Éviter le cimetière des POCs :

Architecturer vos apps IA pour la production

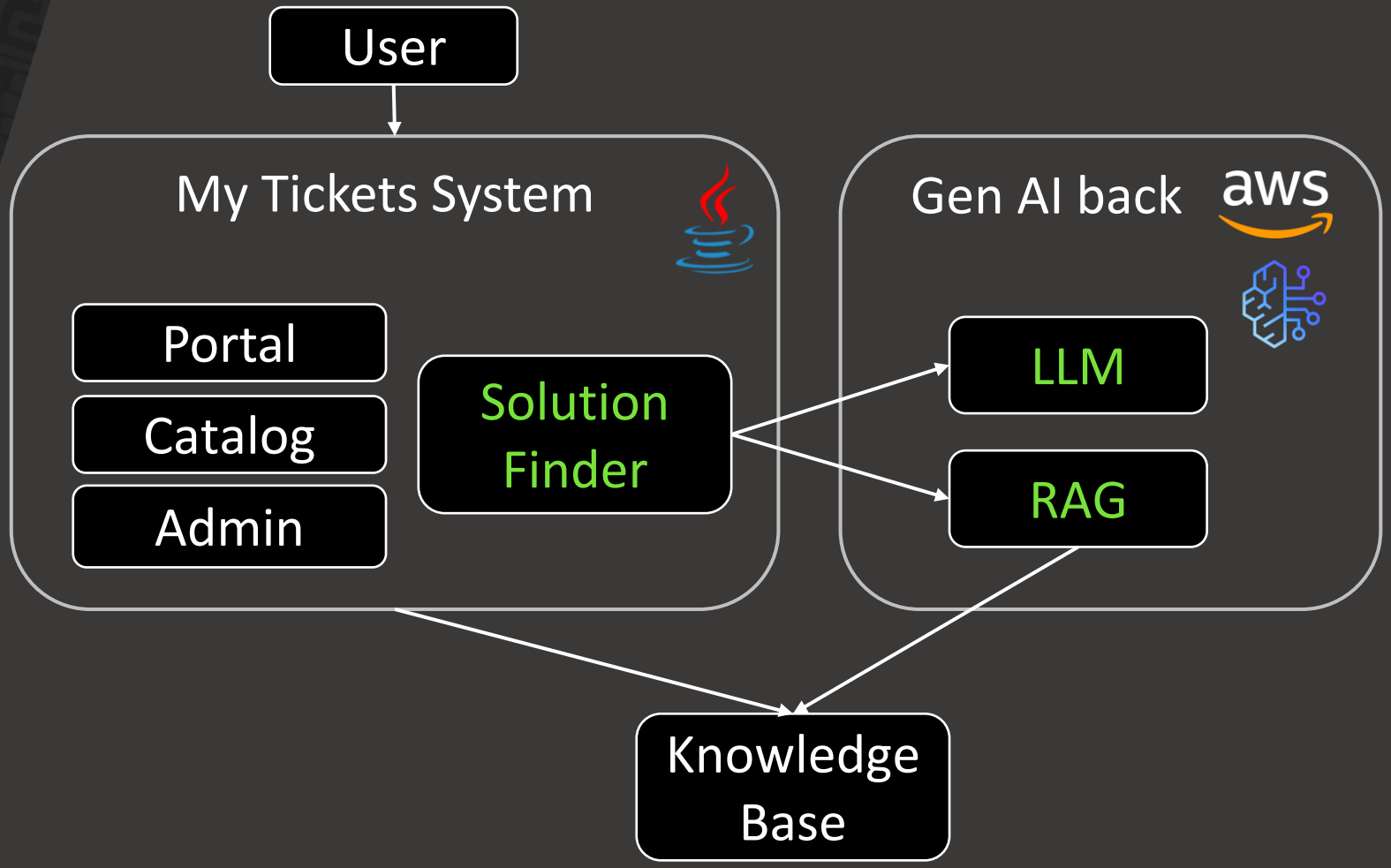
Jérôme Gauthier

DevoxX France 2025

1001 RAISONS D'ÉCHOUER

SOLUTION FINDER

Super POC
parti en prod
trop tôt 🕊️



#11 01.01.2025

DEVOXX NEWS

Guerre commerciale : La France interdit l'usage des outils IA américains

Lorem ipsum dolor sit amet, adipiscing elit. Phasellus cursus et enim ac scelerisque. Nam in urna sed orci gravida dapibus. Sed sed diam

Lorem ipsum dolor sit amet, adipiscing elit. Phasellus cursus et enim ac scelerisque. Nam in urna sed orci gravida dapibus. Sed sed diam

Lorem ipsum dolor sit amet, adipiscing elit. Phasellus cursus et enim ac scelerisque. Nam in urna sed orci gravida dapibus. Sed sed diam

Quisque posuere feugiat ullamcorper. Sed ornare nibh eu enim posuere, gravida cursus augue efficitur. Curabitur ut libero nunc. Cras mollis mi dolor. Proin eleifend finibus porttitor. Cras a lorem nec lectus imperdiet laoreet ut vitae nisl. Aliquam accumsan erat vel commodo egestas. Praesent vitae faucibus nisi.

