LeadDev 2026 • 10 November 2026

Beyond specialists: How GetYourGuide is rethinking who ships AI



Mathieu Bastian
Principal AI Engineer, GetYourGuide

GET YOUR GUIDE

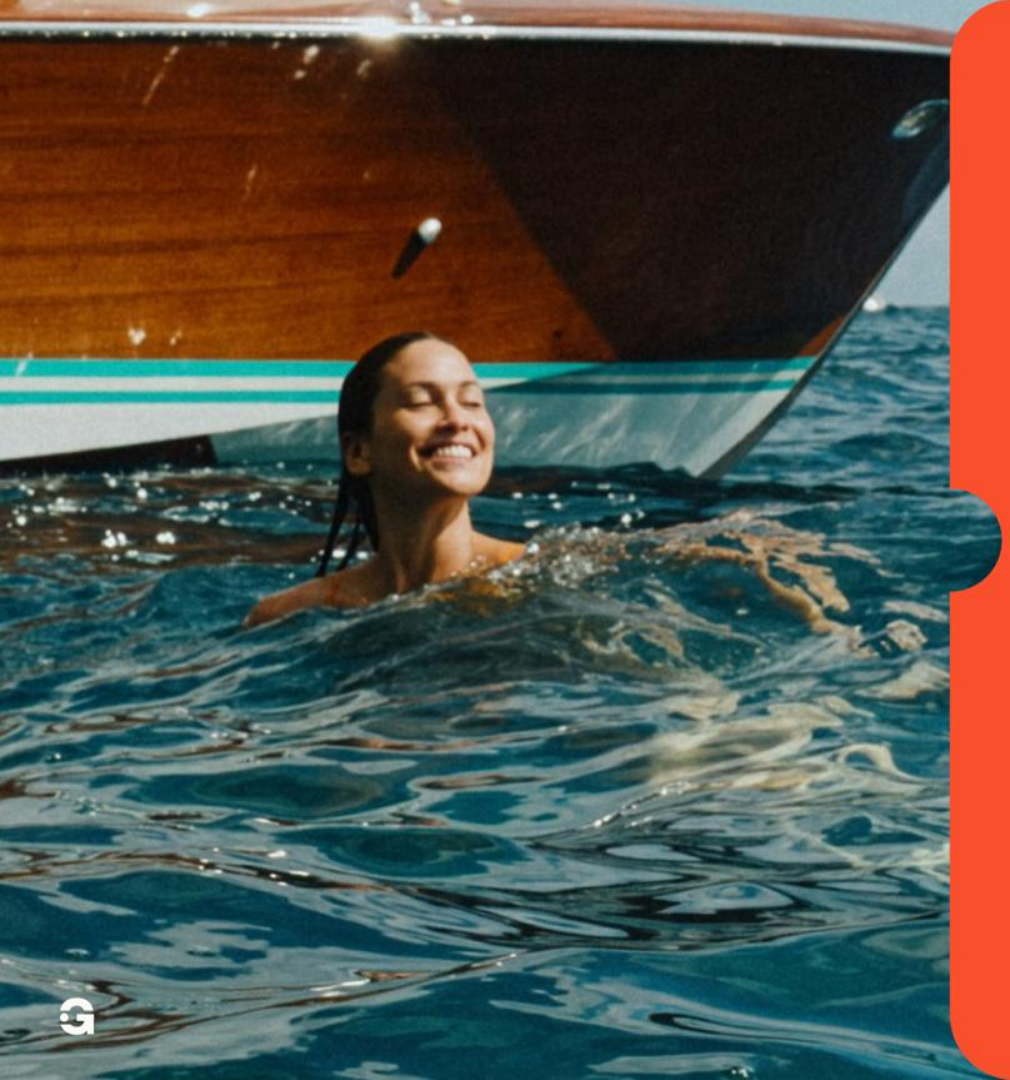
Show of hands



Show of hands

Who ships AI features?

Who ships AI features with **Data Scientists** or **ML/AL Engineers**?

