidia H100 ≈ 15,800 TPP → 12M aggregate TPP is roughly ~750 H100s (or ~300–400 Blackwell GPUs)
  - **Customer Screening** : vet clients for indirect ownership or shell companies
  - **90-day transition** — U.S. hyperscalers (`AWS` , `Azure` , `GCP`) exempt with reporting
- Emphasises my MLSG June-2026 segment :: **"Owning Your AI"**

MLSG
MACHINE LEARNING
SINGAPORE

Stock market action 2

- Anthropic (`ANT`) since IPO
  - pop to `$110+` → back to `~$83`



- Anthropic Model System Card
  - re: usage for frontier AI research

Anthropic on frontier AI research

9 June 2026 • Original safeguards

> "we've implemented new interventions that limit Claude's effectiveness for requests targeting frontier LLM development"
>
> "these safeguards will not be visible to the user."

Anthropic, *Claude Fable 5 & Claude Mythos 5* System Card, p. 13 (original version).
Silent safeguards were subsequently withdrawn on 11 June 2026.

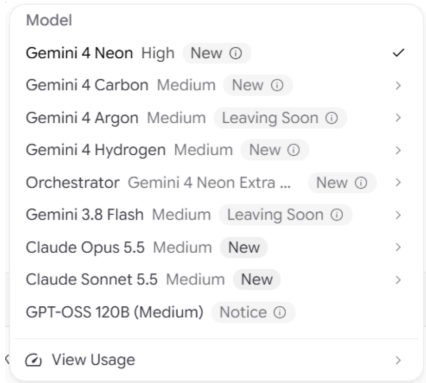
9 November 2026 • Training interventions

> "during the model's training, we took steps to limit advances in the specific research capabilities that could most directly accelerate AI development."

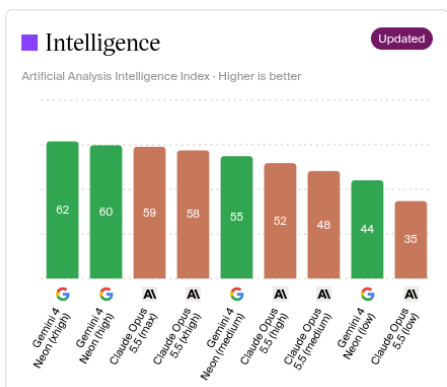
Anthropic, *Claude Fable 5.2 and Mythos 5.2*, 'AI research and development'.
https://www.anthropic.com/claude-fable-and-mythos-5-2
Anthropic says traditional AI/ML research was unaffected.

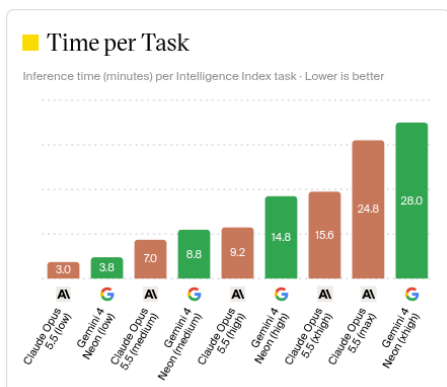
MLSG
MACHINE LEARNING
SINGAPORE

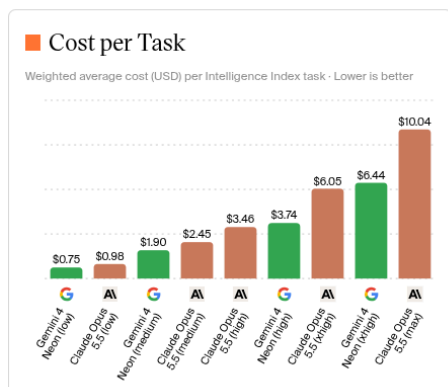
New Model Developments



- "Everything Google does is either **New** or **Leaving Soon**"
  - "There is no in between, no normal. Only **Notice**."
- Charts by Artificial Analysis



