DIGITAL SUMM'R

Jeudi 29 août

avec

Maxime PERRIN
mperrin@doing.fr
Ingénieur projets innovants
DOING

Intelligence artificielle

Intégration de l'IA générative dans les applications métiers : architectures, mutualisation et monitoring efficace

made with ♥ by
DIGITAL LEAGUE
Auvergne-Rhône-Alpes

Connaissez-vous ?...

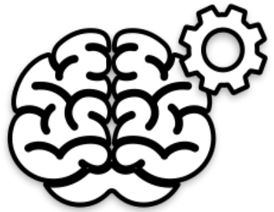
Gemini

Claude

Meta LLAMA 3

Qui suis-je ?

IIoT



IA

Maxime PERRIN
Ingénieur projets innovants
DOING

Web

Qui sommes-nous ?

DOING
GÉNÉRATEUR DE VALEUR NUMÉRIQUE

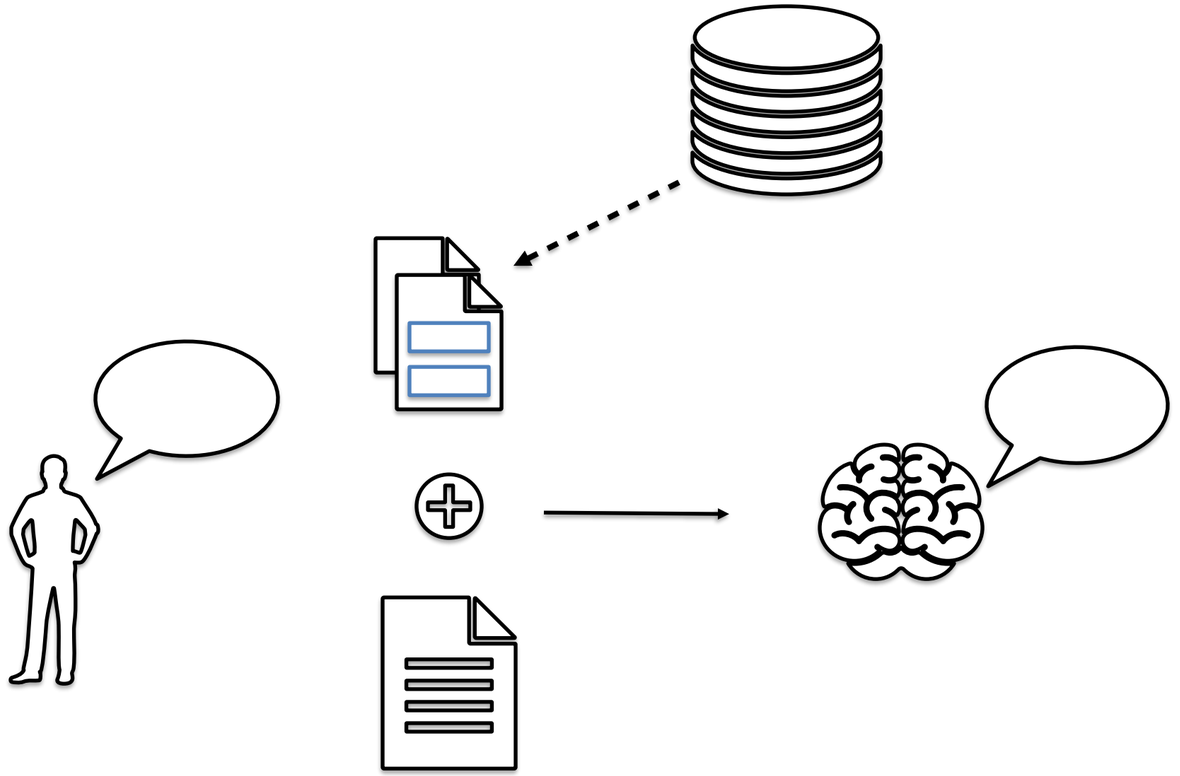
- Entrepreneur numérique sur Saint-Étienne depuis plus de 30 ans
- Une équipe d'une trentaine de personnes
- Spécialisé dans les projets innovants
  - Applications sur mesure
  - Objets connectés industriels (IIoT)

Comment intégrer de l’IA adaptée au métier facilement dans des applications existantes ?