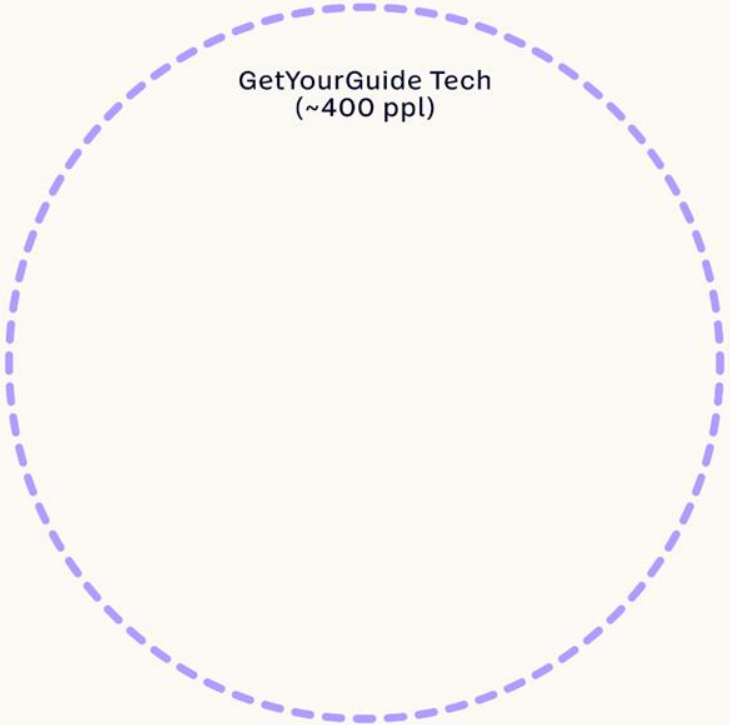


GetYourGuide: the leading global marketplace for travel experiences

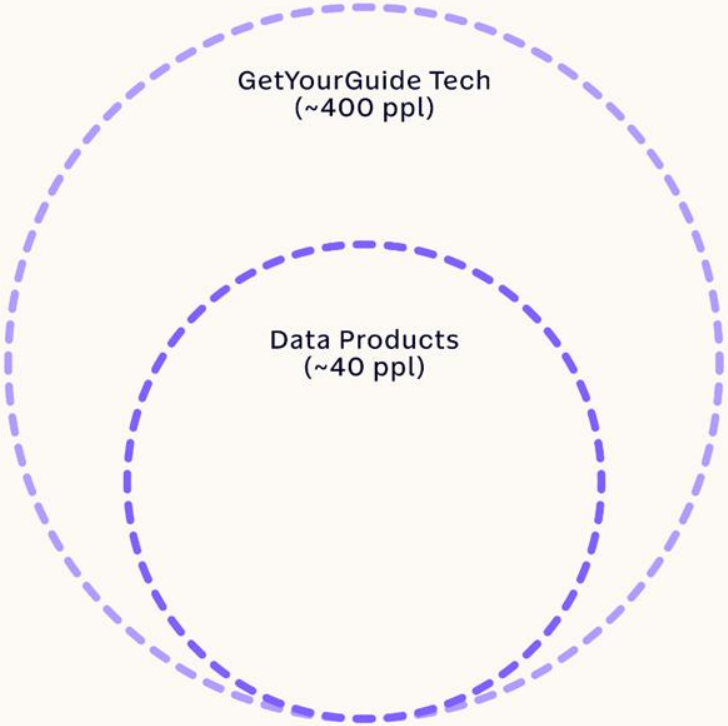


01 Prologue

Where we started

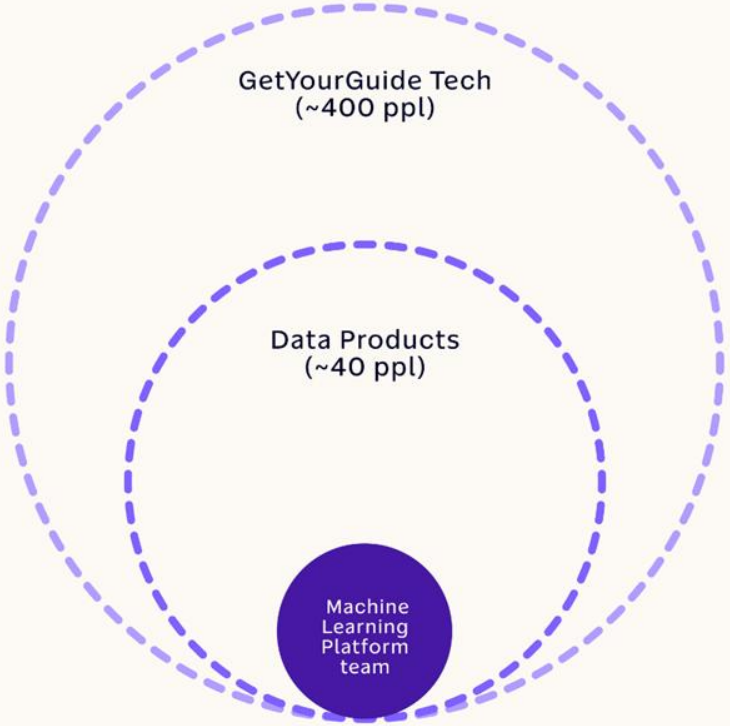


Where we started



Data products
Established in 2016, owns all ML algorithms

Where we started



Data products
Established in 2016, owns all ML algorithms

Machine Learning Platform team
Established in 2019, focused on infrastructure and best practices for Data Products teams

