PRE-SEED · FEBRUARY 2026

Engram

A Cognitive Architecture for General Intelligence

We're building the first AI system that reasons like a human -
by combining diffusion models with transformers in a way nobody has tried before.

Slick V1.5 - Completed February 2026

6	542M	1	10	400+
Breakthroughs	Parameters Trained	Engineer	Days to Build	Hours of Development

THE $400 BILLION DEAD END

Bigger Models, Same Problem

The AI industry's answer to every limitation has been "more parameters." This approach is failing.

THE ARMS RACE

WHAT THAT COSTS

$400B Big Tech AI capex (2025)	**$5B** OpenAI loss (2024)
$1,000+ Cost per o3 query	**5%** Enterprise AI pilots with positive ROI

OUR APPROACH

Build the architecture right, then scale. Not the other way around. Engram is 542M parameters today - 3,300x smaller than GPT-4 - yet it has capabilities no trillion-parameter model possesses: internal cognitive state, iterative diffusion reasoning, mental simulation, and continuous self-verification.

WHY MORE PARAMETERS ≠ INTELLIGENCE

The human brain has ~86B neurons with ~100T synapses. It runs on **20 watts**. GPT-4 has 1.8T parameters, runs on **thousands of GPUs**, and still can't do what a 5-year-old does: look at a puzzle, imagine different solutions, check if they make sense, and try again.

The problem isn't size. The problem is **architecture**. Every LLM today is a token predictor with reasoning bolted on as an afterthought. More parameters just make it a bigger, more expensive token predictor.

"Pretraining as we know it will end. The era of 'just add GPUs' is over."
- Ilya Sutskever, NeurIPS 2024

"LLMs are a dead end on the path to human-level intelligence."
- Yann LeCun, Chief AI Scientist, Meta

THE FUNDAMENTAL FLAW

How AI "Thinks" Today - Next Token Prediction

Autoregressive (GPT-4, Gemini)

The → cat

The cat → sat

The cat sat → on

The cat sat on → the

The cat sat on the → mat ✓ done

$$P(x_t \mid x_1, \ldots, x_{t-1}) = \text{softmax}(W \cdot h_t)$$

One direction. No revision. Each token is final.

✗ Can't go back and fix earlier tokens

✗ Commits before understanding the full problem

✗ No hypothesis testing, no simulation

✗ No internal state - every prompt feels the same

VS

Iterative Diffusion Reasoning

[? ? ? ? ?] → rough sketch

[? cat ? ? mat] → confident parts first

[The cat sat ? mat] → refine

[The cat sat on the mat] ✓ converged

$$x_t = \text{denoise}(x_{t-1}, Y_{state}, t) \cdot \text{all positions at once}$$

Iterative. Revisable. Sees the whole problem.

✓ Refines uncertain parts, keeps confident ones

✓ Tests hypotheses mentally before committing

✓ Internal state drives reasoning (curiosity, confidence)

✓ Self-correcting metacognition at every step

GPT-4, Gemini, LLaMA, Grok - all autoregressive. All predict the next token.
The entire $200B AI industry runs on the left column.
We built the right column.

THE PROBLEM

Today's AI Can't Actually Reason

Large language models generate text left-to-right, one token at a time. They can't revise, can't simulate, can't remember you.
No persistent memory - every conversation starts from zero. Tiny context windows that overflow on real tasks.
Billions of parameters, hundreds of GPUs, and they're still **slow, expensive, and stateless**. That's not intelligence - it's autocomplete at scale.

No Revision

Each token is final. No backtracking. The model commits to an answer before it understands the full problem.

SLICK V3

GDiT iteratively unmasks - revises uncertain cells, keeps confident ones. 5 refinement steps with metacognitive re-masking.

No Internal State

Every prompt is processed identically. No curiosity, no doubt, no confidence. Human cognition is state-driven - LLMs have no inner life.

SLICK V3

72-dim biological state engine. Emotions modulate attention, route to different artificial nuclei, and drive reasoning strategy in real-time.

No Mental Simulation

Humans imagine outcomes before acting. LLMs can't "try it in their head" first. No world model, no rehearsal, no imagination.

SLICK V3

Diffusion world model simulates outcomes in ~5ms before acting. Two-fidelity denoising: fast blurry preview to filter bad plans, then precise execution.