IA contextuelle : RAG (Retrieval-Augmented Generation)

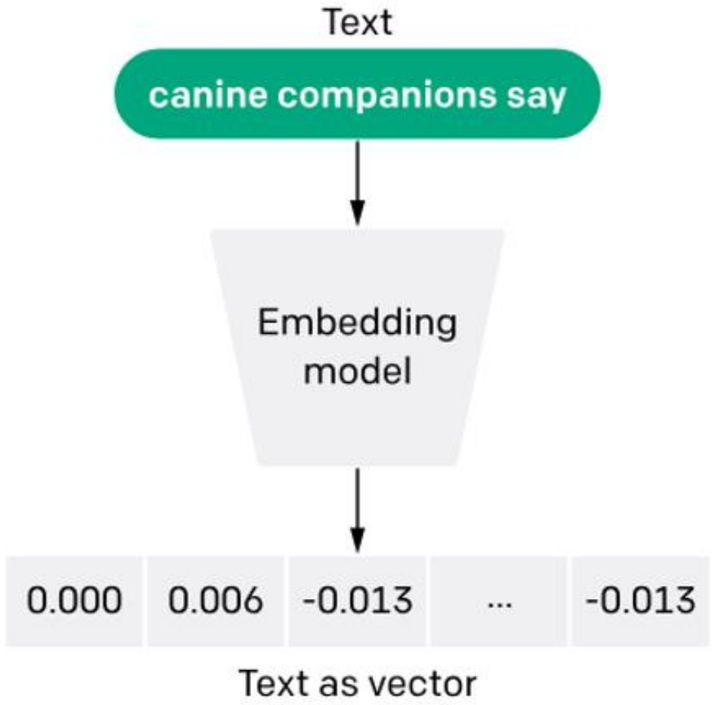
Pourquoi ?

- Avoir des réponses plus précises (données à jour, réduction des hallucinations)
- Réponses contextuelles (ajout de contexte pertinent en fonction de la requête de l'utilisateur)
- Contrôle des données



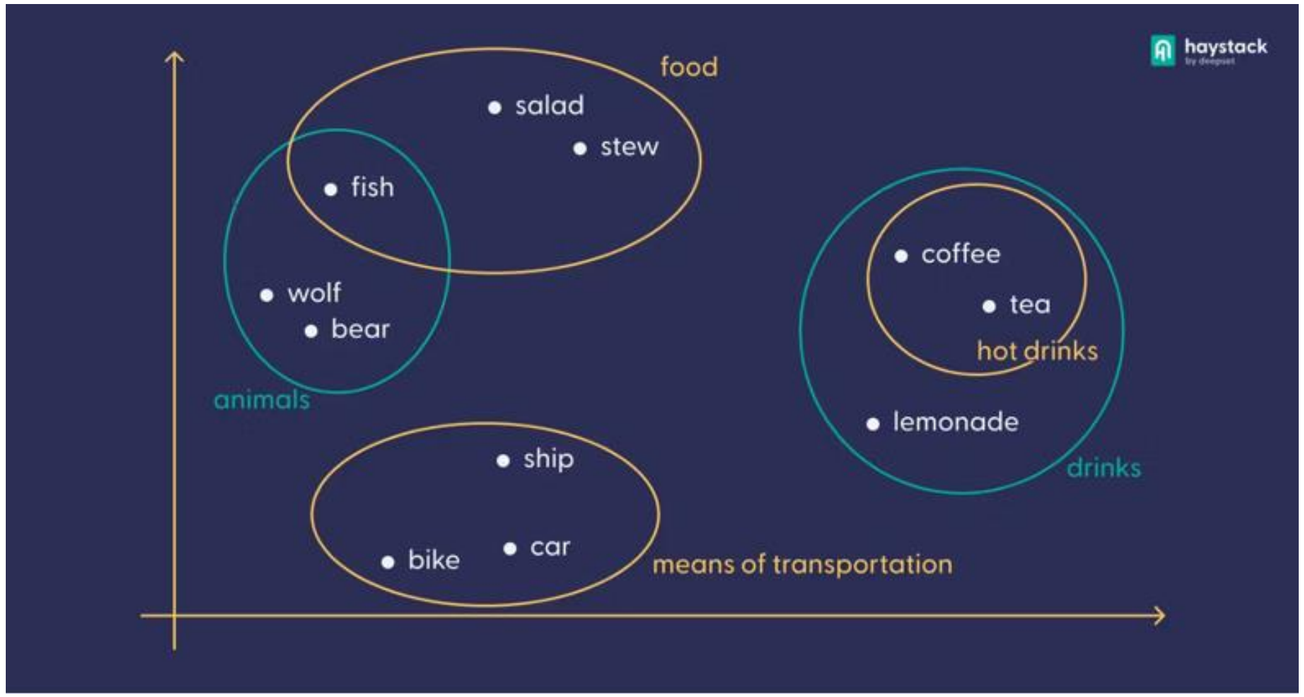
Text embeddings : Qu'est-ce que c'est ?