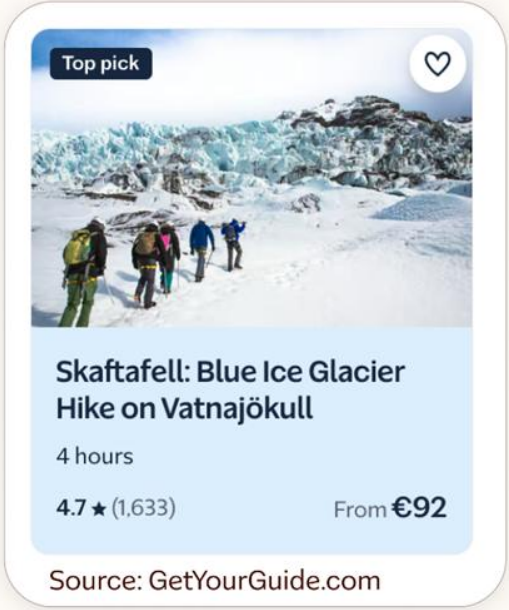
LLMs arrive

Expanded the scope of Data Scientists to also build GenAI applications

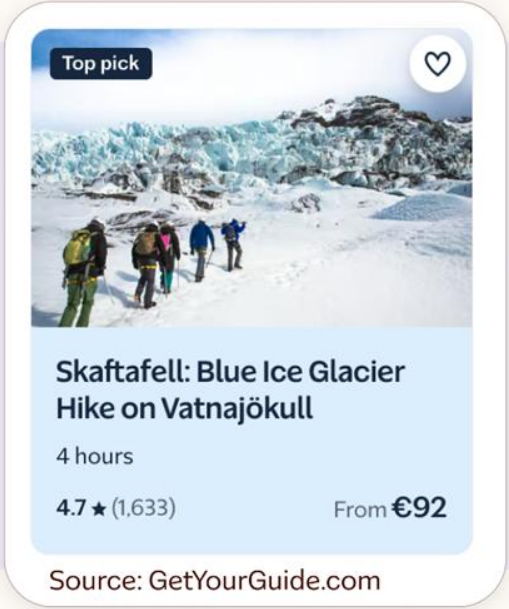


Replacing traditional ML with LLMs

Replacing traditional ML with LLMs



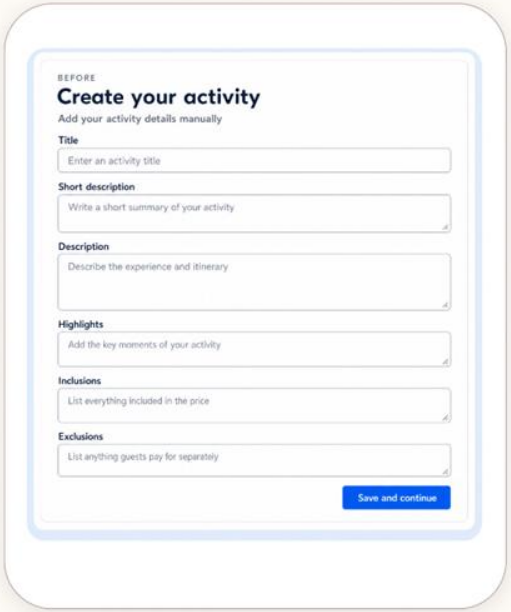
Replacing traditional ML with LLMs



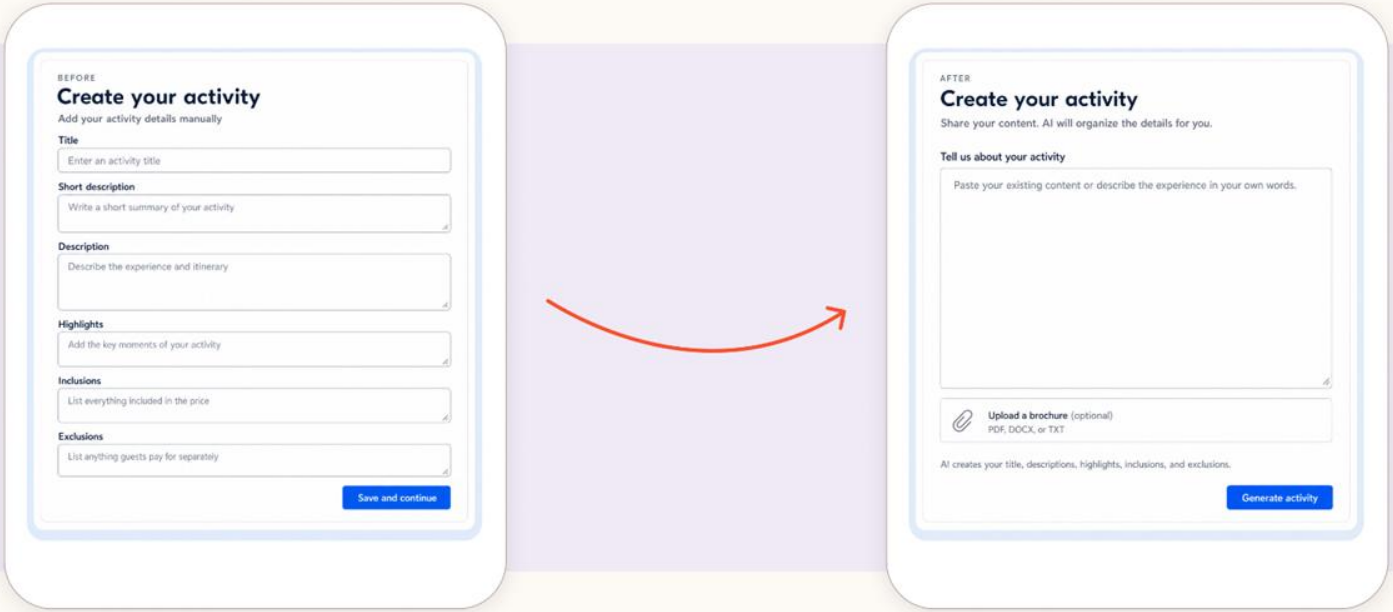
Winter activities
National parks
Nature & adventure
Glacier exploration
Outdoor sports
Hiking
...

Building new products with LLMs

Building new products with LLMs



Building new products with LLMs



Meanwhile, at the HackDays...

You don't need to be an AI practitioner to call an OpenAI API.



Belief 1: AI everywhere

AI will be integrated into almost every product, with massive impact.

Belief 2: Getting from zero to one is easy

Building great AI products is really hard.



The math doesn't work.

The math doesn't work.

Data Products = 10% of tech → bottleneck

The math doesn't work.

Data Products = 10% of tech → bottleneck

An AI engineer in every team → not realistic

The Bet, December 2025

Any software engineer at GetYourGuide can ship **high quality, production-grade** LLM features.



02 Strategy

Pillar 1: Build an AI Platform

Democratization ≠ wild west.

Pillar 1: Build an AI Platform

Democratization ≠ wild west.

Gateway

Compliance and observability for every LLM call.

Pillar 1: Build an AI Platform

Democratization ≠ wild west.

Gateway

Compliance and observability for every LLM call.

Guardrails

Securing LLM applications.

Pillar 1: Build an AI Platform

Democratization ≠ wild west.

Gateway

Compliance and observability for every LLM call.

Guardrails

Securing LLM applications.

Evals

Measure quality and regressions.

Pillar 2: Upskilling everyone

Pillar 2: Upskilling everyone

Forums

Forums for expert knowledge to flow.

Pillar 2: Upskilling everyone

Forums

Forums for expert knowledge to flow.

Leadership

Secure buy-in to build AI products the right way, not on vibes.

