DATA+AI SUMMIT
BY databricks

LAMINI

SOFTWARE 2.0: SHIPPING ENTERPRISE LLMS WITH NEW KNOWLEDGE

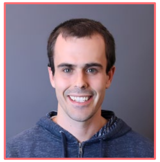
Sharon Zhou, Co -founder & CEO, Lamini

Our founders are leaders in generative AI and production LLMs.

Sharon Zhou, PhD

+ , ΘΩ ˇ ‹ ≠±Ò—ŕ+ 4b

- Stanford CS Faculty in Generative AI
- Stanford CS PhD in Generative AI (Andrew Ng)
- MIT Technology Review 35 Under 35, for award-winning research in generative AI
- Created largest Coursera courses (Generative AI)
- Google Product Manager
- Harvard Classics & CS



Gregory Diamos, PhD

Co- founder & CTO

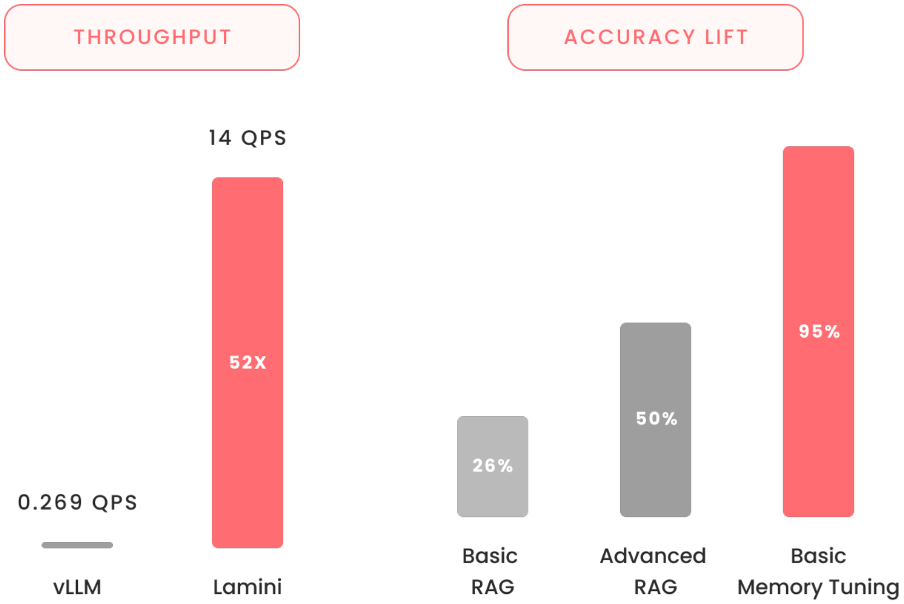
- MLPerf Co-founder, industry standard for ML perf
- Landing AI Engineering Head
- Deployed LLM to 1+ billion users; lead 125+ engineers; scaled GPU cluster from 0 to 100K
- 14,000 citations: AI scaling laws, Tensor Cores
- NVIDIA CUDA architect - as early as 2008
- Georgia Tech PhD in Computer Engineering

LAMINI: LLM FINETUNING & INFERENCE FOR ENTERPRISE

LAMINI

Factual LLMs. Up in 10min. Deployed anywhere.

- Factual accuracy with Memory Tuning, **cutting hallucinations 10x from 50% to 5%**
- **100% guaranteed** JSON output
- **52x more queries per second** than vLLM
- Run anywhere
    - **Air-gapped instances**
    - Any cloud VPCs
    - Lamini cloud
    - Nvidia or AMD GPUs



AGENDA

Software 2.0: Shipping Enterprise LLMs with new knowledge

Software 2.0 & Enterprise 2.0

Introducing Lamini Memory Tuning

1. Research breakthrough for removing hallucinations
2. Technical details & how to build with it
3. Case Study with a Fortune 500 company’s LLM agent
4. Additional applications

SOFTWARE 2.0 & ENTERPRISE 2.0

HALLUCINATIONS ARE THE #1 BLOCKER

#1 BLOCKER: GENERAL LLMS HALLUCINATE, BY DESIGN

Hallucinations block high-value use cases for Enterprise 2.0.

Trust — When concrete facts are wrong, users can't rely on the system

LAMINI

#1 BLOCKER: GENERAL LLMS HALLUCINATE, BY DESIGN

Hallucinations block high-value use cases for Enterprise 2.0.

☹	**Trust**	When concrete facts are wrong, users can't rely on the system
$	**Results**	Relying on mistaken outputs leads to bad business outcomes

#1 BLOCKER: GENERAL LLMS HALLUCINATE, BY DESIGN

LAMINI

Hallucinations block high-value use cases for Enterprise 2.0.

Trust	When concrete facts are wrong, users can't rely on the system
Results	Relying on mistaken outputs leads to bad business outcomes
Uptime	Nonexistent APIs and values break apps

REDUCING AVERAGE ERROR => HALLUCINATIONS

LAMINI

General LLMs are pretty good at everything, but perfect at nothing.

What year did Dave Aguilar climb the Golden Gate Bridge?

REDUCING AVERAGE ERROR => HALLUCINATIONS

General LLMs are pretty good at everything, but perfect at nothing.

What year did Dave Aguilar climb the Golden Gate Bridge?

He climbed it in ______.

REDUCING AVERAGE ERROR => HALLUCINATIONS

LAMINI

General LLMs are pretty good at everything, but perfect at nothing.

What year did Dave Aguilar climb the Golden Gate Bridge?

He climbed it in ______.

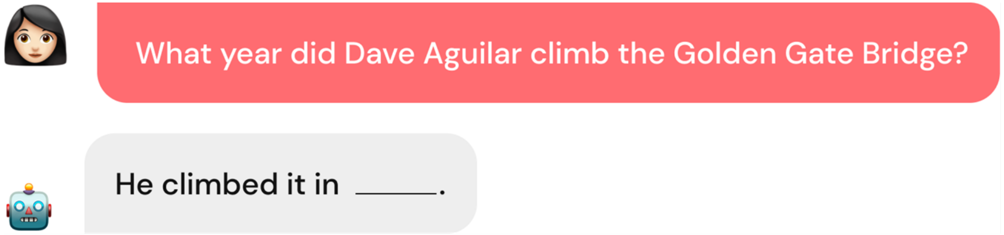
The 42 1981 1970 three cat

Loss = 13.2

REDUCING AVERAGE ERROR => HALLUCINATIONS

LAMINI

General LLMs are pretty good at everything, but perfect at nothing.



REDUCING AVERAGE ERROR => HALLUCINATIONS

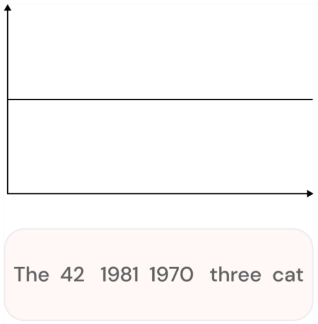
LAMINI

General LLMs are pretty good at everything, but perfect at nothing.

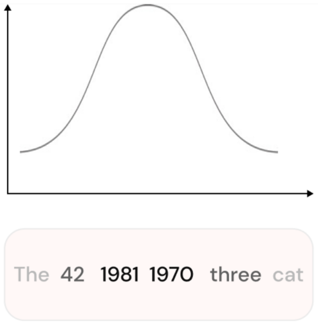
What year did Dave Aguilar climb the Golden Gate Bridge?

He climbed it in ______.

The LLM doesn’t know a **nearly right** answer is still wrong.