Interdum et malesuada fames ac ante ipsum primis in faucibus. In aliquam mauris sed risus fermentum porta. Curabitur quis dolor quis risus accumsan dictum. Nullam efficitur tellus eu est congue maximus. Donec vitae urna sodales enim venenatis vestibulum. Donec gravida varius lectus id accumsan. In tortor orci, molestie quis lorem ut, euismod molestie quam.

Sed nec mauris nec urna efficitur pretium

ARCHITECTURE

ARCHITECTURE

PRINCIPES

Bonnes pratiques à conserver

- Well Architected Framework
- SOLID
- Clean Architecture

ARCHITECTURE

PRINCIPES

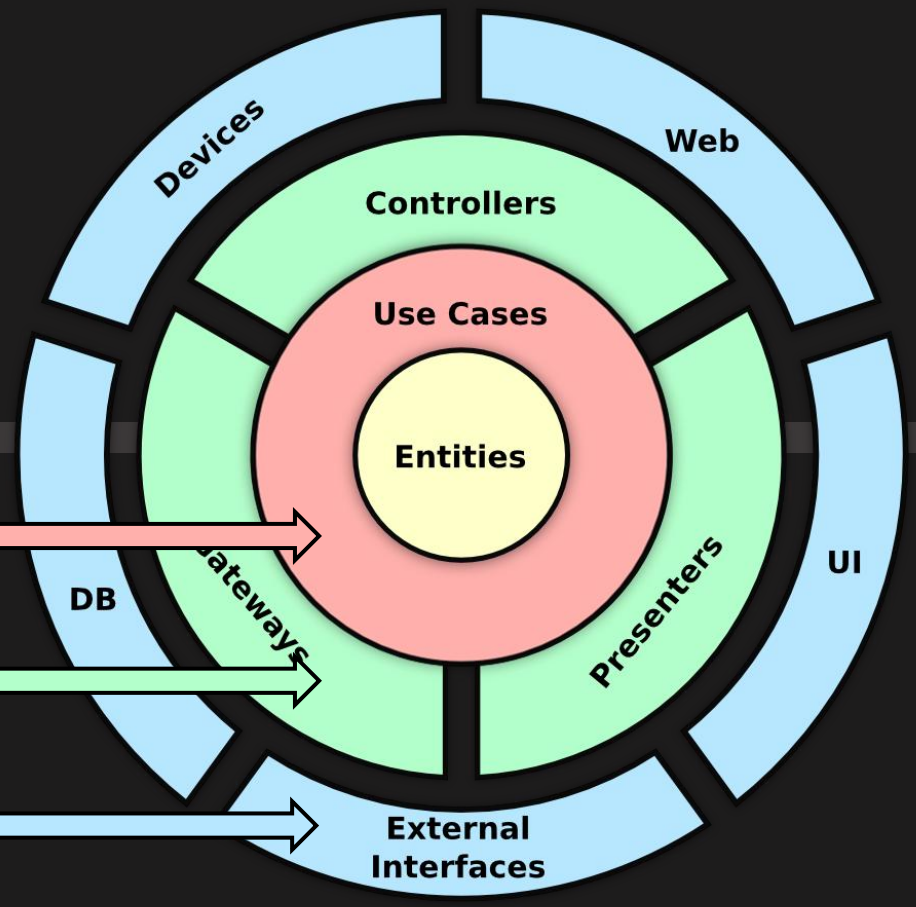
Bonnes pratiques à conserver

- Well Architected Framework
- SOLID
- Clean Architecture

Solution Finder feature

Gen AI controller

GenAI backend interface



ARCHITECTURE

FRAMEWORKS

Frameworks

- SDKs des fournisseurs de modèles
- Frameworks spécialisés LLM / Agents
- Modules stacks principales - Spring AI