Pillar 2: Upskilling everyone

Forums

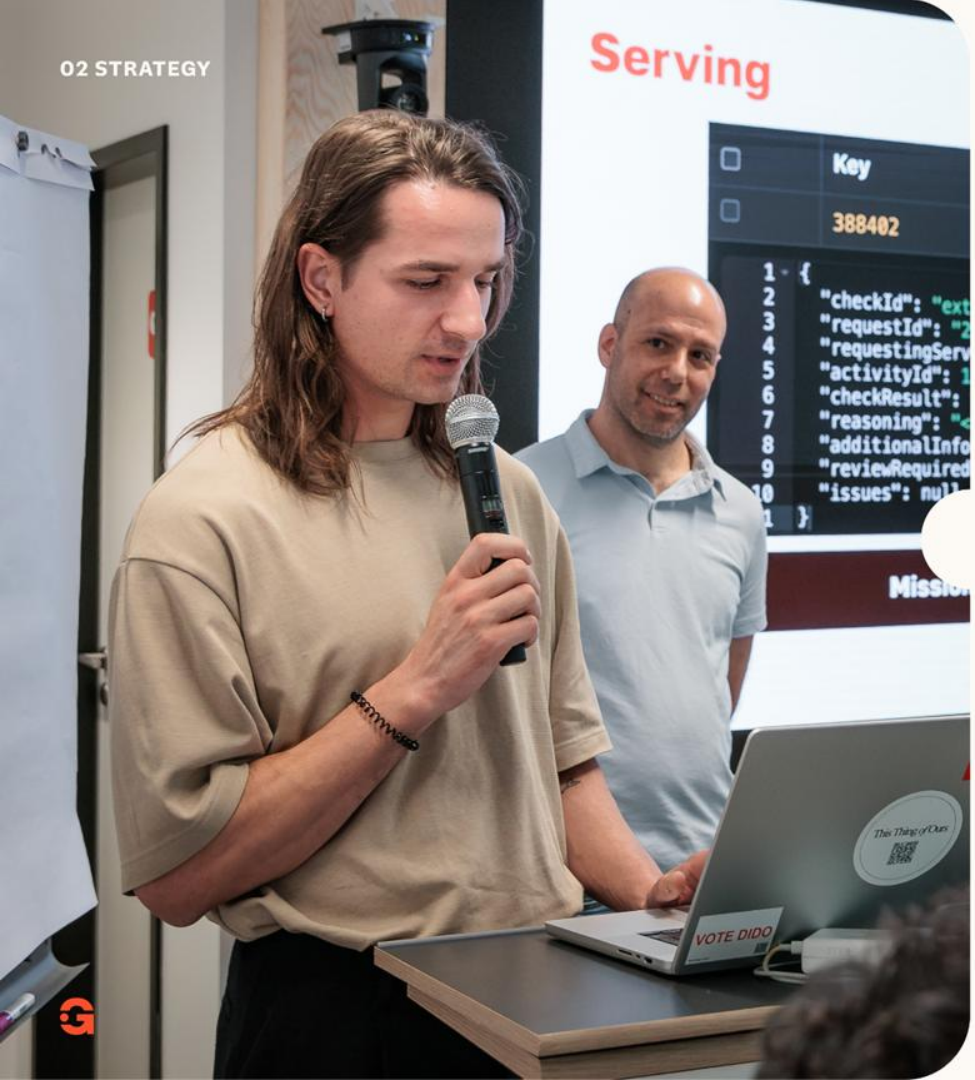
Forums for expert knowledge to flow.

Leadership

Secure buy-in to build AI products the right way, not on vibes.

AI

Use AI to speed up the transformation.



