NimbleEdge

Real-time Reasoning on Edge

NeurIPS 2025

Company Overview

Team

Varun Khare
Co-founder & CEO

AI researcher at:
Berkeley UNIVERSITY OF CALIFORNIA OpenMined

Neeraj Poddar
Co-Founder & CTO

Head of Eng at:
solo.io Istio

Co-founded:
ASPEN MESH

Partners

Microsoft Meta

Backed By

Dawn Song
Professor
Berkeley UNIVERSITY OF CALIFORNIA

Srinivas Narayanan
VP of Engineering
OpenAI

Vibhu Mittal
CTO
Inflection

Swaroop Kolluri
Founder and MD
NEOTRIBE

Event Driven Reasoning at Scale

30M+

Smartphones

10B+

On-device inference calls

15M+

Peak concurrency

3000

Events/min for context engineering

>2500 loc

Python on-device for stateful context

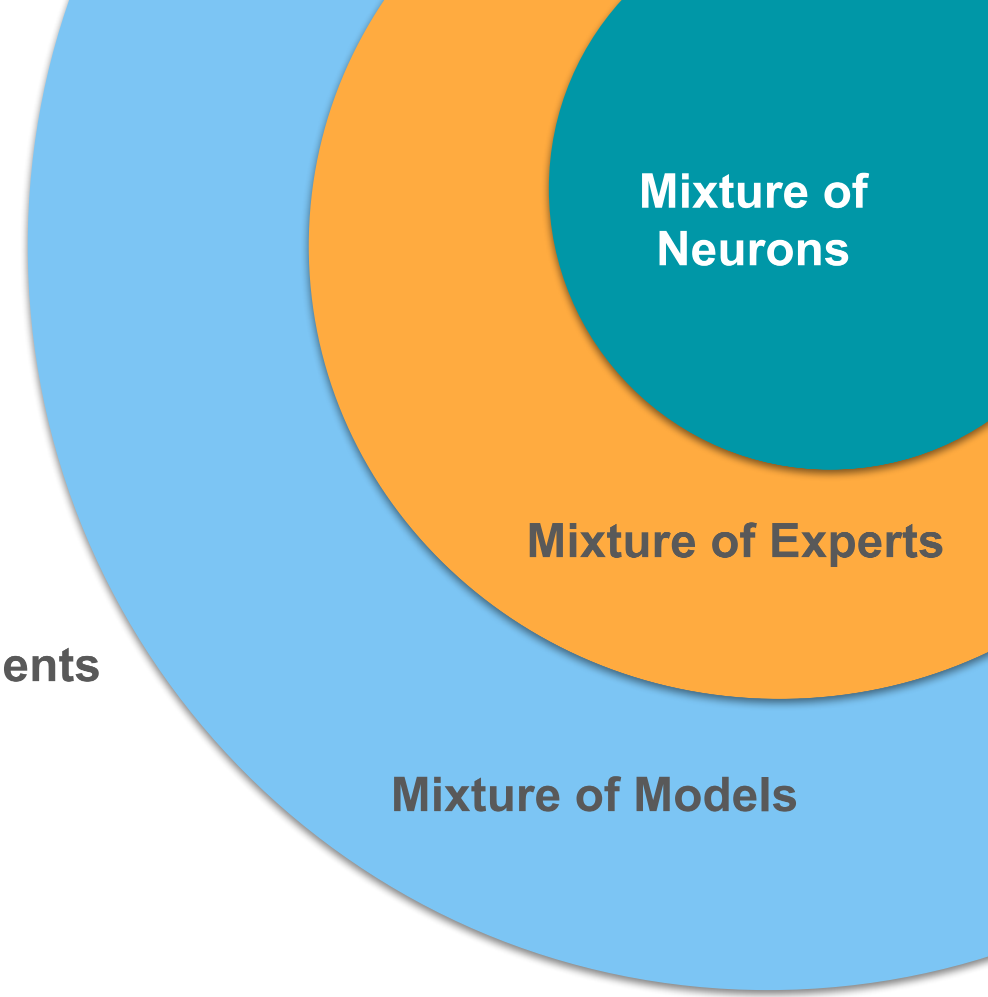
0.5M -> 3M

New concurrent users within a min

Modern UI is built on Event Streams

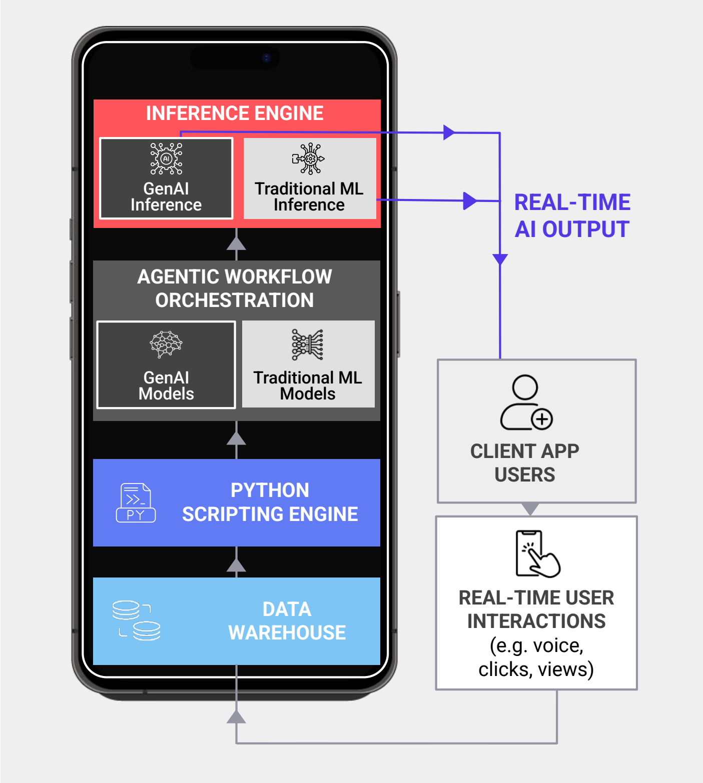
- UI interactions generates continuous events
  - **>3000 user events** generated per min in a typical app
- Real-time Personalization gives ~10% uplift when streaming vs batching
  - Average session **duration < 2 mins**
- Cloud infra spends of **~$30M/year** for batching event processing
  - Delivers subpar results at exorbitant costs!

Large efficiency gains
With smart routing in streams



Mixture of Agents

DeliteAI- Open Source On-Device AI Platform - Demo



INFERENCE ENGINE

Optimized on-device GenAI execution engine with **2x faster performance**

AGENTIC WORKFLOW ORCHESTRATION

Python based prompt chaining/ workflow orchestration enabling lower time to market

PYTHON SCRIPTING ENGINE

On-device Python engine with C++ backend for **real-time event processing** to unlock in-session reasoning

DATA WAREHOUSE

In-memory and persistent **database for events, feature stores and RAG**

Asynchronous In-session Context Engineering

Event Stream Capture

- UI diffs, analytics, clickstreams
- Intent extraction and Trajectory prediction

