

L'IA en production

les réalités que la démo ne montre pas

Raed Chammam

Staff Software Engineer



Agentic | Generative | AI Search

Algolia - Agent Studio

Algolia

Plateforme de recherche et de découverte utilisée par des milliers d'entreprises dans le monde.

Agent Studio

Agents conversationnels IA déployés pour de vrais clients, avec de vrais volumes, en vraie production.

***Tout ce que je vais vous raconter,
on l'a appris à la dure***

Multi-fournisseur

OpenAI · Anthropic · Gemini · Amazon Bedrock · DeepSeek

Multi-région

EU & US

En production

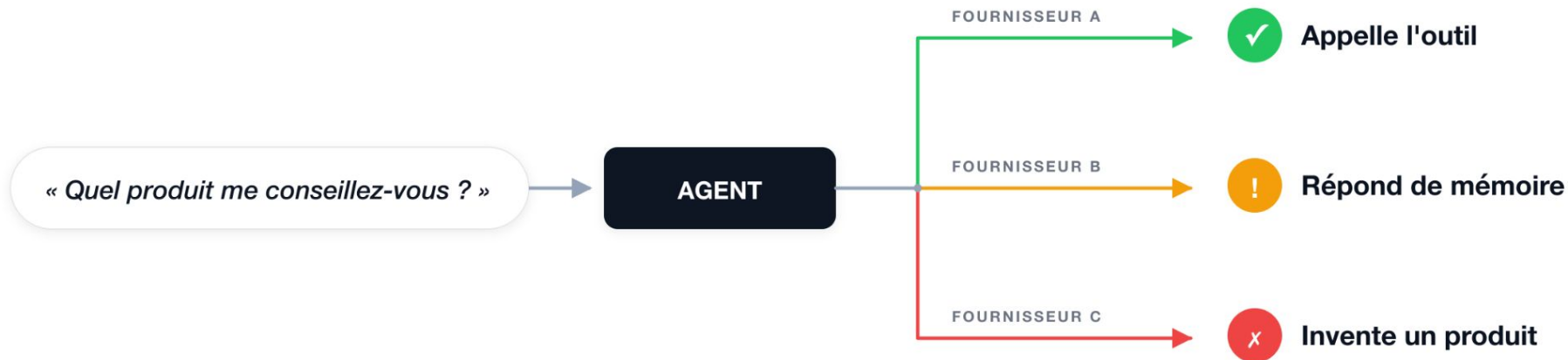
Dizaines de milliers de conversations/jour

Le non-déterminisme change tout

Même input...output différent !

Une question, trois décisions

Même agent. Même question. Trois comportements différents



✓ Nos tests d'acceptance tournent sur chaque fournisseur, à chaque déploiement.

Comment testez-vous un système qui ne répond jamais deux fois pareil ?

On mesure tout sauf la qualité

Avant

Requête → Réponse

- HTTP 200 ? **OK ✓**
- Contenu présent ? **OK ✓**

→ Suffisant

Maintenant

Conversations complexes

- Des heures, des jours
- Itérations multiples & contexte évolutif

HTTP 200 ne dit rien

Nos outils vérifient que le système répond. **Pas qu'il est utile.**

✓ On utilise des outils conçus pour tracer des conversations, pas des requêtes.

Comment mesurez-vous la qualité des réponses de votre agent ?

La panne sans erreur

Le problème

- Le système répond.
- Aucune exception, aucune alerte, aucun ticket.
- Mais les réponses sont... *moins bonnes*.
- Plus vagues, plus longues, moins ancrées dans vos données.

Pourquoi vous ne le voyez pas

- Vos dashboards mesurent la disponibilité, pas la pertinence.
- Une régression silencieuse peut durer des semaines avant qu'un humain la remarque.
- Quand un humain la remarque, c'est souvent un client.

⚠ 0,3 % des utilisateurs cliquent sur 👍 / 👎.

✓ On rejoue en continu des cas de référence et on compare les scores dans le temps.

Comment détectez-vous une dégradation quand il n'y a pas d'erreur ?

L'IA ne vous dispense pas d'ingénierie sérieuse

Les fondamentaux ne disparaissent pas parce que c'est de l'IA

Vos prompts ne sont pas du texte

C'est du code !

Le mythe

Un bloc de texte qu'on copie-colle, géré par un “*prompt engineer*”.

La réalité

Couplé à vos outils, votre modèle de données, vos règles métier.

Changez l'un, les autres cassent

L'incident

Un prompt modifié sans toucher au code.

Non-alignement = pannes

✓ **Prompts versionnés, composés dynamiquement, et déployés comme du code.**

Qui contrôle vos prompts, et comment ?

Votre produit a changé

Personne n'a rien déployé

Le problème

Mise à jour silencieuse du modèle par votre fournisseur

Sans vous prévenir, sans notes de version

La conséquence

L'agent perd ses repères et boucle

+50 appels LLM
pour une seule question !

Imaginez qu'un de vos fournisseurs logiciels change son produit, du jour au lendemain

✓ **Nos tests d'acceptance tournent sur chaque fournisseur**
C'est notre filet de sécurité quand le sol bouge

Que se passe-t-il quand votre fournisseur change de modèle sans vous prévenir ?

Combien coûte une question?

Ce qu'on imagine

Un utilisateur pose une question, le modèle répond, on paie un appel.

Ce qui se passe

L'agent raisonne, appelle des outils, délègue, boucle.

Une question peut déclencher des dizaines d'appels.

L'incident

Une mise à jour de modèle a doublé le nombre d'appels d'outils...

...du jour au lendemain !

*Vous ne laisseriez pas vos APIs s'appeler en boucle sans limite.
Pourquoi le faire avec vos agents ?*

✓ On a implémenté une suite d'outils pour configurer les limites et coûts d'exécution

Savez-vous combien coûte une question utilisateur ?

6 questions avant votre prochain projet IA

1. **Comment testez-vous un système qui ne répond jamais deux fois pareil ?**
2. **Comment mesurez-vous la qualité des réponses de votre agent ?**
3. **Comment détectez-vous une dégradation quand il n'y a pas d'erreur ?**
4. **Qui contrôle vos prompts, et comment ?**
5. **Que se passe-t-il quand votre fournisseur change de modèle sans vous prévenir ?**
6. **Savez-vous combien coûte une question utilisateur ?**

La démo est la partie facile



Raed Chamam

Staff Software Engineer



alg.li/agent-studio