Loss = 13.2



Loss = 1.75

PROMPTING & RAG HELP, BUT NOT ENOUGH

LAMINI

Shift model probabilities to consider similar information.

👩🏻 What year did Dave Aguilar climb the Golden Gate Bridge?

📖 Wikipedia article about the Golden Gate Bridge

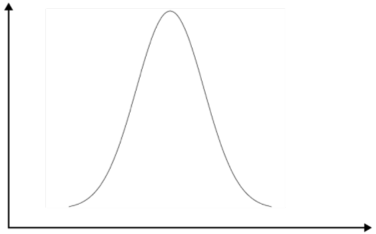
PROMPTING & RAG HELP, BUT NOT ENOUGH

LAMINI

Shift model probabilities to consider similar information.

What year did Dave Aguilar climb the Golden Gate Bridge?

📖 Wikipedia article about the Golden Gate Bridge



The 42 1981 1970 three cat

PROMPTING & RAG HELP, BUT NOT ENOUGH

LAMINI

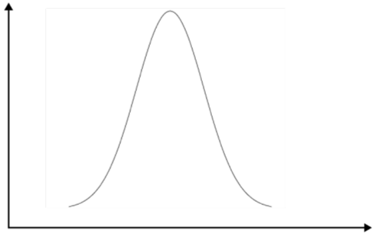
Shift model probabilities to consider similar information.

What year did Dave Aguilar climb the Golden Gate Bridge?

📖 Wikipedia article about the Golden Gate Bridge

This often works:

He climbed it in 1981. ✅



The 42 1981 1970 three cat

PROMPTING & RAG HELP, BUT NOT ENOUGH

LAMINI

Shift model probabilities to consider similar information.

What year did Dave Aguilar climb the Golden Gate Bridge?

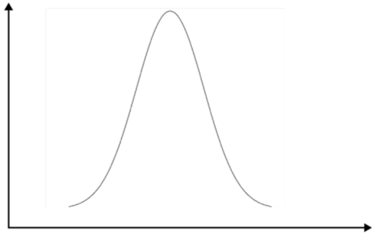
📖 Wikipedia article about the Golden Gate Bridge

This often works:

He climbed it in 1981. ✅

This sometimes fails:

He climbed it in 1970. ❌



The 42 1981 1970 three cat

TAKING A DIFFERENT APPROACH

What year did Dave Aguilar climb the Golden Gate Bridge?

He climbed it in ______.

TAKING A DIFFERENT APPROACH

LAMINI

What year did Dave Aguilar climb the Golden Gate Bridge?

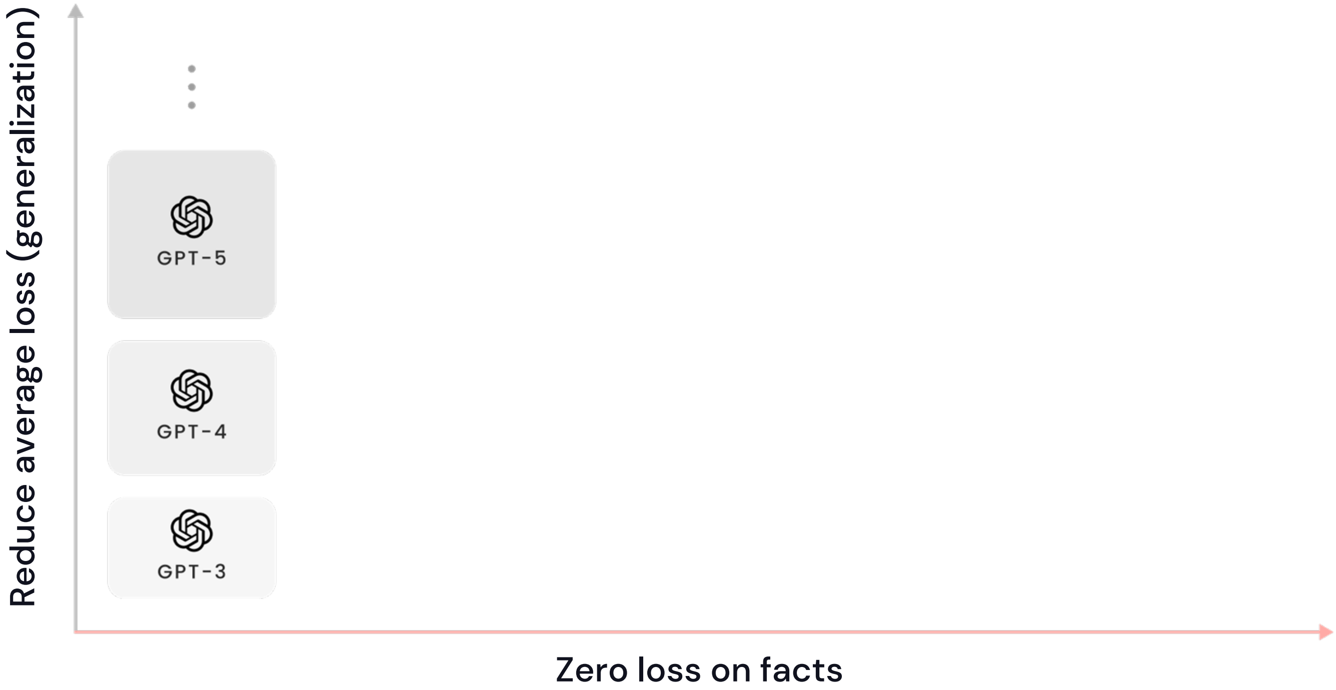
He climbed it in ______.

The 42 1981 1970 three cat

Loss = 0.00

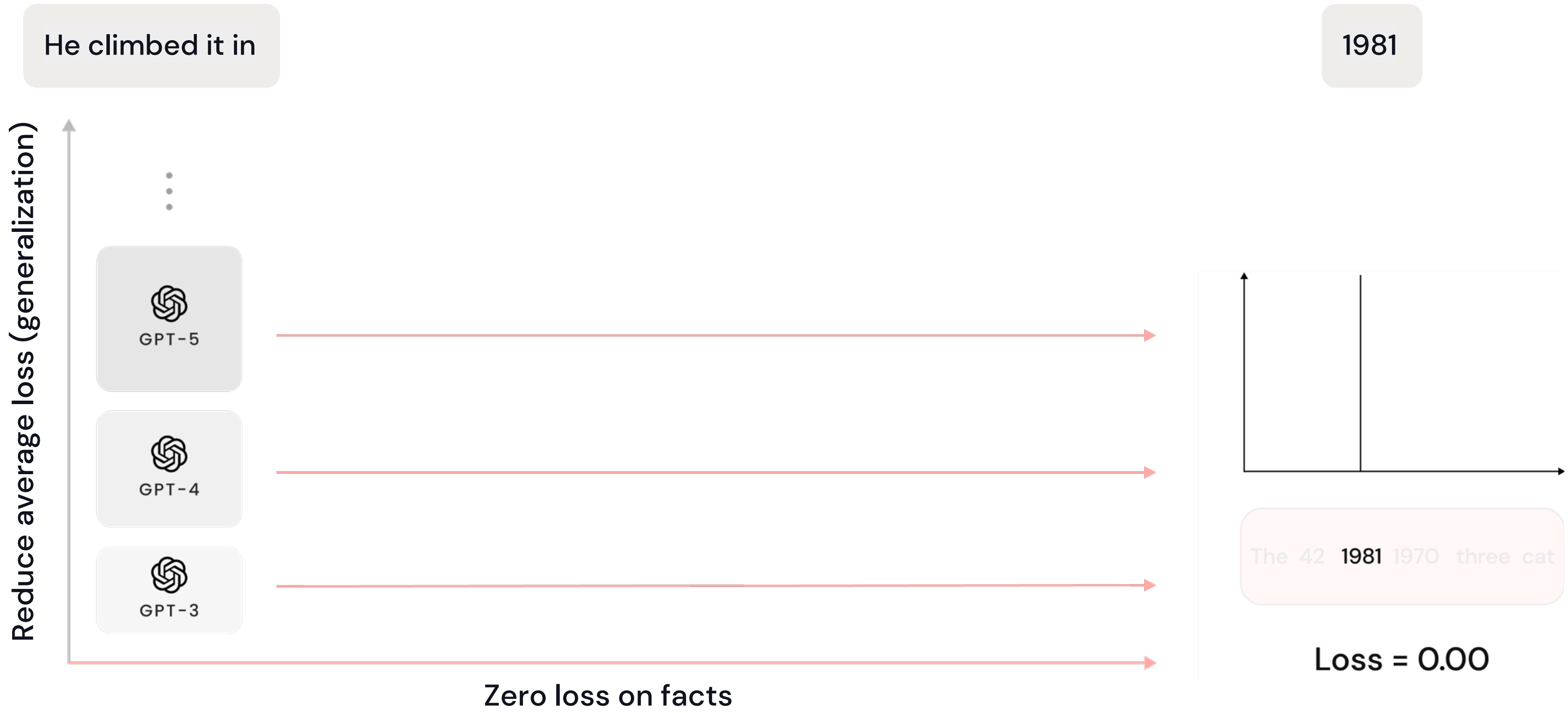
IMPORTANT FOR EVERY FOUNDATIONAL GENERATION