Stateful Context Engineering

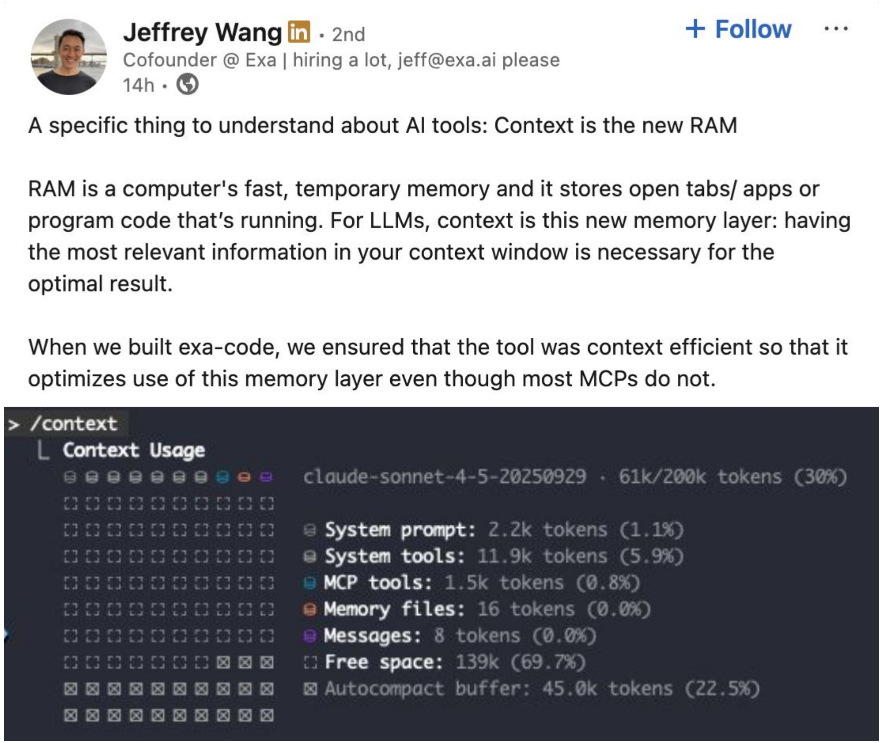
- Offload & compact
- Summarize when needed
- Isolate contexts across agents

KV cache and Memory Prefill as Bottleneck

Limited RAM for Multi-agents

- Message list as base input format ineffective

- Multiplexing requires hot reload



Jeffrey Wang in · 2nd
Cofounder @ Exa | hiring a lot, jeff@exa.ai please
14h ·

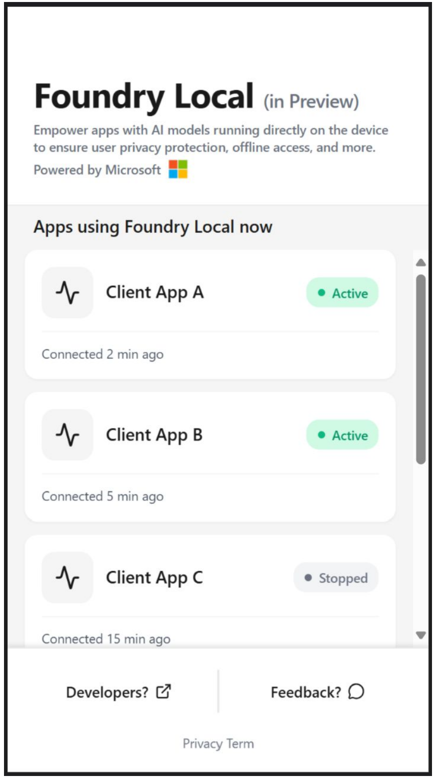
A specific thing to understand about AI tools: Context is the new RAM

RAM is a computer's fast, temporary memory and it stores open tabs/ apps or program code that’s running. For LLMs, context is this new memory layer: having the most relevant information in your context window is necessary for the optimal result.

When we built exa-code, we ensured that the tool was context efficient so that it optimizes use of this memory layer even though most MCPs do not.

```
> /context
  ⎿ Context Usage
    ⛁ ⛁ ⛁ ⛁ ⛁ ⛁ ⛁ ⛁ ⛁ ⛁   claude-sonnet-4-5-20250929 · 61k/200k tokens (30%)
    ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶
    ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶   ⛁ System prompt: 2.2k tokens (1.1%)
    ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶   ⛁ System tools: 11.9k tokens (5.9%)
    ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶   ⛁ MCP tools: 1.5k tokens (0.8%)
    ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶   ⛁ Memory files: 16 tokens (0.0%)
    ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶   ⛁ Messages: 8 tokens (0.0%)
    ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛶ ⛝ ⛝ ⛝   ⛶ Free space: 139k (69.7%)
    ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝   ⛝ Autocompact buffer: 45.0k tokens (22.5%)
    ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝ ⛝
```

Foundry Local for Android with DeliteAI & Microsoft- Demo



Foundry Local (in Preview)

Empower apps with AI models running directly on the device to ensure user privacy protection, offline access, and more.