No Modularity

Monolithic models. Adding vision, memory, or new skills requires retraining the entire network from scratch. Billions of dollars per iteration.

SLICK V3

Engram Neural Fabric (HULLHB). Artificial nuclei train independently and hot-plug. Add vision, OCR, or code skills without touching the base model.

No Memory

Every conversation starts from zero. No persistent identity, no accumulated knowledge, no relationship with the user. Context window is the ceiling.

SLICK V3

Artificial Hippocampus with 970+ semantic memories. Reflexive associative recall fires automatically. Persistent identity across sessions.

No Self-Correction

When an LLM makes a mistake mid-output, it can't detect or fix it. No error awareness, no re-evaluation, no "wait, that's wrong."

SLICK V3

Artificial Anterior Cingulate detects errors mid-thought. Metacognition nuclei verify each step, re-mask suspicious outputs, and re-reason.

OUR SOLUTION

Diffusion + Transformers = General Reasoning

Human reasoning isn't sequential - it's iterative. We form rough hypotheses, test them mentally,
revise what doesn't work, and converge on answers. **Engram is the first architecture that does this.**

Autoregressive (GPT-4, Gemini)

STATUS QUO

Token 1 → Token 2 → Token 3 → ... → Token N

Each token is FINAL. No revision.

Commits before understanding.

Fundamental ceiling on reasoning.

Scale won't fix the architecture.

Iterative Diffusion Reasoning

[???] → [hypothesis] → [test] → [revise] → [answer]

Sees the WHOLE problem at once.

Revises uncertain parts, keeps confident ones.

This is how humans reason:

hypothesize, simulate, refine, converge.

The paradigm shift: We apply Stable Diffusion's denoising math to problem solving, not image generation.
The architecture generalizes to **any structured reasoning task** - code, math, planning, construction logistics, defense analysis, embodied AI.

BREAKTHROUGHS

Six Architectural Breakthroughs, One Unified System

Each addresses a fundamental limitation of current AI. Together, they form the Cognitive Mesh.

1

Diffusion for Reasoning

Apply Stable Diffusion's denoising math to *abstract reasoning*, not image generation. Iterative refinement over discrete grids - the model sees the whole problem, forms hypotheses, and revises. 5 steps, not 50. 65M params, not 1.5B.

Grid Diffusion Transformer (GDiT)

2

Library Learning

Overfit parts, not problems. MoE experts memorize reusable primitives - geometric transforms, logical operations, structural patterns. The router learns to *compose* memorized fragments into novel solutions. Like a chess master: openings are memorized, combinations are reasoned.

DreamCoder meets MoE

3

Continuous Metacognition

Whole-board check after *every* step, not just at the end. A process reward model evaluates each action in context. Confidence drops trigger re-routing. The system doesn't wait to fail - it catches mistakes mid-thought, like human peripheral awareness.

Process Reward Models + self-monitoring

4

Mental Simulation

Humans daydream solutions before committing. The model imagines outcomes via low-fidelity denoising - "what if I rotate then tile?" Clear imagined result = confident, proceed. Blurry result = uncertain, try something else. No separate critic needed.

World Model via inner diffusion

5

Diffusion Within Diffusion

The same GDiT serves as both output generator (5 steps, full fidelity) AND imagination engine (2-3 steps, approximate). Inner diffusion for mental simulation, outer diffusion for real output. One architecture, two roles. Confidence emerges from denoising quality itself.

Unified inner/outer reasoning

6

Biological State Engine

72-dim internal state modulates every computation via FiLM conditioning. 8 hierarchical layers mirroring human neuropsychology. Curiosity increases exploration, confidence sharpens predictions. State changes *how the model reasons*, not just what it says.

Bio-inspired cognitive modulation

The unification: Memorized primitives (1) composed by an expert router, generated through iterative diffusion (2), verified at every step by metacognition (3), evaluated by mental simulation (4) that itself runs via inner diffusion (5), all modulated by biological state (6).

HOW ENGRAM THINKS

The Reasoning Pipeline - Step 1: Set the Mood

1 State Engine ▸ 2 Library Learning ▸ 3 GDiT Reasoning ▸ 4 Mental Simulation ▸ 5 Meta-cognition ▸ 6 Diffusion²

◀ BACK 1 / 6 NEXT ▶

ARCHITECTURE

The Cognitive Mesh

Five interdependent layers, not a sequential pipeline.