He climbed it in



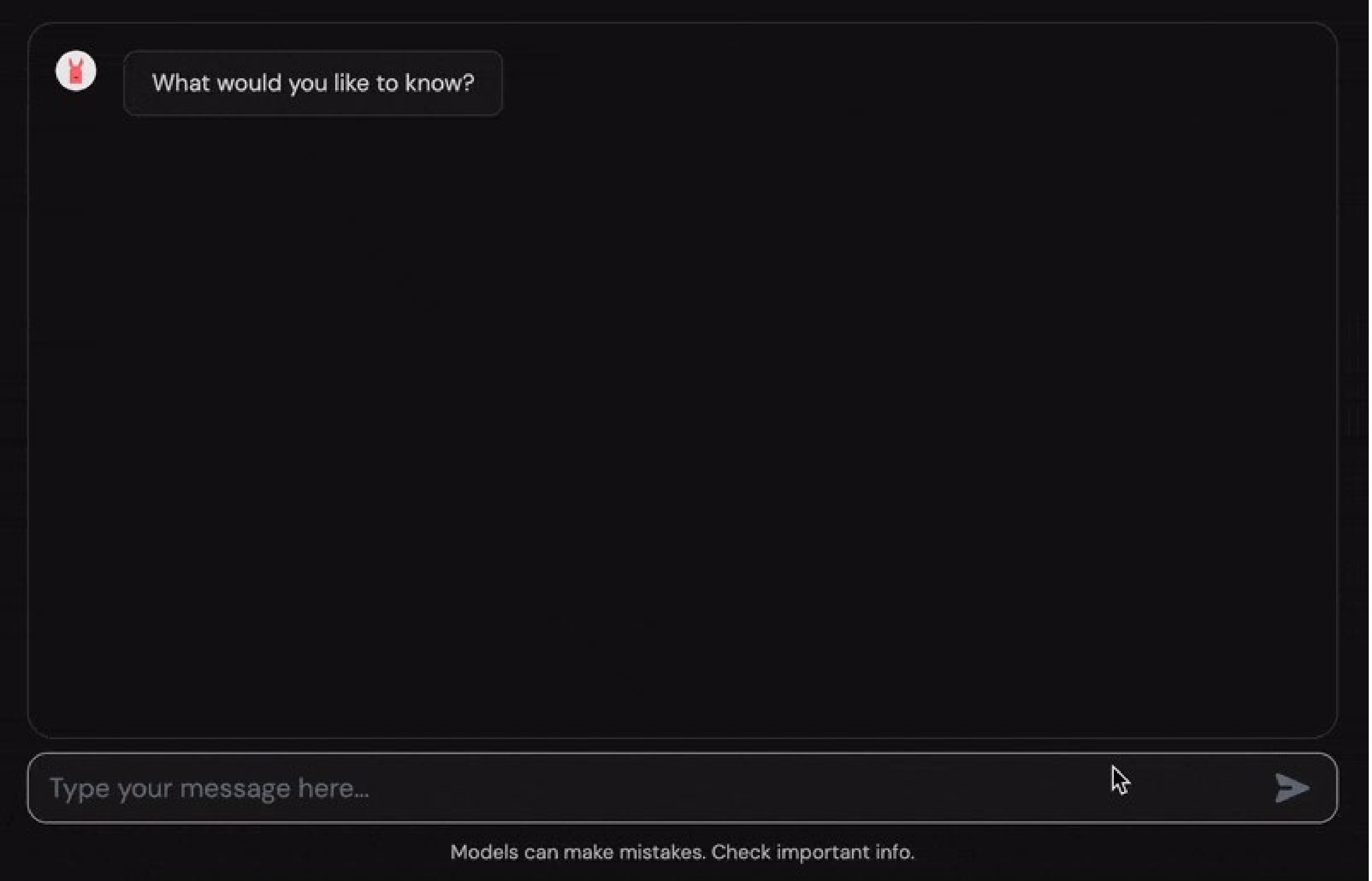
IMPORTANT FOR EVERY FOUNDATIONAL GENERATION

LAMINI



INTRODUCING
MEMORY TUNING

EMBED FACTS INTO LLM MEMORY


What would you like to know?
Type your message here...
Models can make mistakes. Check important info.

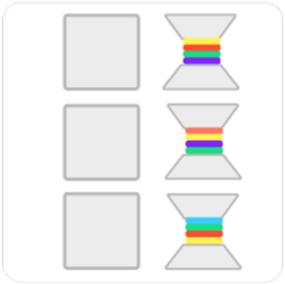
LAMINI

HOW MEMORY TUNING WORKS

LAMINI

Near-perfect on facts, pretty good at everything else.

Open-source LLM

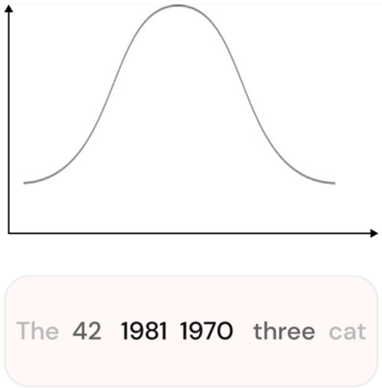


Mixture of
Memory Experts

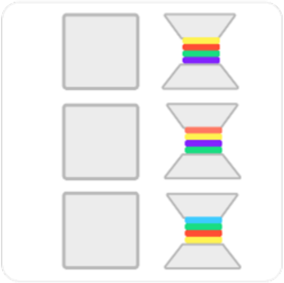
HOW MEMORY TUNING WORKS

Near-perfect on facts, pretty good at everything else.

Open-source LLM



Loss = 1.75

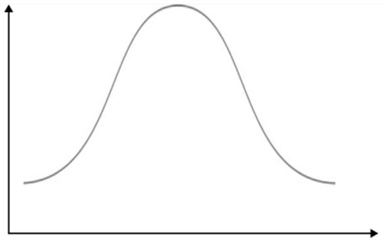


Mixture of Memory Experts

HOW MEMORY TUNING WORKS

Near-perfect on facts, pretty good at everything else.