50

LLM applications in production

65%

of those projects are owned by teams outside Data Products

210B

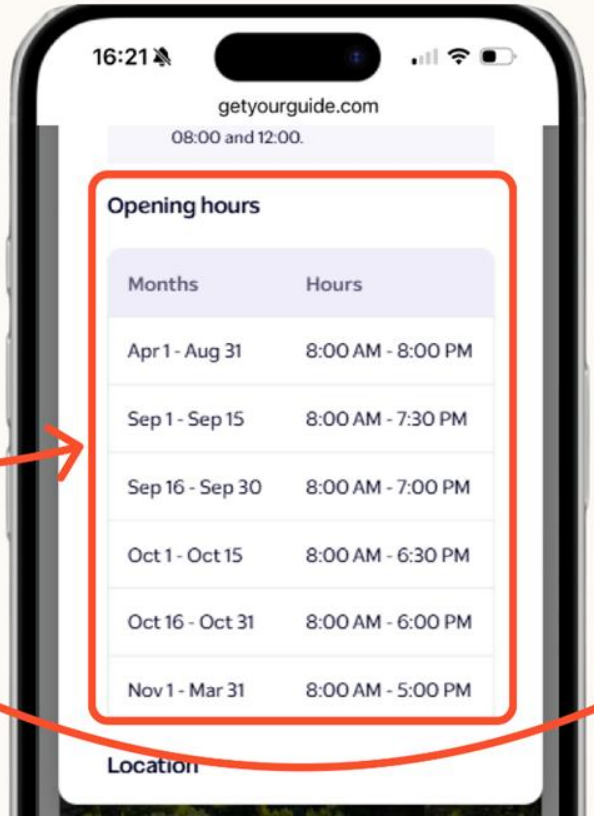
tokens for 53M requests last month

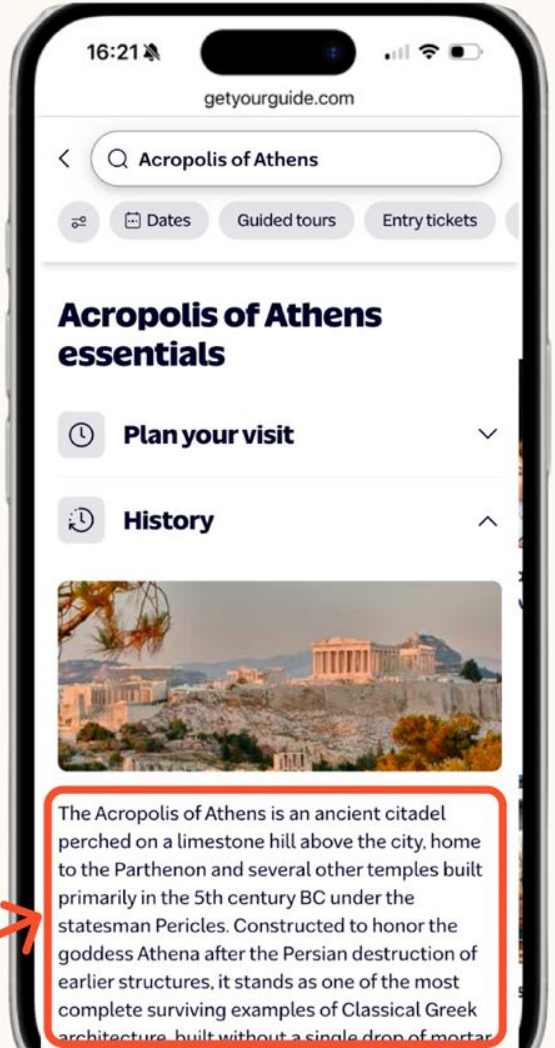
A success story



Gianluca Veschi
Mobile Engineer at
Landing pages

AI Generated







03 Learnings

Evals

Evals were our #1 AI Platform priority this year.



Evals

Evals were our #1 AI Platform priority this year.

Our job:

1) Explain why



Evals

Evals were our #1 AI Platform priority this year.

Our job:

1) Explain why

2) Explain how



What are Evals

Evals are quality tests for AI systems. They help align the output with what's expected. They make building quality AI systems measurable.

What are Evals

Evals are quality tests for AI systems. They help align the output with what's expected. They make building quality AI systems measurable.

Similar to software testing, but for LLM Output

What are Evals

Evals are quality tests for AI systems. They help align the output with what's expected. They make building quality AI systems measurable.

Similar to software testing, but for LLM Output

Being Evals first is like being TDD (Test Driven Development)

What are Evals

Evals are quality tests for AI systems. They help align the output with what's expected. They make building quality AI systems measurable.

Similar to software testing, but for LLM Output

Being Evals first is like being TDD (Test Driven Development)

Defining what good looks like

What are Evals

Evals are quality tests for AI systems. They help align the output with what's expected. They make building quality AI systems measurable.

Similar to software testing, but for LLM Output

Being Evals first is like being TDD (Test Driven Development)

Defining what good looks like

Collecting ground truth data

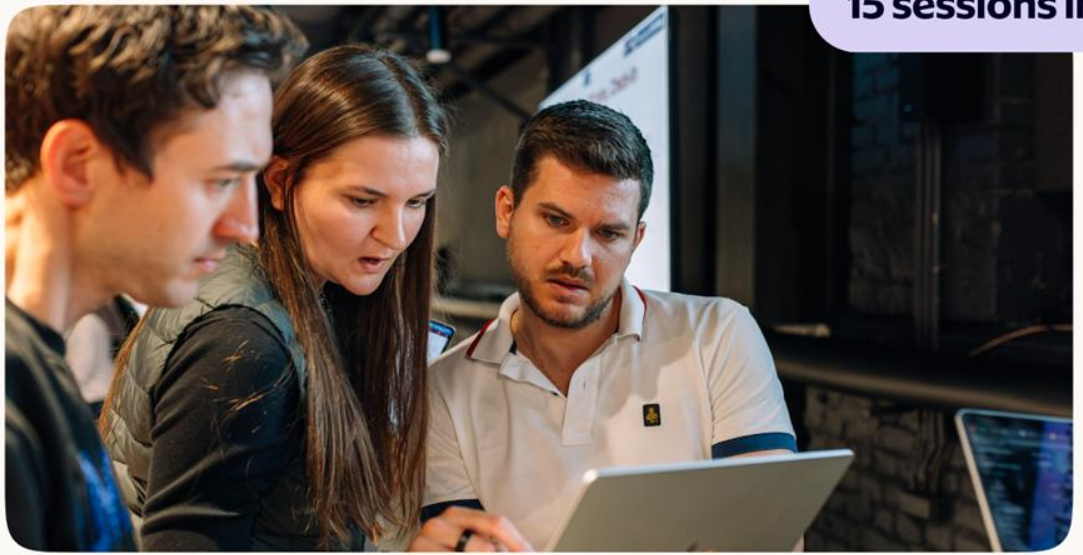
Reality check: Eval coverage



Intervention 1: Train the leaders

Intervention 1: Train the leaders

15 sessions in 3 weeks.