Powered by Microsoft

Apps using Foundry Local now

Client App A ● Active
Connected 2 min ago

Client App B ● Active
Connected 5 min ago

Client App C ● Stopped
Connected 15 min ago

Developers?

Feedback?

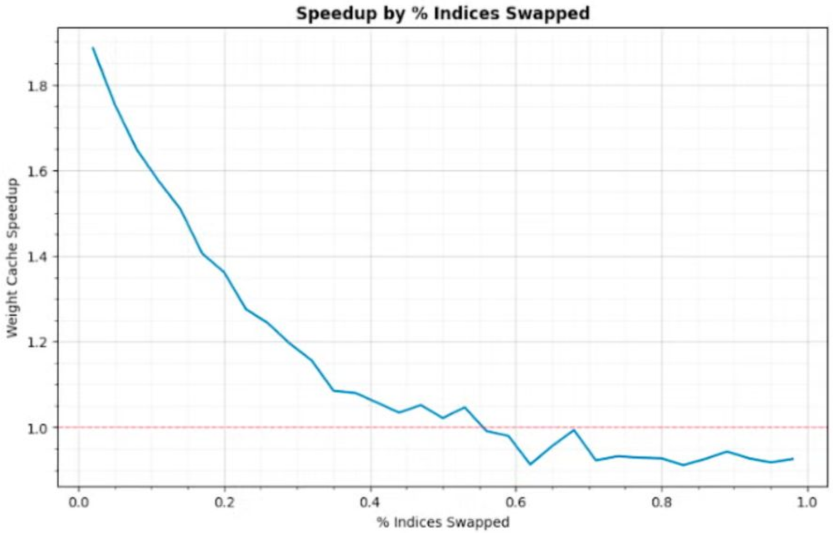
Privacy Term

On-Device AI Service multiplexing requests across apps

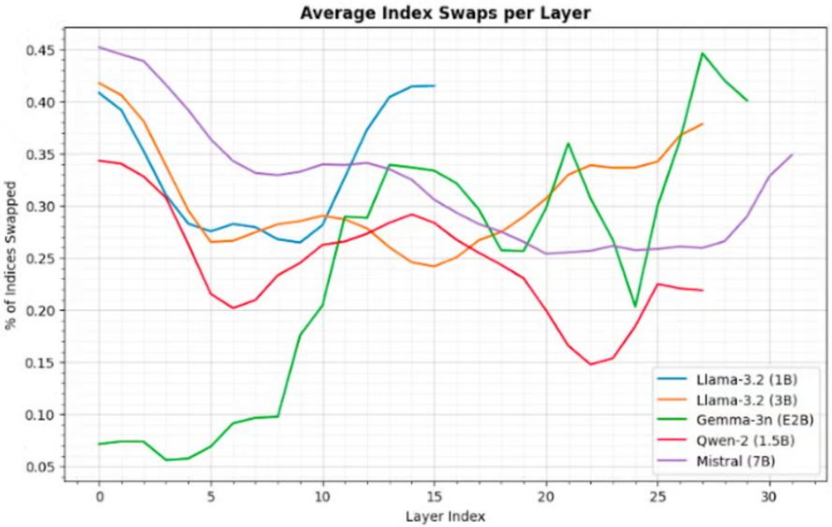
Pre-shipped SOTA LLMs custom built for app workflows