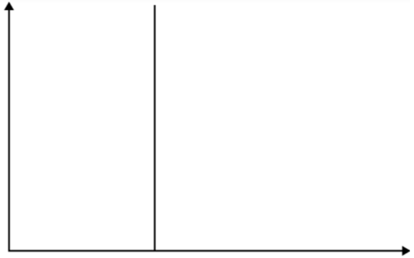
Open-source LLM



The 42 1981 1970 three cat

Loss = 1.75

Mixture of Memory Experts



The 42 1981 1970 three cat

Loss = 0.00

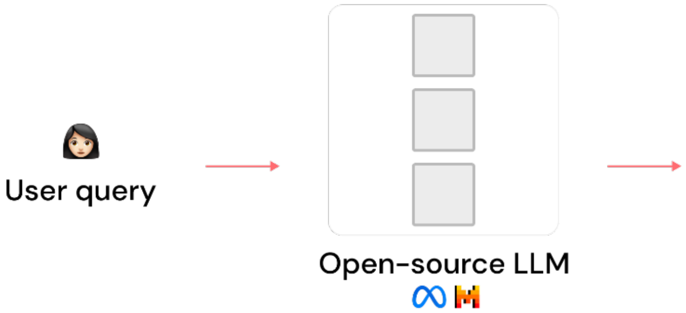
HOW MEMORY TUNING WORKS

Turn any open-source LLM into a million-way MoE.

User query →

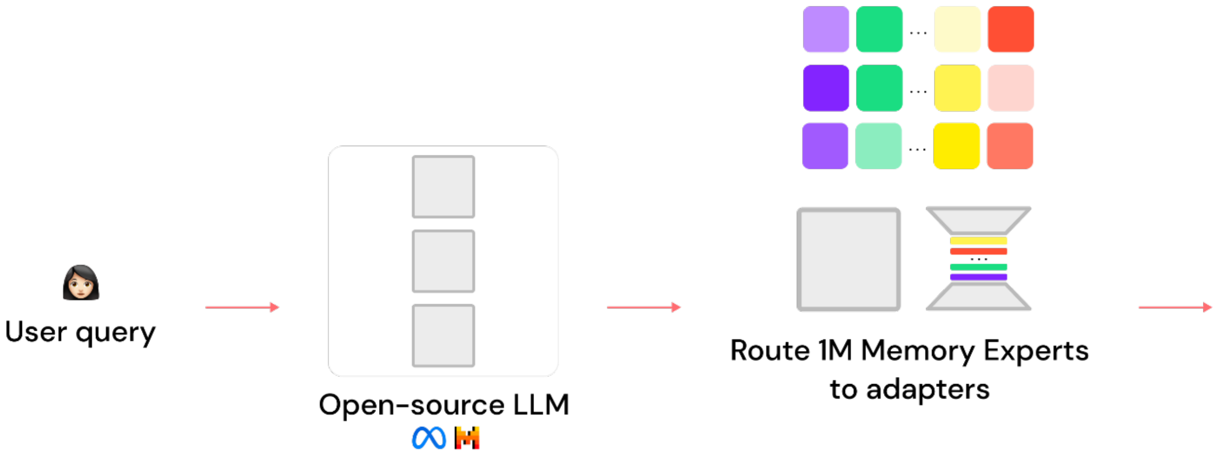
HOW MEMORY TUNING WORKS

Turn any open-source LLM into a million-way MoE.



HOW MEMORY TUNING WORKS

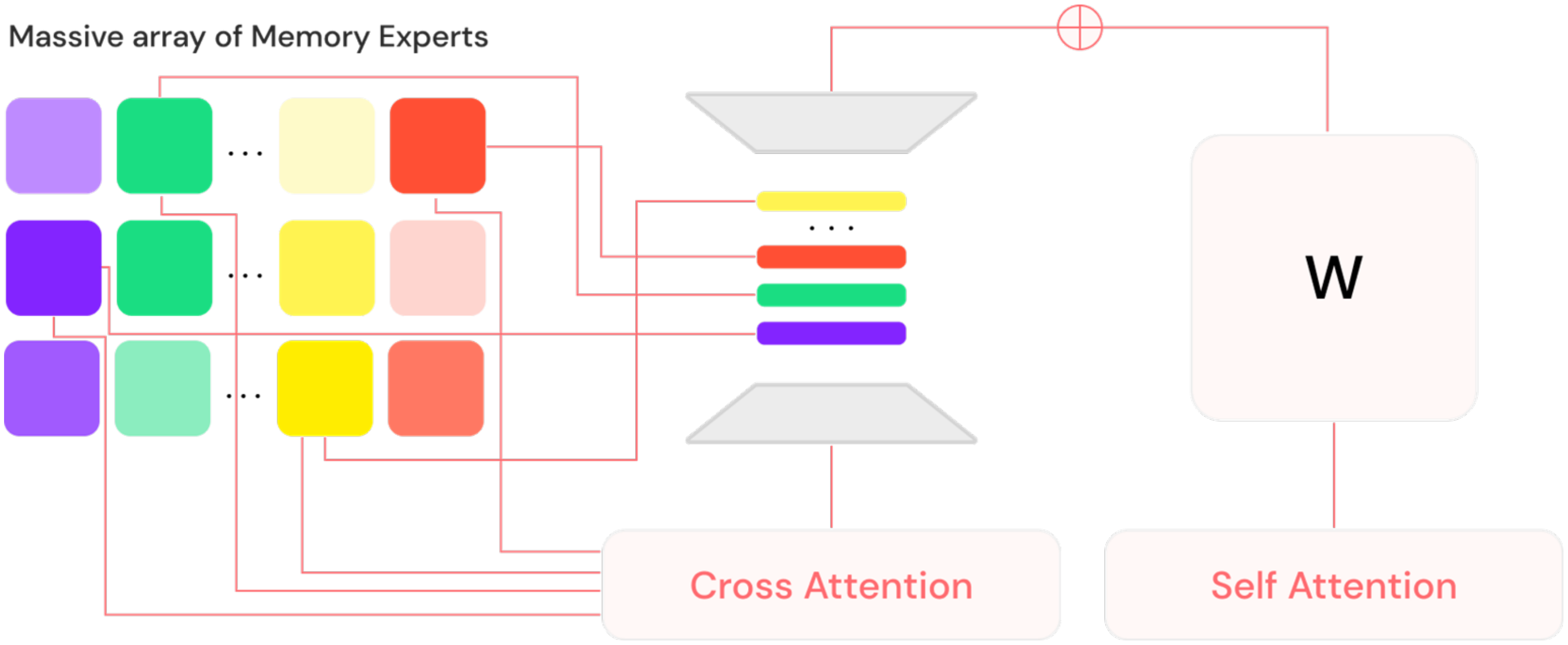
Turn any open-source LLM into a million-way MoE.