	Github	Langages
LangChain	105k ⭐	
LlamaIndex	40k ⭐	
Semantic Kernel	24k ⭐	

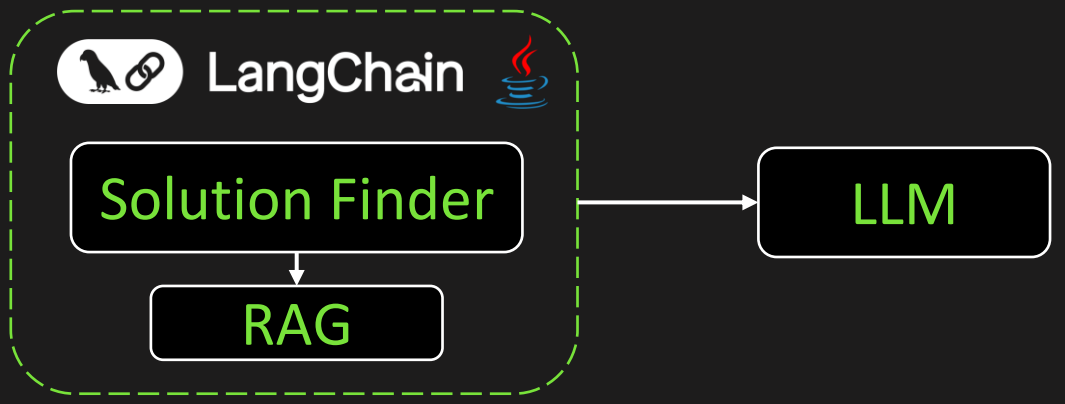
ARCHITECTURE

FRAMEWORKS

Frameworks

- SDKs des fournisseurs de modèles
- Frameworks spécialisés LLM / Agents
- Modules stacks principales - Spring AI

	Github	Langages
LangChain	105k	
LlamaIndex	40k	
Semantic Kernel	24k	



#11 01.01.2025

DEVOXX NEWS

Scandale après une fuite de données dans « Solution Finder »

Lorem ipsum dolor sit amet, adipiscing elit. Phasellus cursus et enim ac scelerisque. Nam in urna sed orci gravida dapibus. Sed sed diam

Lorem ipsum dolor sit amet, adipiscing elit. Phasellus cursus et enim ac scelerisque. Nam in urna sed orci gravida dapibus. Sed sed diam

Lorem ipsum dolor sit amet, adipiscing elit. Phasellus cursus et enim ac scelerisque. Nam in urna sed orci gravida dapibus. Sed sed diam

Quisque posuere feugiat ullamcorper. Sed ornare nibh eu enim posuere, gravida cursus augue efficitur. Curabitur ut libero nunc. Cras mollis mi dolor. Proin eleifend finibus porttitor. Cras a lorem nec lectus imperdiet laoreet ut vitae nisl. Aliquam accumsan erat vel commodo egestas. Praesent vitae faucibus nisi.

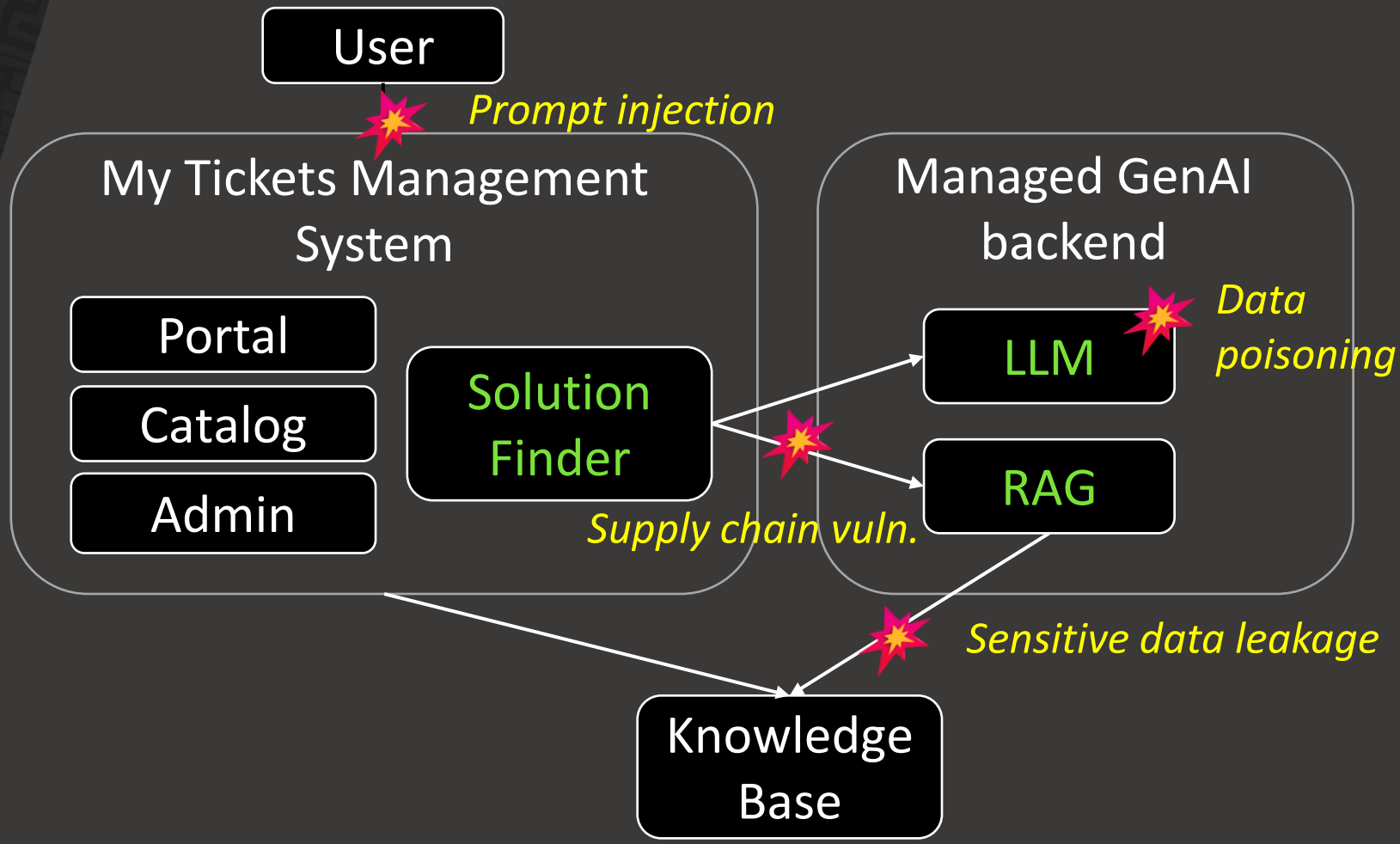
Interdum et malesuada fames ac ante ipsum primis in faucibus. In aliquam mauris sed risus fermentum porta. Curabitur quis dolor quis risus accumsan dictum. Nullam efficitur tellus eu est congue maximus. Donec vitae urna sodales enim venenatis vestibulum. Donec gravida varius lectus id accumsan. In tortor orci, molestie quis lorem ut, euismod molestie quam.