Intervention 1: Train the leaders

Before

Did we do a bug bash before shipping?

Intervention 1: Train the leaders

Before

Did we do a bug bash before shipping?

After

“What Evals do we have in place? 80% accurate? Ok, how did we measure that? How was the dataset built?”

Intervention 2: Establish a minimum bar

A clear set of minimum requirements each project has to meet

Intervention 2: Establish a minimum bar

A clear set of minimum requirements each project has to meet

It became clear what teams had to accomplish

Intervention 2: Establish a minimum bar

A clear set of minimum requirements each project has to meet

It became clear what teams had to accomplish

We could write a playbook on how to get to the minimum bar

Intervention 2: Establish a minimum bar

A clear set of minimum requirements each project has to meet

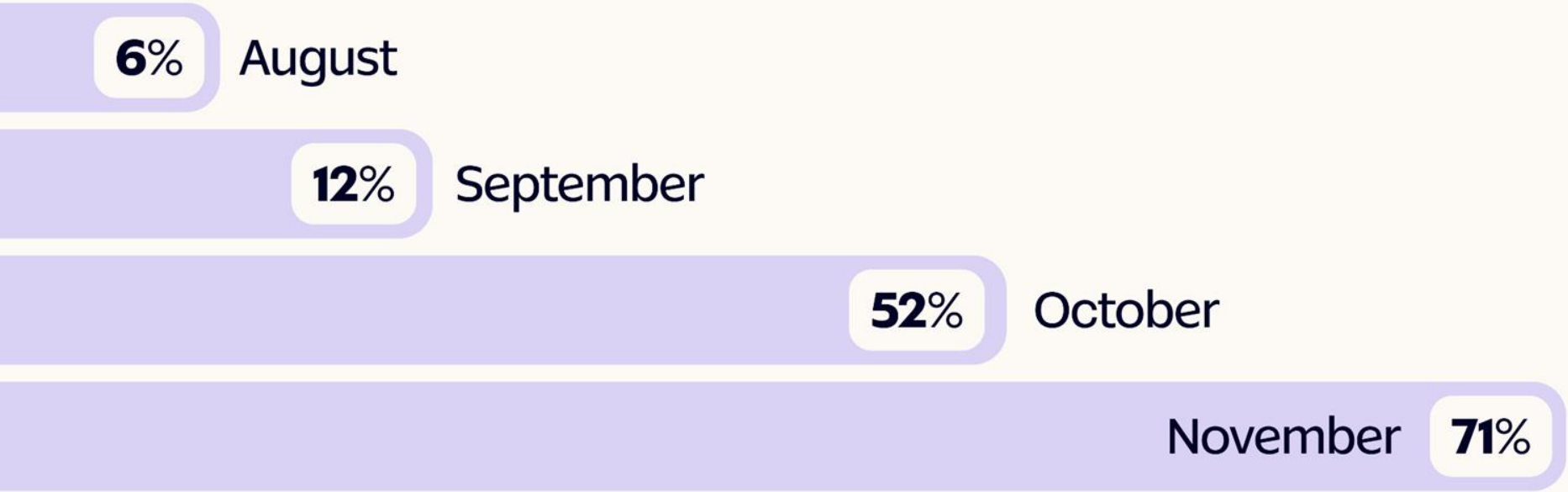
It became clear what teams had to accomplish