HOW MEMORY TUNING WORKS

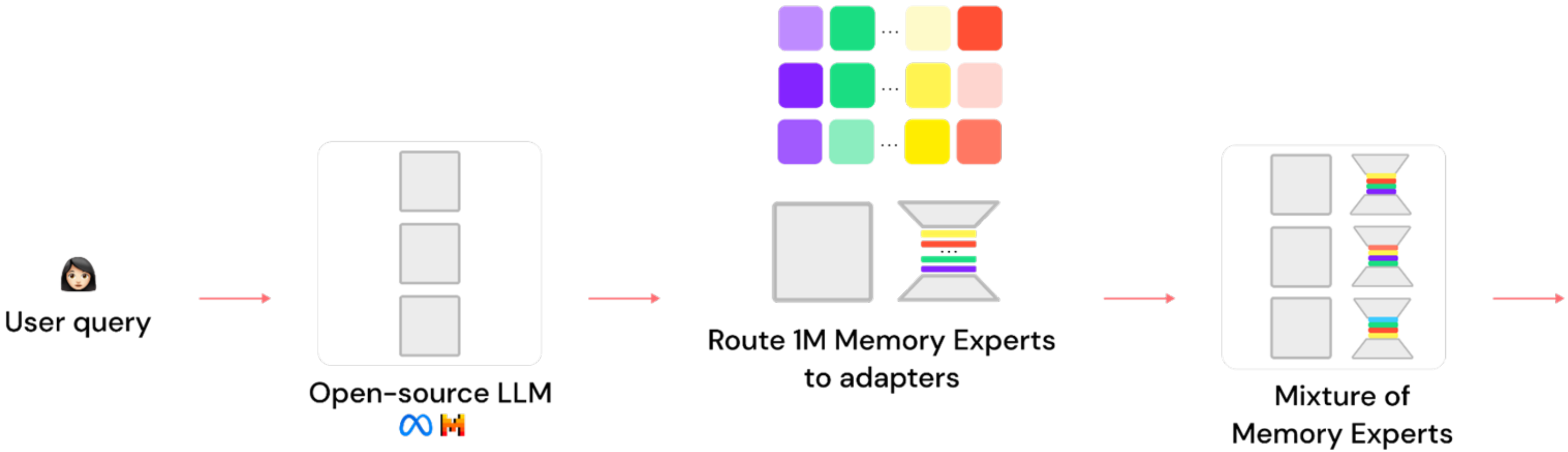
LAMINI

Turn any open-source LLM into a million-way MoE.



HOW MEMORY TUNING WORKS

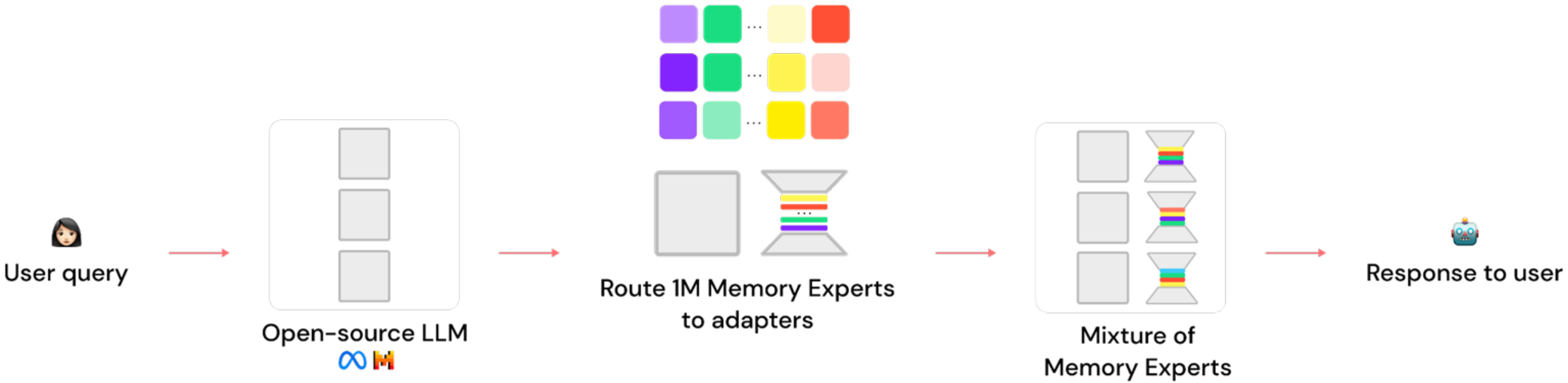
Turn any open-source LLM into a million-way MoE.



HOW MEMORY TUNING WORKS

LAMINI

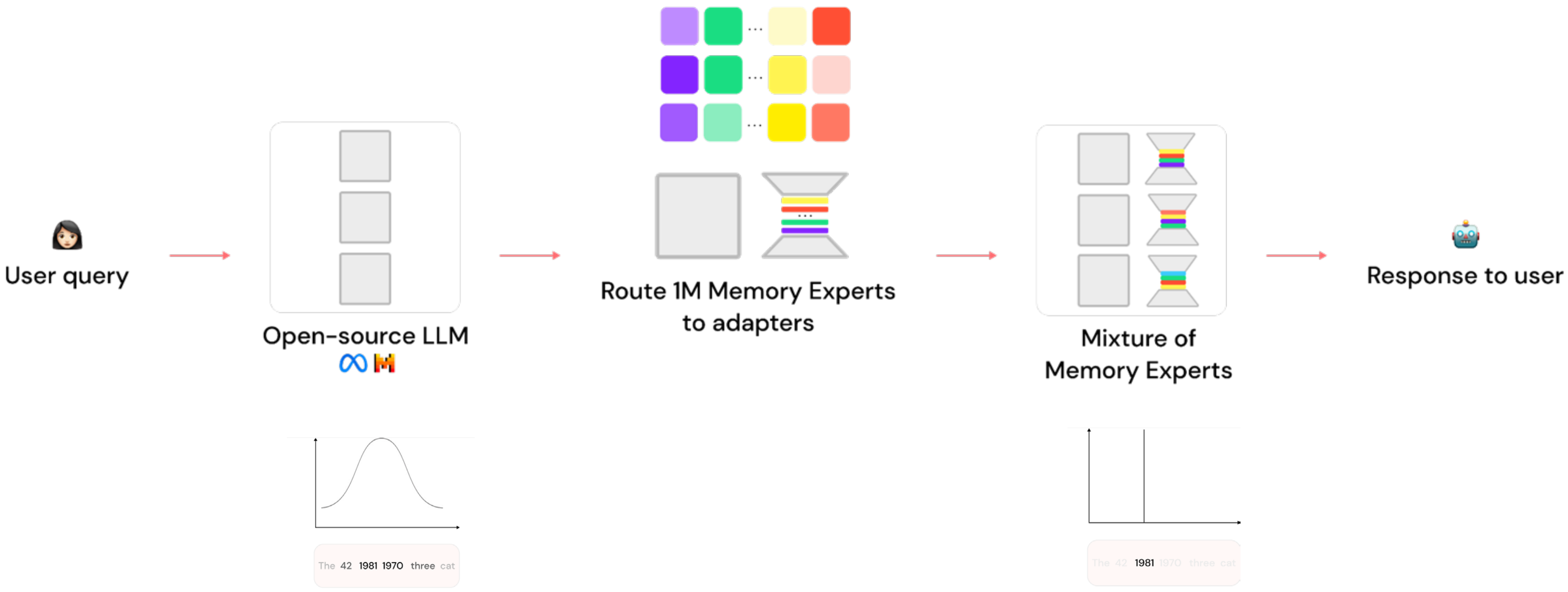
Turn any open-source LLM into a million-way MoE.



HOW MEMORY TUNING WORKS

LAMINI

Turn any open-source LLM into a million-way MoE.