Azure AI Foundry compatible for hybrid inference

Optimizing LLM Inference for CPU and NPU

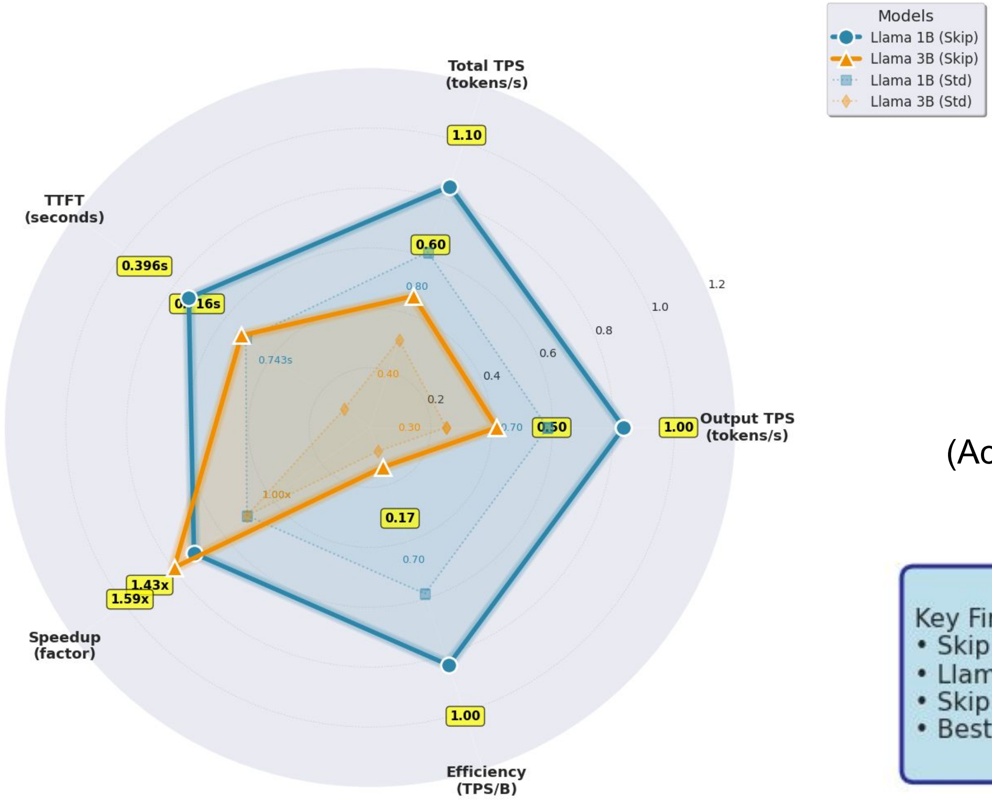


Overall speedup in MLP block operations as a function of index swaps per cache update.