Sed nec mauris nec urna efficitur pretium

SÉCURITÉ

SÉCURITÉ

NOUVELLES SURFACES D'ATTAQUE



SÉCURITÉ

OWASP TOP 10 FOR LLM

01 : Prompt injection

02 : Sensitive info. disclosure

03 : Supply chain

04 : Data and model poisoning

05 : Improper output handling

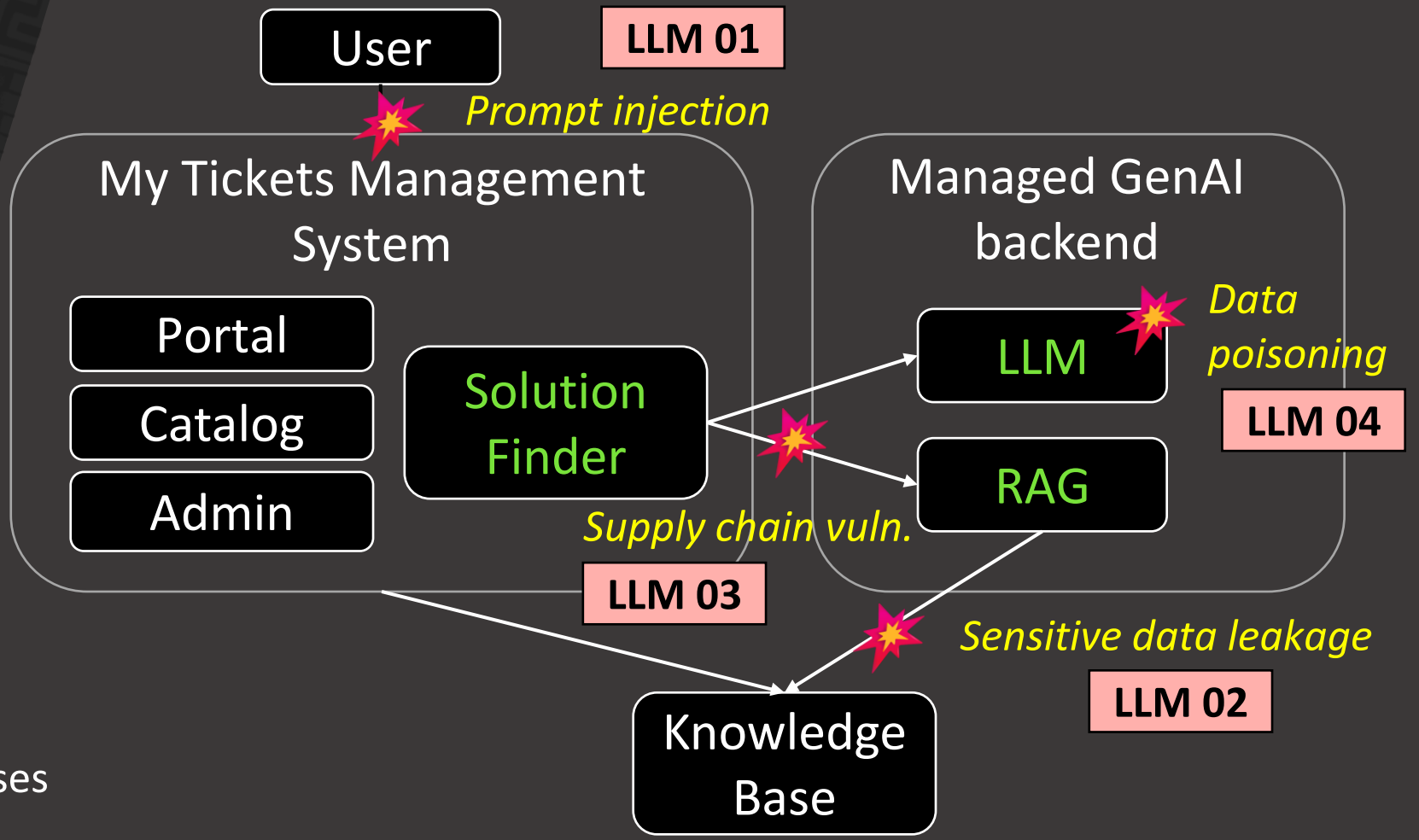
06 : Excessive agency

07 : System prompt leakage

08 : Vector & embedding weaknesses

09 : Misinformation

10 : Unbounded consumption



SÉCURITÉ

TESTER

Grille d'audit

- **OWASP** Top 10 for LLM
- 35 *recommandations de sécurité pour un système d'IA générative* de l'**ANSSI**

Outils : approche Red Team

PyRIT

garak

promptfoo

DeepEval.

SÉCURITÉ

TESTER

Grille d'audit

- **OWASP** Top 10 for LLM
- 35 *recommandations de sécurité pour un système d'IA générative* de l'**ANSSI**

Outils : approche Red Team

PyRIT

garak

promptfoo

DeepEval.

PROTÉGER #GUARDRAILS

Services managés

- **Azure** AI Content Safety
- **Amazon** Bedrock Guardrail
- **Google Cloud** Vertex AI Safety Filters

Librairies / Frameworks / APIs

- Solutions intégrées aux frameworks
- **Nvidia** NeMo Guardrails
- **OpenAI** Moderation API

#11 01.01.2025

DEVOXX NEWS

J-Corp chute en bourse après une indisponibilité de 12h de « My Tickets »

Lorem ipsum dolor sit amet, adipiscing elit. Phasellus cursus et enim ac scelerisque. Nam in urna sed orci gravida dapibus. Sed sed diam

Lorem ipsum dolor sit amet, adipiscing elit. Phasellus cursus et enim ac scelerisque. Nam in urna sed orci gravida dapibus. Sed sed diam

Lorem ipsum dolor sit amet, adipiscing elit. Phasellus cursus et enim ac scelerisque. Nam in urna sed orci gravida dapibus. Sed sed diam

Quisque posuere feugiat ullamcorper. Sed ornare nibh eu enim posuere, gravida cursus augue efficitur. Curabitur ut libero nunc. Cras mollis mi dolor. Proin eleifend finibus porttitor. Cras a lorem nec lectus imperdiet laoreet ut vitae nisl. Aliquam accumsan erat vel commodo egestas. Praesent vitae faucibus nisi.

Interdum et malesuada fames ac ante ipsum primis in faucibus. In aliquam mauris sed risus fermentum porta. Curabitur quis dolor quis risus accumsan dictum. Nullam efficitur tellus eu est congue maximus. Donec vitae urna sodales enim venenatis vestibulum. Donec gravida varius lectus id accumsan. In tortor orci, molestie quis lorem ut, euismod molestie quam.

Sed nec mauris nec urna efficitur pretium

OPS

OPS

AVANT LE DÉPLOIEMENT

Tests des outputs LLM

- ⚠️ Outputs non déterministes
- Outils dédiés ou extensions
- Human-in-the-loop

promptfoo

DeepEval.

- ✓ **Assertions** et **métriques** : contains, schémas réponses, …
- ✓ **LLM as a judge** : cohérence factuelle, pertinence, fidélité, …

OPS

AVANT LE DÉPLOIEMENT

Tests des outputs LLM

- ⚠️ Outputs non déterministes
- Outils dédiés ou extensions
- Human-in-the-loop

promptfoo DeepEval.

- ✓ **Assertions** et **métriques** : contains, schémas réponses, …
- ✓ **LLM as a judge** : cohérence factuelle, pertinence, fidélité, …

Tests de performances

- Risques principaux : latence et limites
- Couvrir les **Guardrails** et **Fallbacks**
- ⚠️ Coûts et capacités -> mocks / bouchons