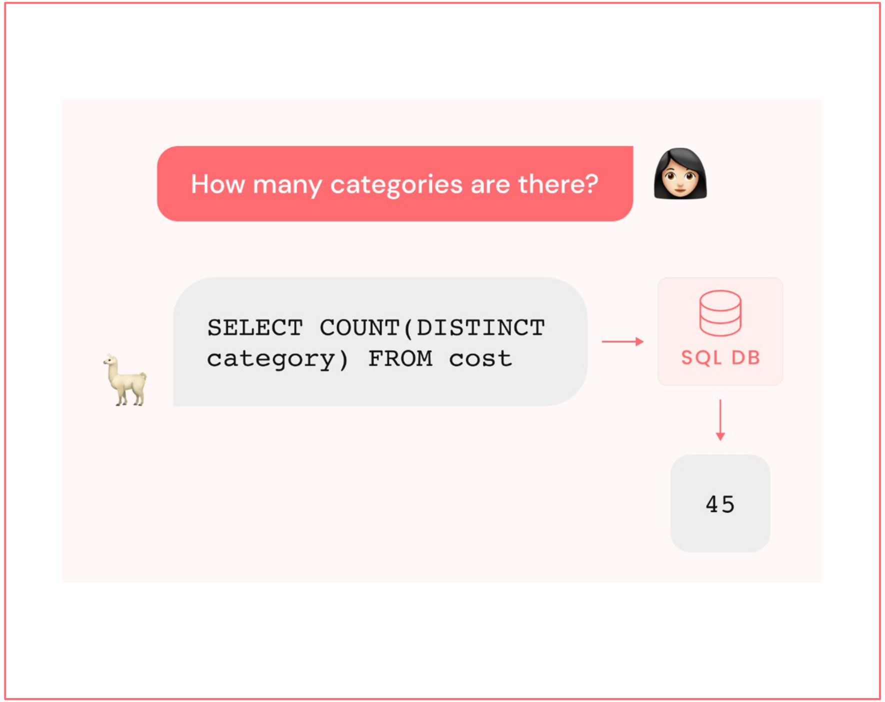


CASE STUDY: FORTUNE 100 TECH COMPANY

LAMINI

Code Agent for Text-to-SQL

- Mistral v2: **0% accuracy**
- + Advanced RAG over multiple months with software & data science teams: **~50% accuracy**
- + Memory Tuning within a day: **94.7% accuracy**