Average index swaps for common models using contextual sparsity

Faster Inference w/Sparse Transformers ~2x faster 30% less memory



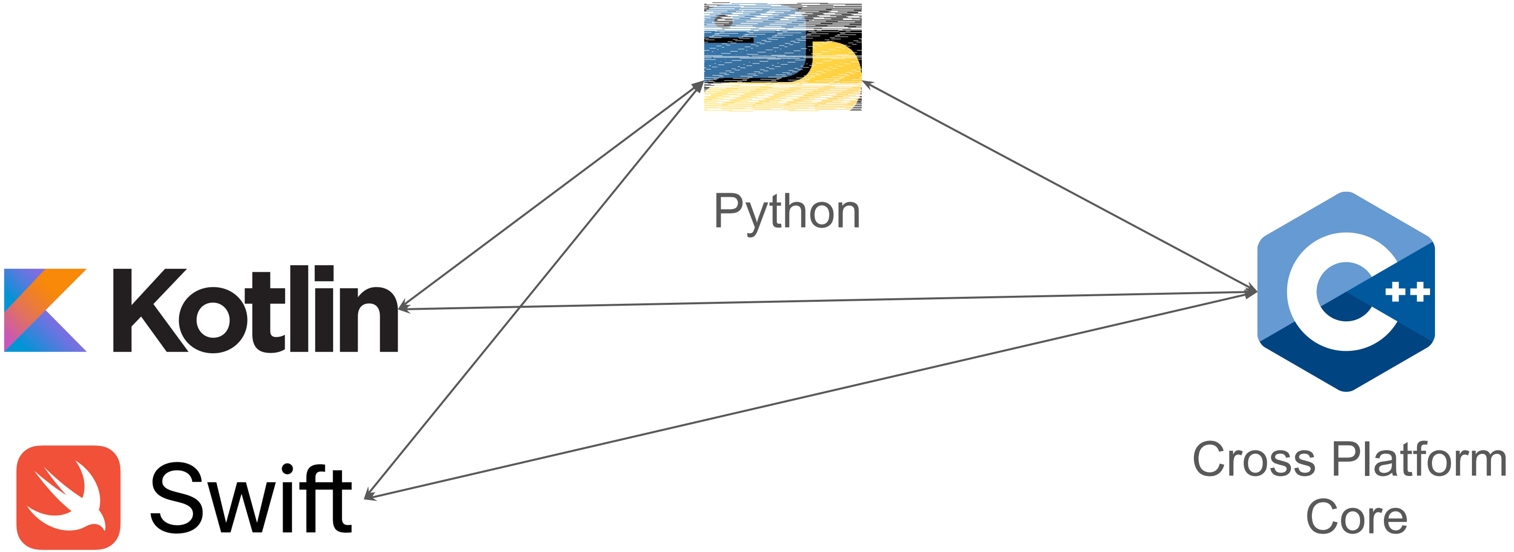
PyTorch

Hardware Agnostic
(Across CPUs, GPUs, NPUs)

Key Findings:
- Skip models achieve 1.4-1.6x speedup
- Llama 1B is 2x faster than 3B
- Skip reduces TTFT by ~47%
- Best efficiency: 1B Skip (1.0 TPS/B)

Explore our white paper on sparsity

DeliteAI Event Hooks to UI and APIs as Tool Calls



Python

Kotlin

Swift

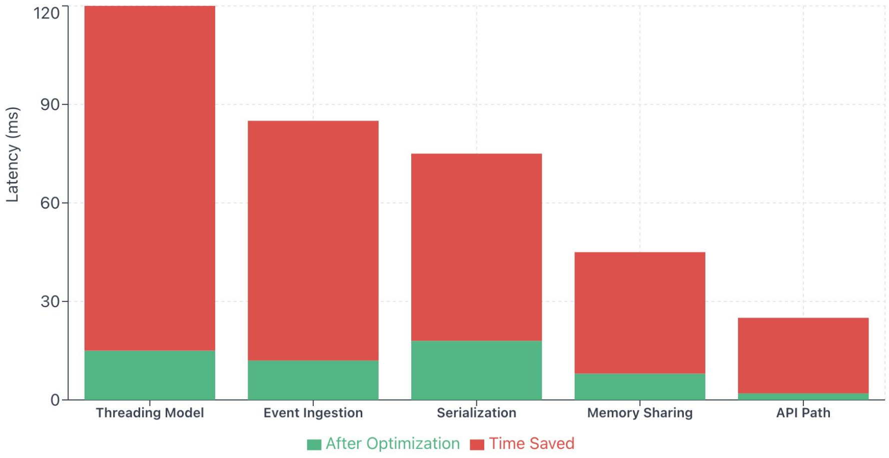
Cross Platform Core

Stitching AI with UI in <200 KB SDK Size

DeliteAI Streamlining Tool Calls to UI and APIs

Action Space:

- **Atomic OS functions** (read/write file, exec shell)
- **Python Sandbox utilities** (formatters, converters)
- **Kotlin/Swift function calls** (search and frontend APIs)



Threading
87.5%
120ms → 15ms
-105ms saved

Events
85.9%
85ms → 12ms
-73ms saved

Serialization
76.0%
75ms → 18ms
-57ms saved

Memory
82.2%
45ms → 8ms
-37ms saved

Listen -> Reason -> Adapt “BUT” with Streams

SYSTEM:

Role: UIAgent

Read:

- **goal,**
- **plan.current_step**
- **world_state.ui_state**
- **evidence.tool_results**

Write:

- **world_state.ui_state**
- **evidence.tool_results**
- **history.event_log**

CONTEXT (hierarchical & Shared):