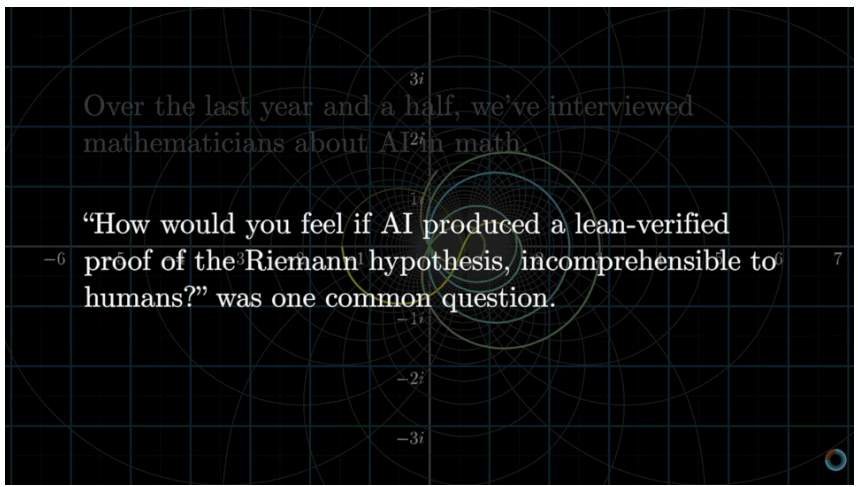




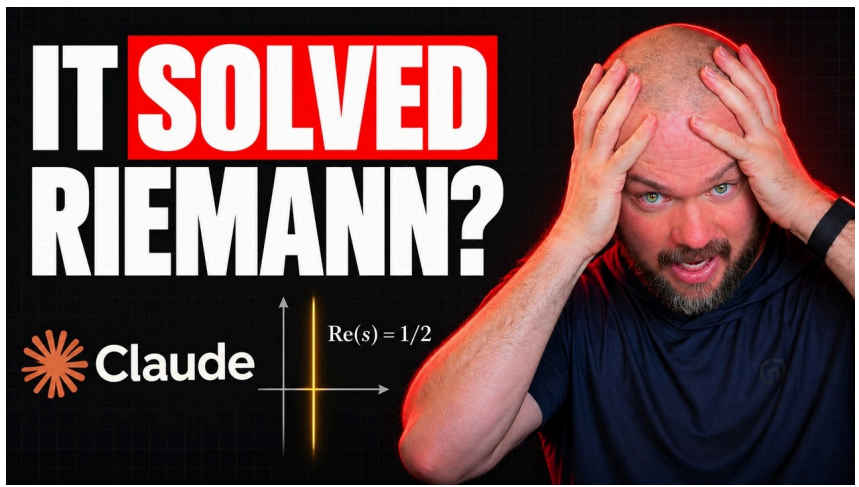
MLSG
MACHINE LEARNING
SINGAPORE

Riemann Hypothesis - (likely) Proved in Lean

- 3Blue1Brown : *AI in Math*
  - *"How would you feel if AI produced a Lean-verified proof..."*



- Wes Roth : *YouTube Thumbnail Energy*
  - meanwhile, on the less technical side of YouTube...



MLSG
MACHINE LEARNING
SINGAPORE

Wrap-Up

- "System 1" Decision Models are cool!
  - single-pass calibrated readouts
    - replace slow, brittle generative parsing
  - `Gemini Decision 1` :
    - clean, multimodal & `~1.2¢` for 600 decisions
- Local CPU Students can learn *live*
  - `EmbeddingGemma 2` + 5 Dirichlet-weighted ridge heads
    - `<20 ms` on CPU
  - committee disagreement = *know when to ask the Teacher*
    - exact incremental updates in `~17 ms`
- Best of both worlds : fast & adaptive
  - and you can build this too!
    - Demo link & Colab Link

MLSG
MACHINE LEARNING
SINGAPORE



Link to Slides

- https://bit.ly/MLSG_2026-11



MLSG
MACHINE LEARNING
SINGAPORE