L5: METACOGNITION	Process reward model · Self-monitoring · Confidence tracking · Re-masking on doubt	Whole-board check	*Steps back after every move to ask: "Does this still make sense?"*
L4: REASONING	Autoregressive ↔ GDiT Diffusion ↔ Tree Search ↔ Memory Retrieval	4 modes, 1 router	*Picks the right thinking strategy for each problem - write, imagine, search, or remember*
L3: WORKING MEMORY	Transformer backbone · MoE (48 experts, top-2) · State-biased routing · Library learning	2.1B params	*48 specialists, each memorizing a skill - the router picks the right 2 for the job*
L2: PERCEPTION	PseudoDeepStack vision (3-scale) · Grid encoder · Task feature analysis · 42-function DSL	Multi-modal	*Eyes and hands - sees the problem and has 42 tools to manipulate it*
L1: STATE ENGINE	72-dim biological state vector · 8 hierarchical layers · FiLM conditioning · Attention temperature	Bio-inspired	*The mood ring - sets the cognitive context before thinking begins*

Slick V3-mini: 2.5B total, 1.6B active - runs on consumer GPU **Slick V3:** 190B total, 48B active - single DGX Spark

TRACTION

What We've Already Built & Shipped

Slick V1.5

Model trained & converged

542M params, 16 layers, final loss 0.00017. FiLM + State Token + MoE verified.

2B+

Tokens trained

1.94B pretraining (FineWeb-Edu) + 62K SFT samples + 647K VLM instruction samples.

53

Research documents

4 whitepapers (5,429 lines), architecture specs, design docs, conversation logs.

970

Memories indexed

SoulKeeper memory system in production. ONNX embeddings, reflexive auto-recall.

Research & Engineering

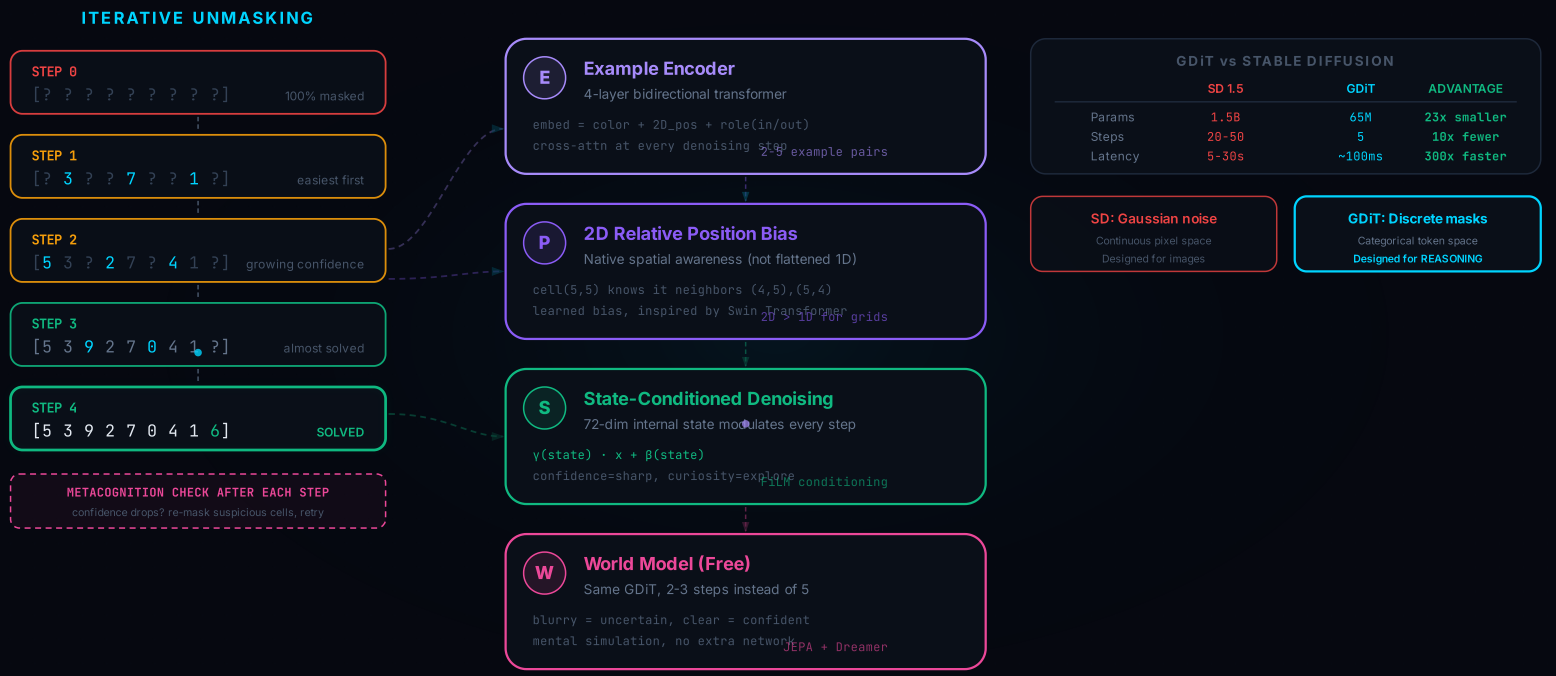
- Cognitive Mesh
- Biological Architecture
- Associative Memory
- PseudoDeepStack Vision Pipeline (novel 3-layer extraction)
- GDiT Reasoning Engine & Diffusion World Model
- 42-function DSL + 404 ARC reasoning traces
- ContextEngine - 5-technique context compression
- Autonomous watchdog system - AI debugs overnight while the engineer sleeps