APACHE JMeter™

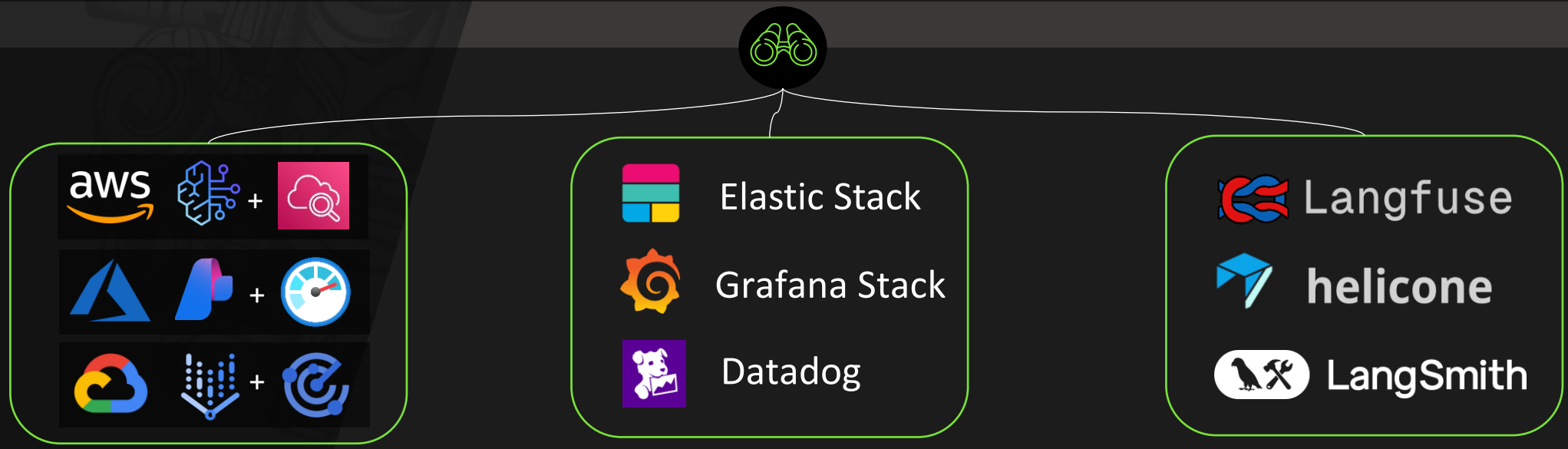
Gatling

OPS

APRÈS LE DÉPLOIEMENT

Observabilité

- **Nouvelles métriques** : Sessions, coûts, qualité, latence, erreurs, …
- 3 types de solutions :



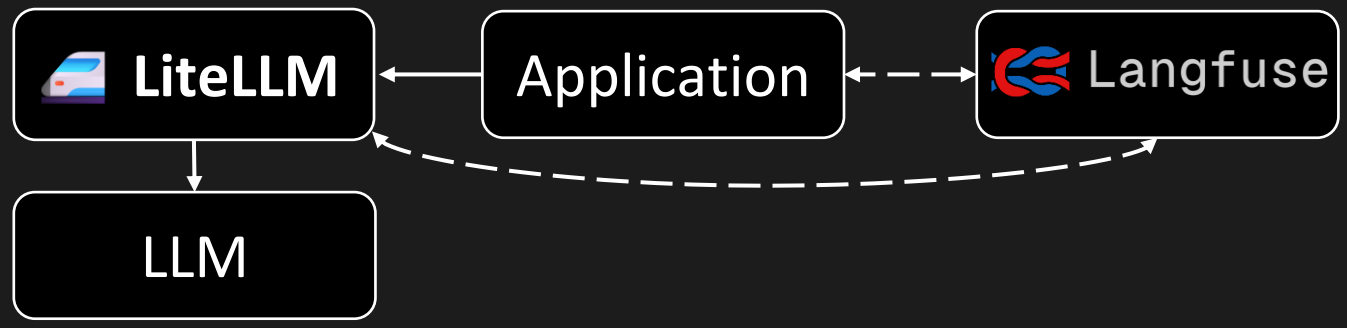
Solutions hyperscalers *Spécialistes observabilité* *Spécialistes observabilité IA*

OPS

APRÈS LE DÉPLOIEMENT

Stack LLMOps

- Cloud ou **Open Source**
- **LiteLLM** : LLM Gateway
- **Langfuse** : Observabilité