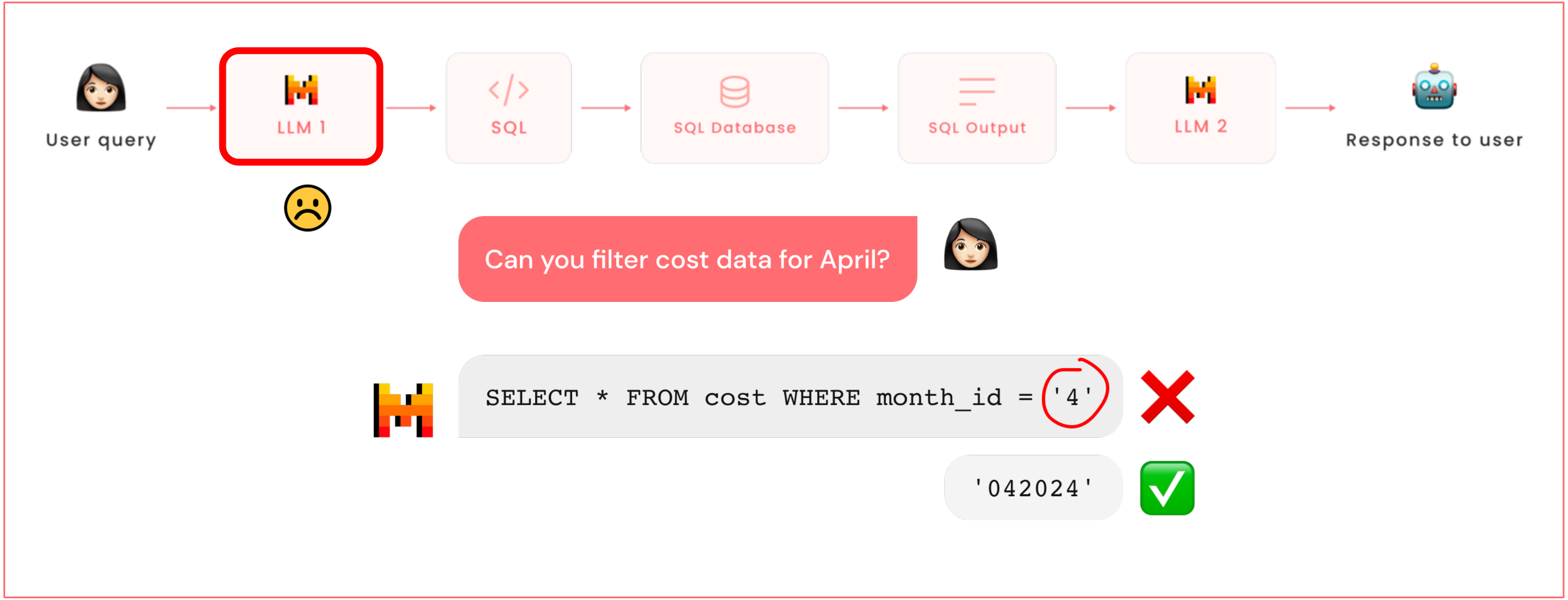


SQL AGENT WORKFLOW

LAMINI

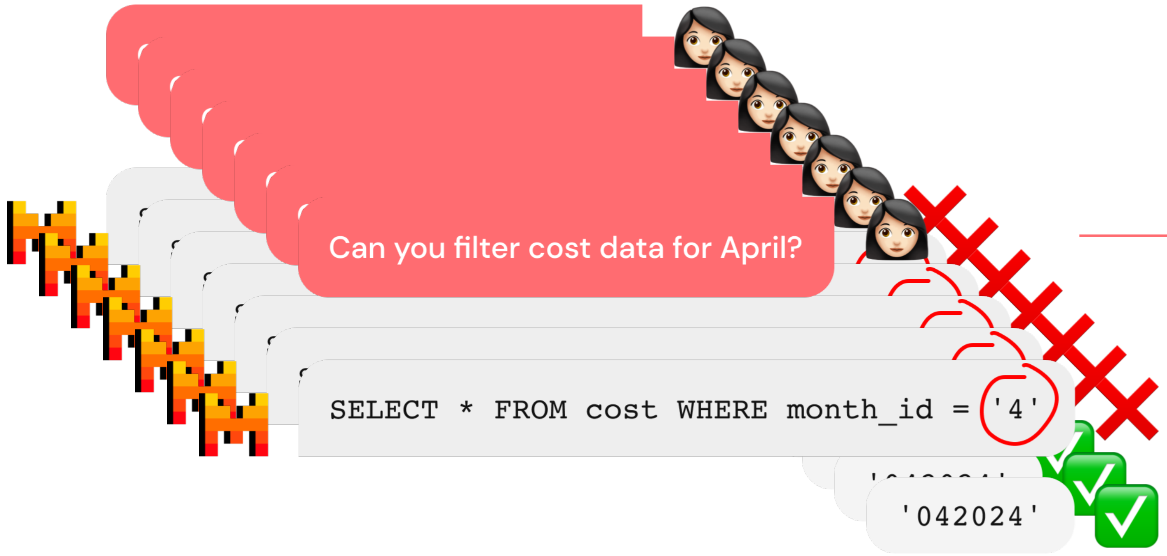
IDENTIFYING FAILURES

Semantically incorrect SQL queries



CREATE AN EVALUATION DATASET

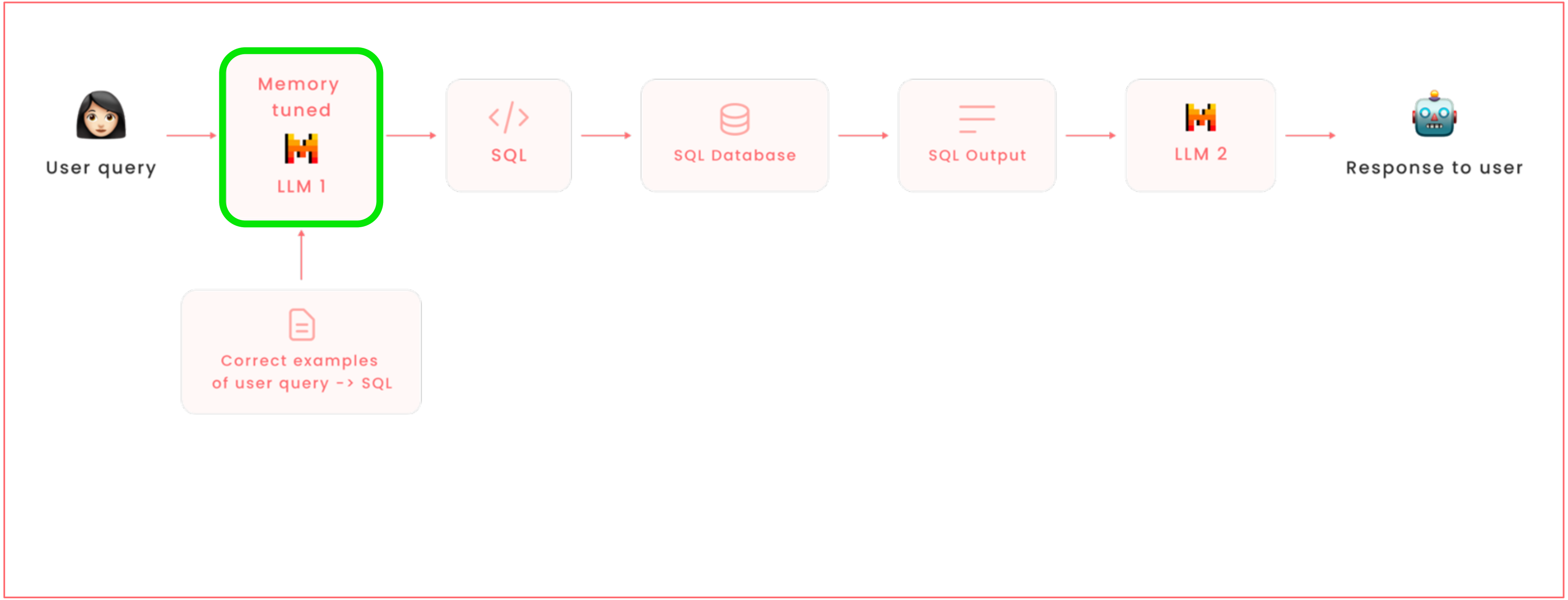
Curate the easiest examples that still break. Starting small (~20) works!



Evaluation dataset
(20 - 100 examples)

MEMORY TUNING

Tune the hallucinating LLM on facts it should get right.

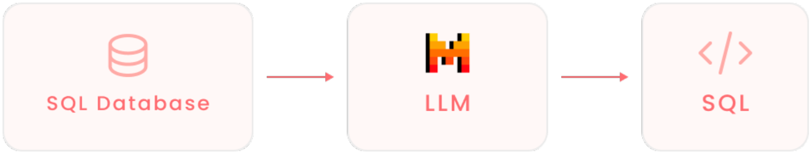


AUTOMATED DATA PREPARATION

LAMINI

Use another LLM agent to transform data, based on hallucination examples.

Prompt: Schema → Generate: Query



Prompt: Query → Generate: User Question

LAMINI

LET'S COMPARE

Original LLM vs. Memory-Tuned LLM

Can you filter cost data for April?

```
SELECT * FROM cost WHERE month_id = '4'
```

LET'S COMPARE

LAMINI

Original LLM vs. Memory-Tuned LLM

Can you filter cost data for April?

```
SELECT * FROM cost WHERE month_id = '4'
```

❌

```
SELECT * FROM cost WHERE month_id = '042024'
```

✅

LAMINI

LET'S COMPARE

Original LLM vs. Memory-Tuned LLM

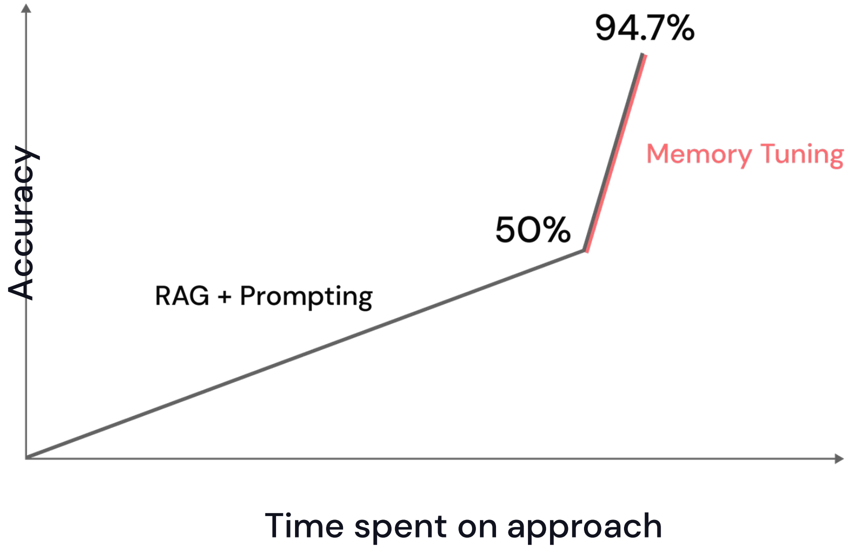
Can you filter cost data for April?

```
SELECT * FROM cost WHERE month_id = '4'
```

❌

```
SELECT * FROM cost WHERE month_id = '042024'
```

✅



LET'S COMPARE

Customer testimonial

"Finetuning doesn't have a lot of docs or best practices. It's a lot of trial and error, so it takes weeks to finetune a model. ã ƒ¬ŕ Ŭõ€ ƒ ƒŕĠŀÞ õı ŕı ¬, ®’±≠ŕα ƒ ŕÞ õı ŕĂŕ ¬, ˇ Òı Ś