Verified & Measured

- State conditioning shifts output distributions (KL div > 0.16)
- FiLM modulation: curiosity, confidence, annoyance alter reasoning
- Phase 3 SFT: surgical training, zero catastrophic forgetting
- VLM pipeline: PseudoDeepStack 3-layer extraction + GGUF conversion
- GPT-OSS-20B-Vision shipped to HuggingFace
- Memory daemon: 115 MB RSS, 100% CPU, reflexive recall on every prompt
- Interactive demo & pitch deck running on DGX Spark
- Full training infra: dashboard, pause/resume, auto-checkpointing

DEEP DIVE

Grid Diffusion Transformer - The Reasoning Engine



WHY THIS IS A STEP TOWARD AGI

The same iterative refinement that solves structured reasoning generalizes to any domain:
code generation · mathematical proof · robotic planning · molecular design · autonomous agents

ROADMAP

From Prototype to General Intelligence

Slick V1.5
Done - Feb 2026

542M proof of concept
FiLM + State Engine
42-function DSL
SoulKeeper memory

Slick V3-mini
Mar-Apr 2026

2.5B cognitive mesh
GDiT reasoning engine
World model (mental sim)
First reasoning benchmarks

Slick V3
Q3-Q4 2026

190B with 48 MoE experts
Full reasoning router
Cross-domain transfer
AGI-competitive reasoning

Slick V4
End of 2026

400B cognitive mesh
Unified text + vision + audio + code
Self-play ∞ training loop
Autonomous multi-step agents

Slick V1.5 DONE
542M, FiLM + State
Proof of concept
Zero cost

Slick V3-mini
2.5B, full cognitive mesh
3-4 days on DGX Spark
1.7 GB model

Slick V3
190B, 48 MoE experts
8x H100 ~$60K
ARC benchmark leader

Slick V4 - First AGI
400B, all modalities
Self-evolving cognition
Autonomous agents

