Building Trust in LLM Products: A Production-Ready Evaluation Framework

LREC2026 – Industry day

SCIAM

Speakers

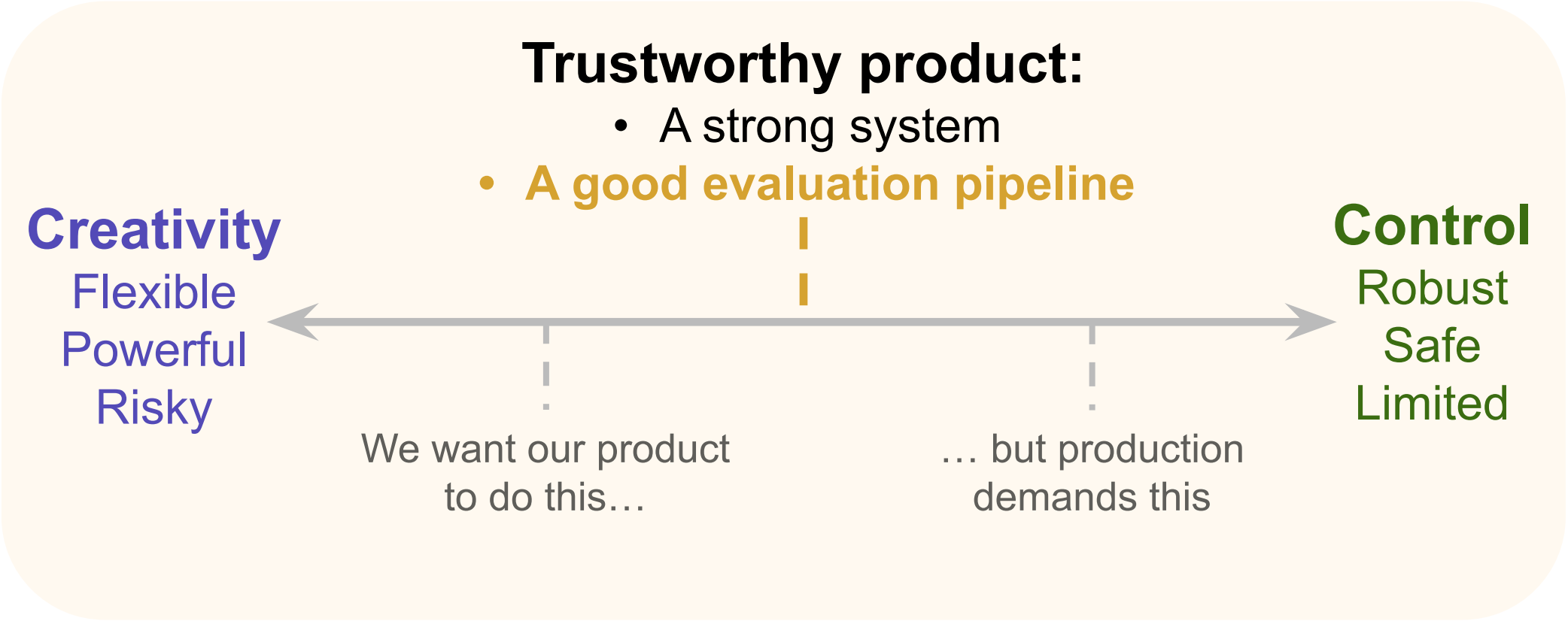
Erin Pacquetet
PhD
Senior AI Scientist

Christophe Séjourné
Head of AI & Data

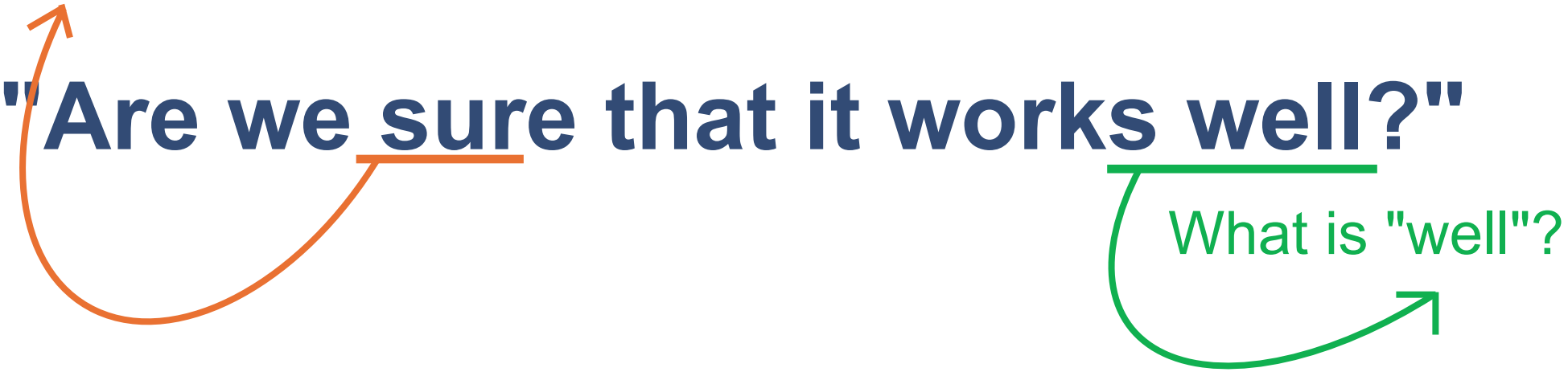
SCIAM

Evaluation is important

- It's easy to do an AI POC... and it's difficult to ship to production



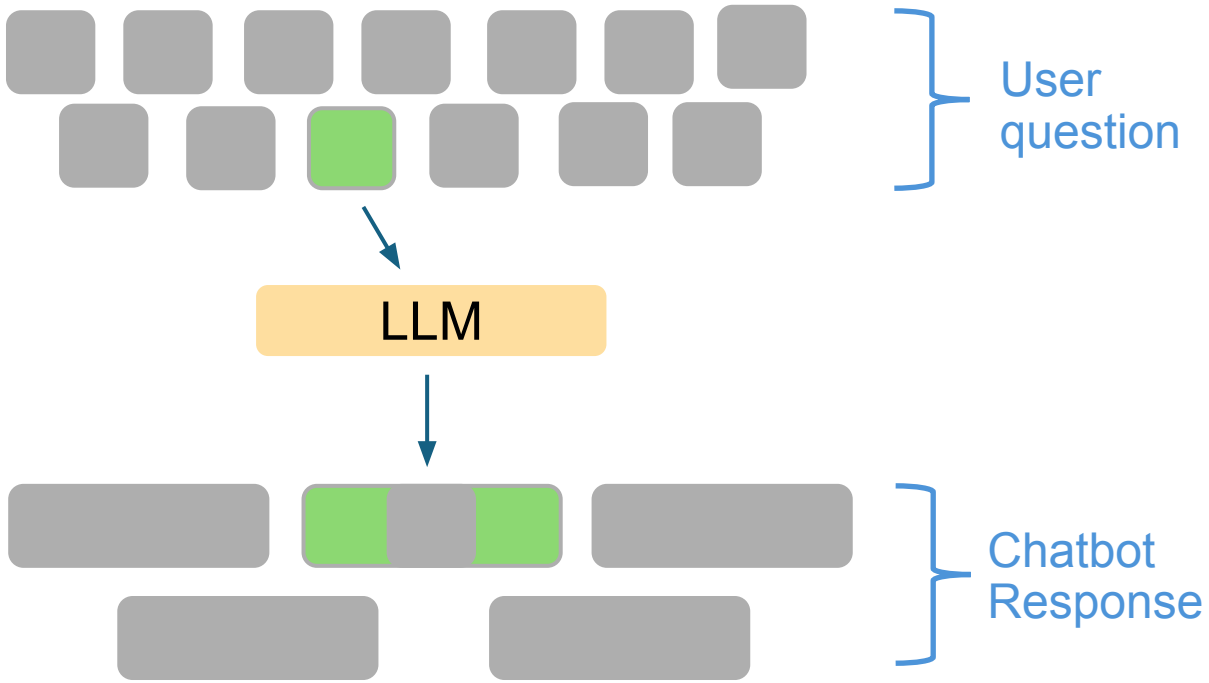
How do we trust the system?



SCIAM

LLMs in products

- The difficulty in evaluating products with LLM comes from the very reason we use them.



LLMs are used when we have:

- A wide variety of **inputs**:
  - That we don't control
  - That the system has never seen
    Potentially of poor quality

- Outputs tailored to inputs

LLMs in products

- The difficulty in evaluating products with LLM comes from the very reason we use them.