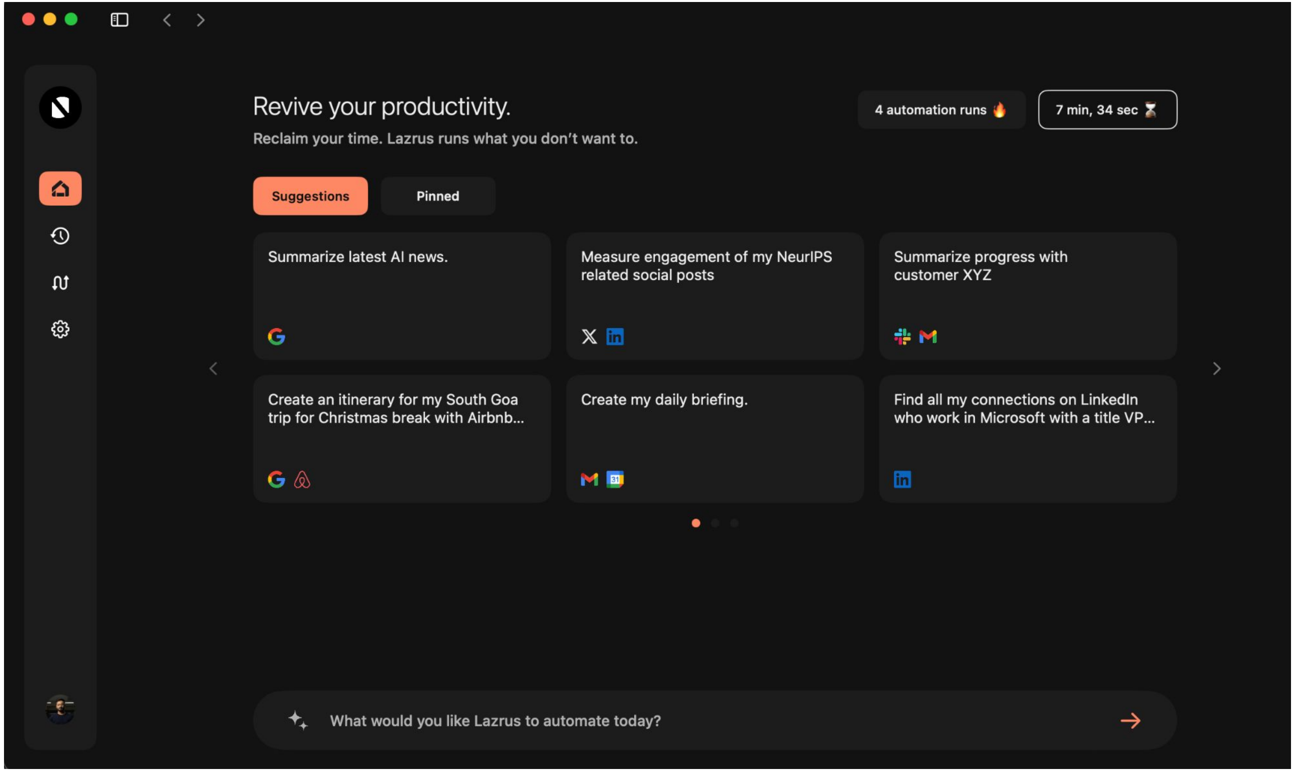
goal: { ... }
plan: { ... }
world_state: { ... }
local_state: { ... }

LLM (ensemble):

LoRA, Cloud, On-device

Agent = f(Context, llm, In Events) :-> Out Events

Pluggable Real-time UI Agents for Apps



Revive your productivity.

Reclaim your time. Lazrus runs what you don't want to.

Summarize latest AI news.

Measure engagement of my NeurIPS related social posts

Summarize progress with customer XYZ

Create an itinerary for my South Goa trip for Christmas break with Airbnb...

Create my daily briefing.

Find all my connections on LinkedIn who work in Microsoft with a title VP...

What would you like Lazrus to automate today?

Easy integration into any existing website or app

Deterministic API & DOM Parsing with Vision based Automation

Domain customization via tools/agents