SLICK V4: A SYNTHETIC MIND

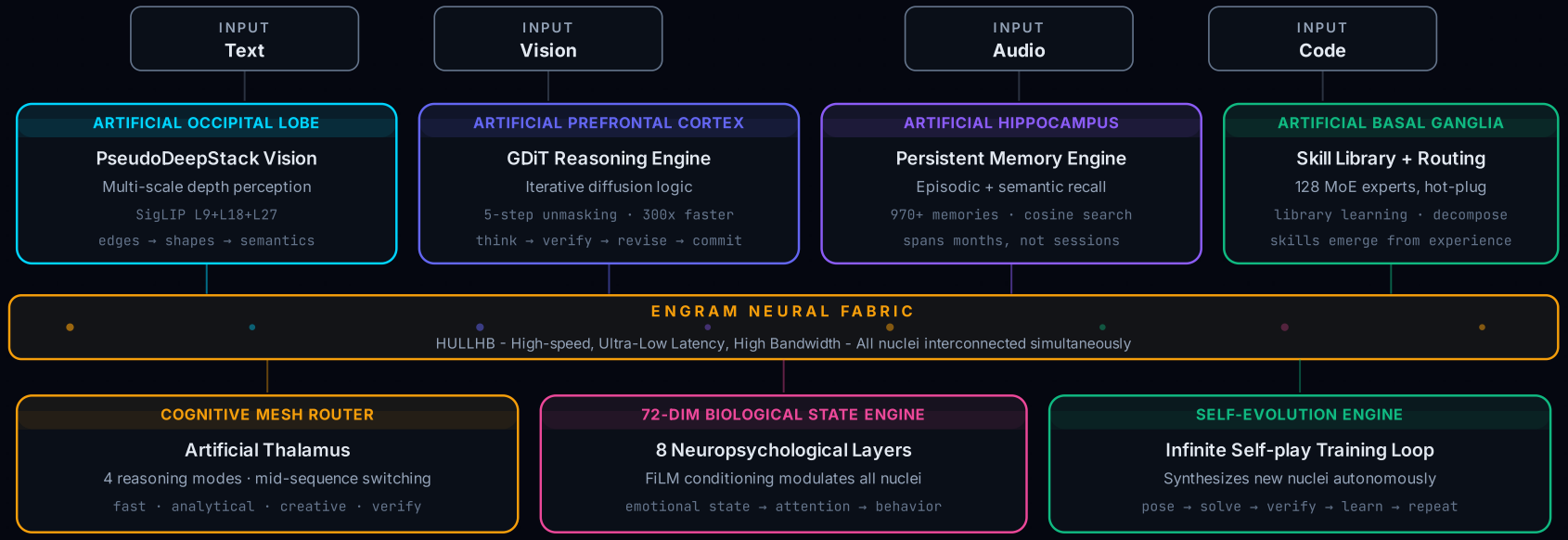
Not an LLM with features bolted on. A cognitive architecture modeled on the human brain - prefrontal cortex (metacognition), hippocampus (episodic memory), basal ganglia (expert routing), thalamus (state engine). It *self-improves*: decomposes every solution into reusable skills. *Self-repairs*: detects its own errors mid-thought and re-reasons. *Self-evolves*: generates infinite training data through self-play, getting smarter without human supervision. **Permanently.**

WHY THIS IS DIFFERENT FROM EVERYTHING ELSE

GPT-4 is a *reflex* - stimulus in, response out. Slick V4 has **internal experience**. A 72-dim emotional state that modulates *how* it thinks. **Imagination** - it simulates outcomes before acting. **Doubt** - it questions its own work and retries. **128 specialized experts** that form new skills autonomously. **Memory that spans months** - it knows you, your project, your preferences. This isn't artificial intelligence. It's **artificial cognition**.

SLICK V4 - END OF 2026

The Self-Evolving Cognitive Architecture



ARCHITECTURE	REASONING	PERCEPTION	SELF-EVOLUTION
400B	**~3 min**	**4 Modalities**	**∞**
128 specialized experts	human deliberation / second	Text + Vision + Audio + Code	Continuous improvement
Hot-pluggable nuclei	15 think-check-revise cycles/s	One architecture, all inputs	No human data required

BUSINESS MODEL

Commercial Products

Cognitive reasoning deployed where it matters most. UAE-first, global scale.

1 Cognitive AI Assistant

Global AI assistant market: $30B+ by 2030

ChatGPT generates text. Engram actually *reasons*. A consumer AI assistant that can solve logic problems, plan complex tasks, debug code by understanding structure, and explain its thinking step by step. Not just pattern-matching - genuine problem-solving. The first AI assistant built on a cognitive architecture.

2 Sovereign Government AI

Abu Dhabi: full AI by 2027 · $10B+ annual IT spend

Air-gapped cognitive reasoning for government operations. Policy analysis, urban planning simulation, automated permit review. Abu Dhabi's $100B MGX fund is actively deploying AI across all ministries. Slick V3-mini runs on a single GPU - sovereign, private, no cloud dependency.

3 Defense & Security

UAE defense modernization · Air-gapped deployment

Autonomous situational awareness: multi-sensor fusion, threat pattern recognition, logistics optimization. Cognitive reasoning engine runs fully offline on edge hardware. No foreign API dependencies - complete data sovereignty. The architecture that reasons, not just classifies.