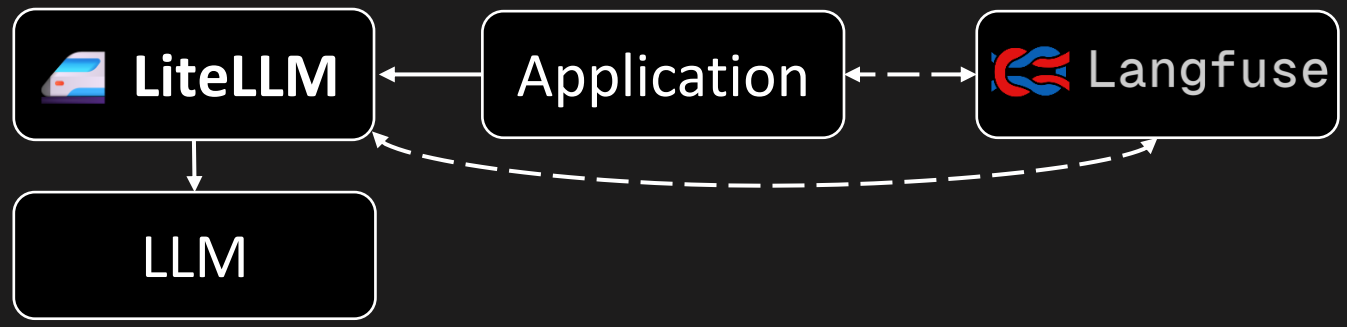


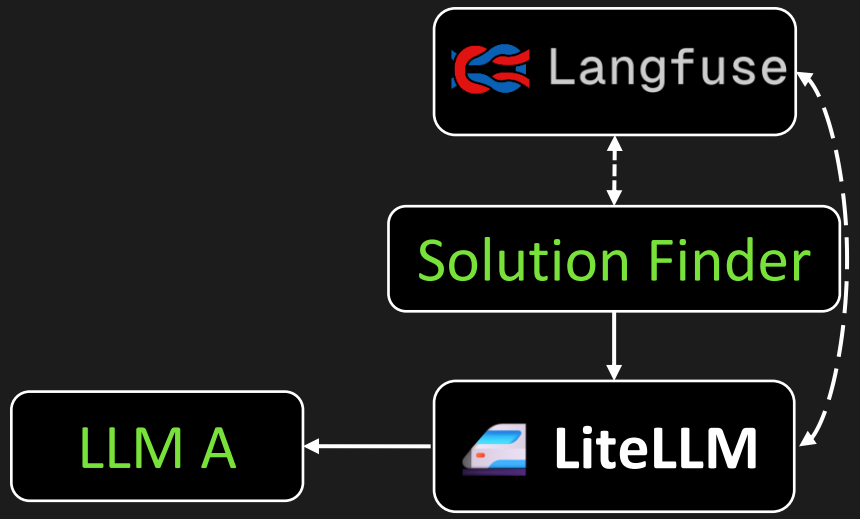
OPS

APRÈS LE DÉPLOIEMENT

Stack LLMOps

- Cloud ou **Open Source**
- **LiteLLM** : LLM Gateway
- **Langfuse** : Observabilité