We don’t have control over the inputs.

We don’t have control over the outputs.

LLMs are used when we have:

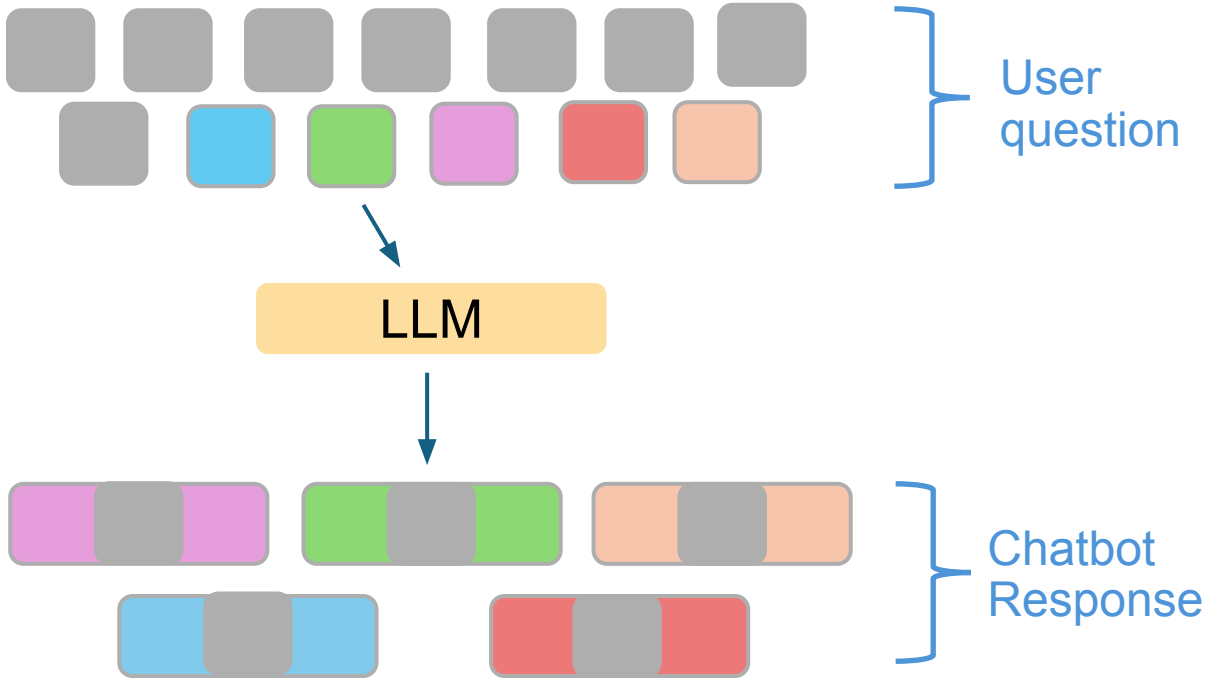
- A wide variety of **inputs**:
  - That we don't control
  - That the system has never seen
    Potentially of poor quality

- Outputs tailored to inputs

SCIAM

Evaluation pipeline

- Transform a manual qualitative assessment into a **quantitative and comprehensive measure** that can be run whenever you want, independently.



Evaluation :

1. *Simulate inputs*
2. *Generate outputs*
3. *Evaluate outputs*

When:

- *Before shipping to prod*
- *During the development phase*

SCIAM

Why evaluation is hard

① **Identify usage scenarios**

② **Judge product behavior**

Why evaluation is hard

① **Identify usage scenarios**

Identifying usage scenarios is hard because…

- LLMs accept flexible, uncontrolled inputs; the input space is **effectively infinite**

→ *Address by **decomposing the system** into testable components with well-scoped input boundaries*

Why evaluation is hard

② **Judge product behavior**

Judging behavior is hard because…

- LLMs produce open-ended, generative outputs; correctness is **rarely binary**
- **Hallucinations** can occur and are hard to detect

→ *Address by combining heuristics, model-based graders, and human review tiered by stakes and severity*

What is in our poster :

WHY Trust is a **production prerequisite**: PoCs are easy, production is hard

WHAT **Decompose into components** with bounded input spaces, then cover ideal, realistic, and adversarial usage scenarios

HOW **Score automatically**, calibrated against human judgment : scalable & repeatable

ROADMAP A **practical pipeline** built jointly by product & dev teams, never complete, always improving