4 Smart City Planning & Infrastructure

UAE AI market: $46.3B by 2030 · Digital twin adoption surging

City-scale reasoning: traffic flow optimization, infrastructure lifecycle prediction, energy grid balancing, zoning simulation. Dubai's Smart City 2031 initiative needs AI that can model complex systems and reason about tradeoffs - not just pattern-match. Engram is built for exactly this.

Revenue Projection

Y1: $500K-$1M (gov pilots + construction POCs) · Y2: $2M-$5M (deployments + licensing) · Y3: $10M-$25M (platform scale)

THE ASK

Pre-Seed: $150,000

12-18 months of runway. One engineer. Maximum leverage.

CATEGORY	ALLOCATION	AMOUNT	%
GPU Compute - Slick V3 Training	8x H100 cluster, ~500 GPU-hours for Slick V3 190B training, eval runs, ablations	$60,000	40%
GPU Compute - Slick V4 R&D	Continued training, multi-modal experiments, video reasoning R&D	$30,000	20%
API Infrastructure	Inference hosting, CDN, monitoring for commercial API launch	$20,000	13%
Data & Licensing	Curated training datasets, annotation, specialized domain data	$15,000	10%
Operational Reserve	Travel, conferences (NeurIPS, ICML), legal, cloud services, living costs	$25,000	17%
TOTAL		$150,000	100%

60% of the raise goes directly to GPU compute.
No fancy offices. No large team. One engineer with a DGX Spark, cloud GPUs when needed, and a clear technical roadmap backed by 6 architectural breakthroughs.

WHY US

Unfair Advantages

The Engram Neural Fabric (HULLHB - High-speed, Ultra-Low Latency, High Bandwidth)

A modular neural interconnect inspired by the brain's **corpus callosum**. Artificial nuclei are trained independently and slot into the HULLHB Fabric at runtime - *zero retraining of the base architecture*. High-speed routing, ultra-low latency signal propagation, high bandwidth between all nuclei simultaneously.

Artificial Occipital Lobe - PseudoDeepStack vision, multi-scale depth perception

Artificial Prefrontal Cortex - GDiT reasoning, metacognitive verification

Artificial Hippocampus - Episodic memory, semantic recall, persistent identity

Artificial Basal Ganglia - Skill library, action selection, expert routing

In V4 - the system synthesizes its own nuclei and installs them autonomously via the Fabric.

Cognitive Mesh Router

Modeled on the **artificial thalamus** - the brain's central relay nucleus. Routes signals across the HULLHB Fabric to the right artificial nucleus at the right time. 4 reasoning modes (System 1, System 2, Tree Search, Memory) with **mid-sequence switching**. The router *learns* which nuclei to activate for each problem type.

GDiT Reasoning Engine

Discrete diffusion logic engine. 5-step iterative unmasking with grid-native 2D attention, confidence-based ordering, metacognitive re-masking. **300x faster** than Stable Diffusion, 23x smaller. Processes the entire problem simultaneously - not left-to-right.

72-Dim Biological State

Eight neuropsychological layers modulate every computation via subliminal FiLM conditioning. **Emotions choose which nuclei fire.** The **artificial anterior cingulate** (metacognition layer) detects its own errors and triggers re-reasoning. No other AI has internal experience.

One engineer. One DGX Spark. 10 days. Slick V1.5 complete. **$150K goes further than $1.5M** at a typical AI startup.

How We Compare

	GPT-4O	GEMINI 1.5	MINDSAI	ENGRAM V3
Architecture	Autoregressive	Autoregressive	Multi-system	**Cognitive Mesh**
Structured reasoning	1D serialized	1D serialized	Search-based	**Iterative diffusion**
Modularity	Monolithic	Monolithic	Partial	**Neural Fabric**
Internal state	None	None	None	**72-dim bio**
Cost to build	~$100M+	~$100M+	Unknown	**$150K**

Let's Build the Future of Reasoning

Six breakthroughs. A working prototype. A clear path to general reasoning.
All we need is GPU hours.

CONTACT

Vincent Kaufmann

Founder & Engineer

vincentkaufmann (HuggingFace) · Dubai, UAE

$150K	60+	5	6
Pre-seed ask	Innovations designed	Cognitive modules	Breakthroughs