We could write a playbook on how to get to the minimum bar

Progress could be measured

Intervention 2: Minimum bar

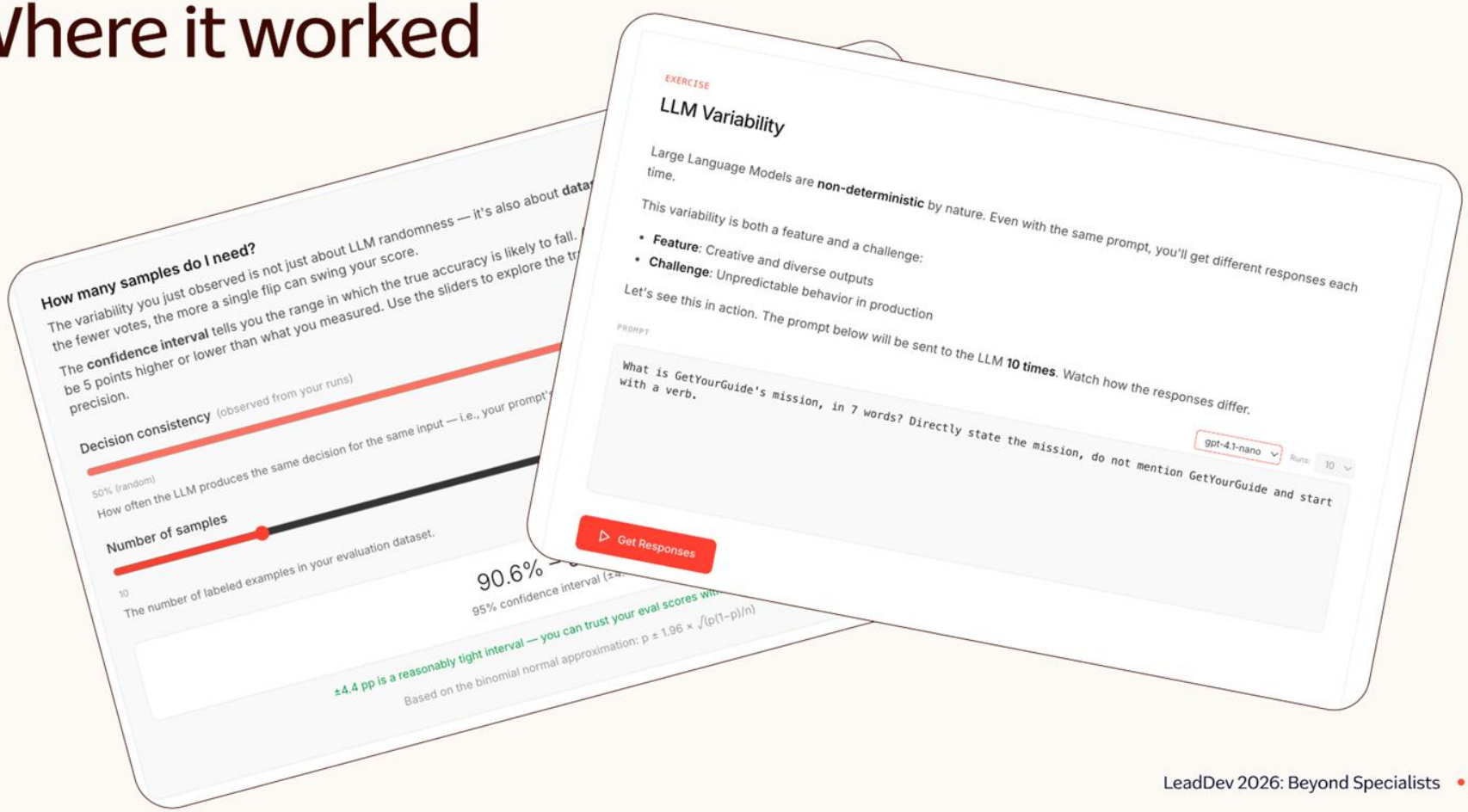
Show progress in the last few months



Use AI to teach AI

We can't be in every team, in every project.

Where it worked



Where it worked

01 **AI-native tooling**
Pick vendor with a great MCP server, CLI and skills.

Where it worked

01 **AI-native tooling**
Pick vendor with a great MCP server, CLI and skills.

02 **Platform-owned skills**
Best practices and playbook encoded centrally, shared via our internal Claude marketplace.

Where it didn't

01 **Project had Evals, but not the right ones**

Unhealthy shortcuts:

- AI-generated eval datasets leading to biases
- Prompts stuffed with examples leading to poor generalization

Where it didn't

01 **Project had Evals, but not the right ones**

Unhealthy shortcuts:

- AI-generated eval datasets leading to biases
- Prompts stuffed with examples leading to poor generalization

02 **Gap in understanding**

Metric, accuracy, baseline didn't resonate.

Learning: Evals need data science skills



04 Conclusion



Reasons to believe

- Platform established.
- Engineers with no AI background ship LLM features and build with evals.
- A clear bar moved reluctant teams.



Reasons to doubt

- Without data skills, teams take unhealthy shortcuts.
- Most LLM spend still comes from Data Products.
- We don't expect this to change soon.


TAKEAWAY
To those of you shipping AI without Data Science practitioners, you can.
But you'll need some of their skills.

Thank you LeadDev Berlin.



Mathieu Bastian
Principal AI Engineer, GetYourGuide