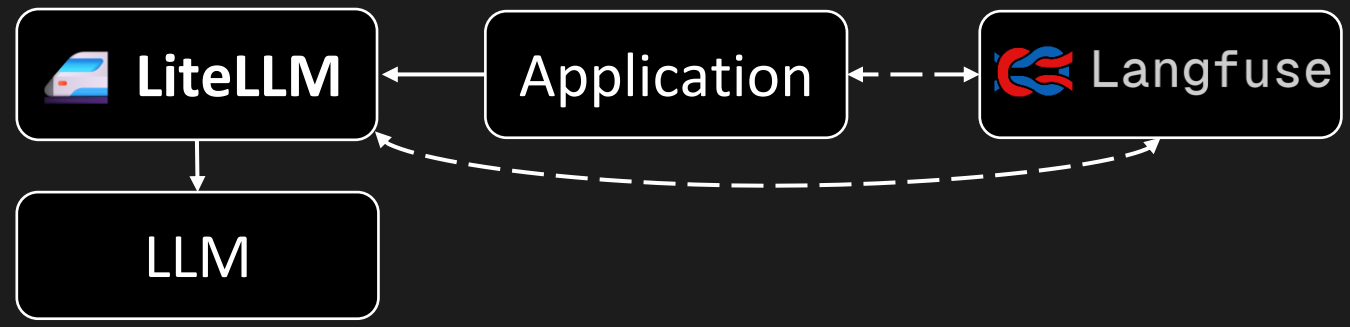


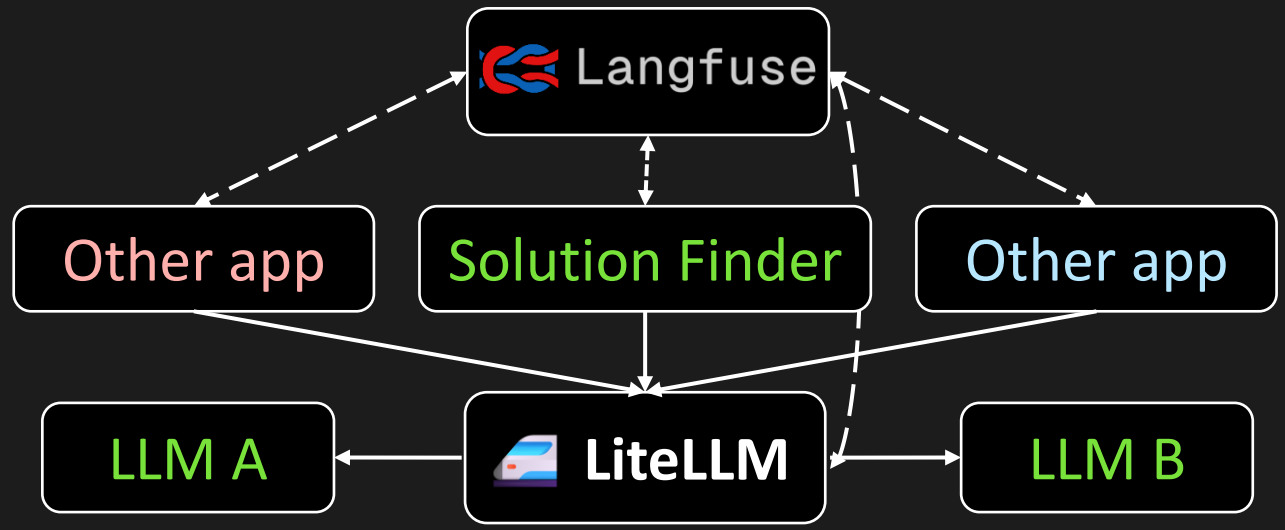
OPS

APRÈS LE DÉPLOIEMENT

Stack LLMOps

- Cloud ou **Open Source**
- **LiteLLM** : LLM Gateway