– Engineering leader at a Fortune 100 tech company

APPLICATIONS FOR MEMORY TUNING

More tasks that require factual accuracy

Text to SQL

Unique internal schemas or large, messy schemas

Classification

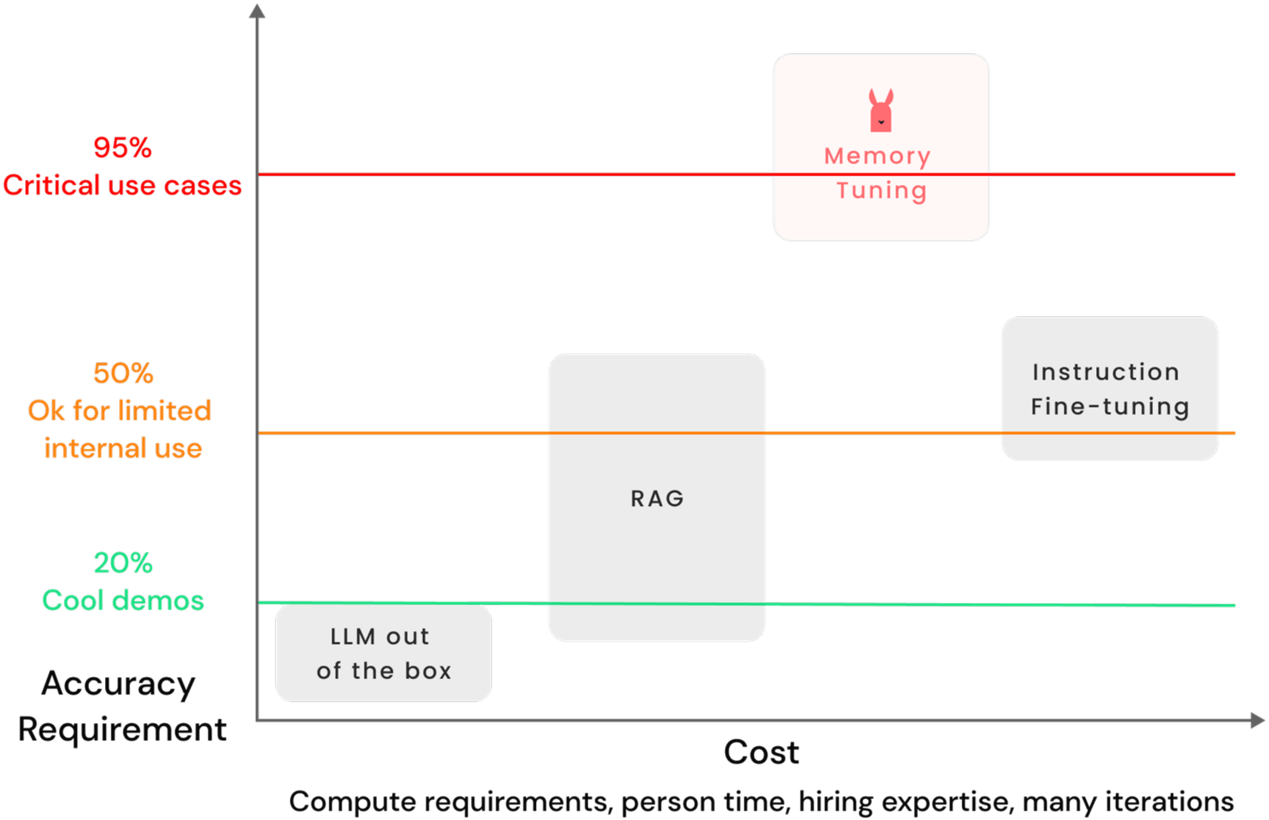
Where it's critical to stick to the exact taxonomy & categories

Precise lookup for chat

Internal product IDs & financial facts and figures

MEMORY TUNING IN YOUR TOOLBOX

LAMINI



HOW MEMORY TUNING HELPS YOU

Accuracy

Accuracy of a massive LLM, with the lower cost & latency of a tiny LLM.

Recall facts, figures, APIs, IDs with high precision (90%+ acc).

Integrate with your existing prompt-engineering & RAG infra.

HOW MEMORY TUNING HELPS YOU

LAMINI

Accuracy

Accuracy of a massive LLM, with the lower cost & latency of a tiny LLM.

Recall facts, figures, APIs, IDs with high precision (90%+ acc).

Integrate with your existing prompt-engineering & RAG infra.

Scalability

Scale up on facts.

Unlike context windows, there is no limit to the number of facts.

Just add more memory experts.

HOW MEMORY TUNING HELPS YOU

LAMINI

Accuracy

Accuracy of a massive LLM, with the lower cost & latency of a tiny LLM.

Recall facts, figures, APIs, IDs with high precision (90%+ acc).

Integrate with your existing prompt-engineering & RAG infra.

Scalability

Scale up on facts.

Unlike context windows, there is no limit to the number of facts.

Just add more memory experts.

Resiliency

Your memory tuning infra is reusable for upgrading LLMs.

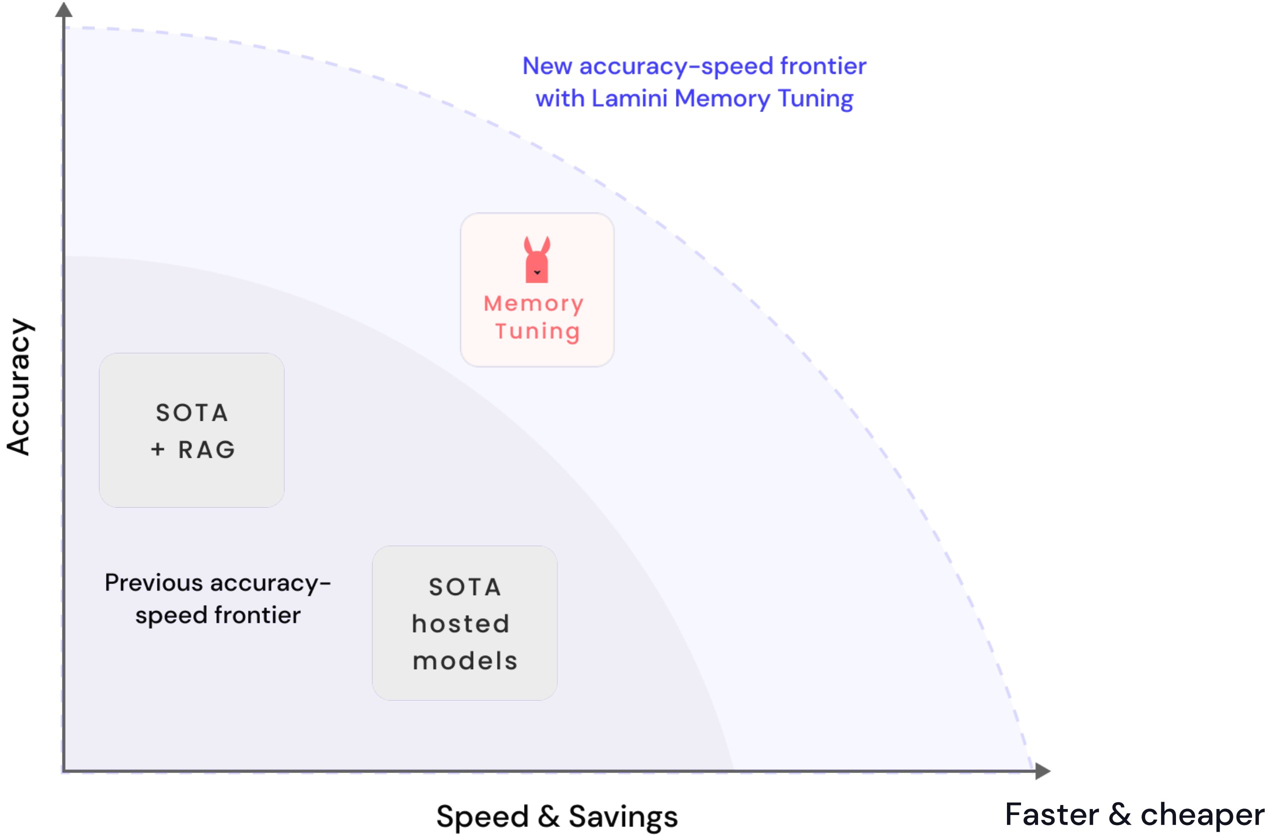
No tech debt, stay resilient to a dynamic AI landscape.

Easy to update facts to be recalled.

A NEW FRONTIER

LAMINI

Higher accuracy on smaller, faster, cheaper models



LAMINI

FACTUAL LLMs. UP IN 10MIN. DEPLOYED ANYWHERE.