- **Langfuse** : Observabilité



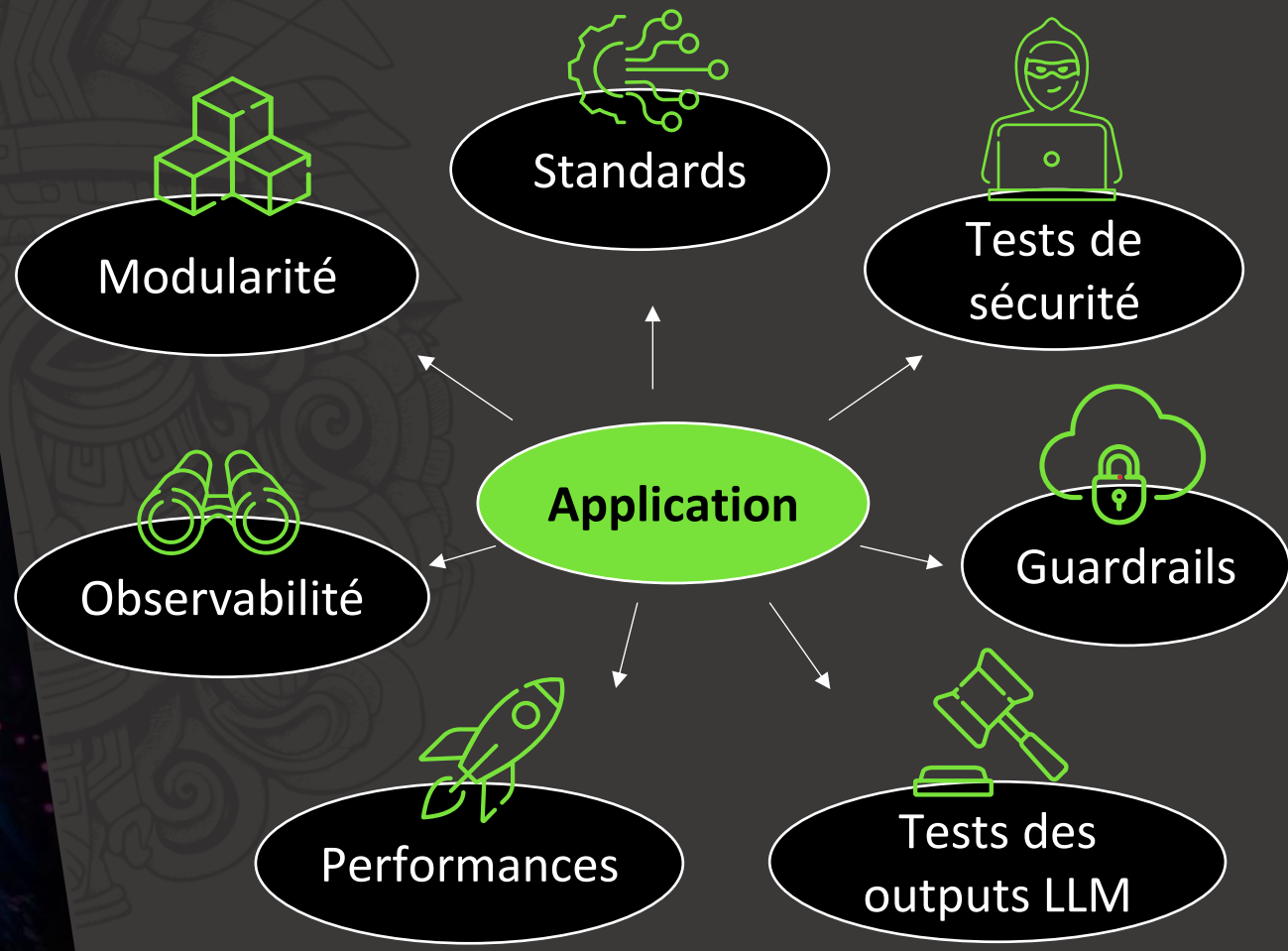


1001 AUTRES RAISONS D'ÉCHOUER

Slides

ROAD TO PROD

TAKE AWAY



Jérôme Gauthier
jerga
jerga.bsky.social