- Représentation sémantique de texte sous forme de vecteur
- La transformation se fait via un modèle spécialement entrainé pour *vectoriser*



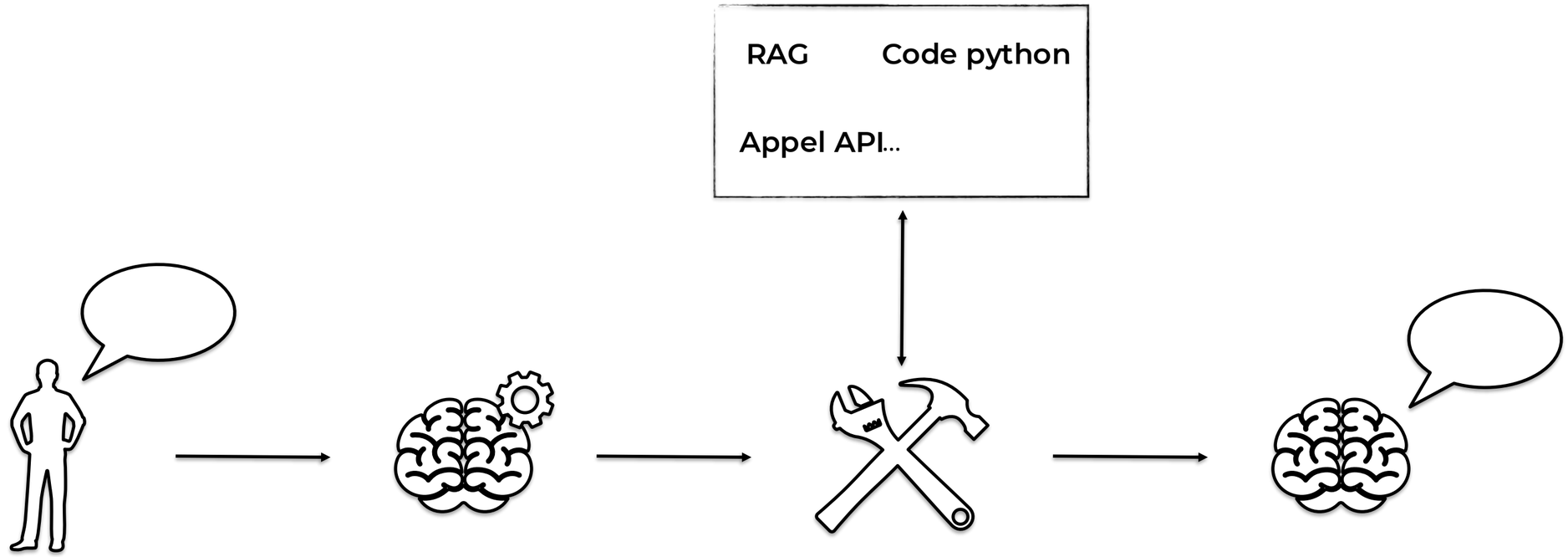
Text embeddings : Qu'est-ce que c'est ?

- Pour faire une recherche par similarité, des calculs de distance sont effectués entre les vecteurs
- Exemple : si on cherche *hot chocolate* (chocolat chaud), on le trouvera entre café et thé

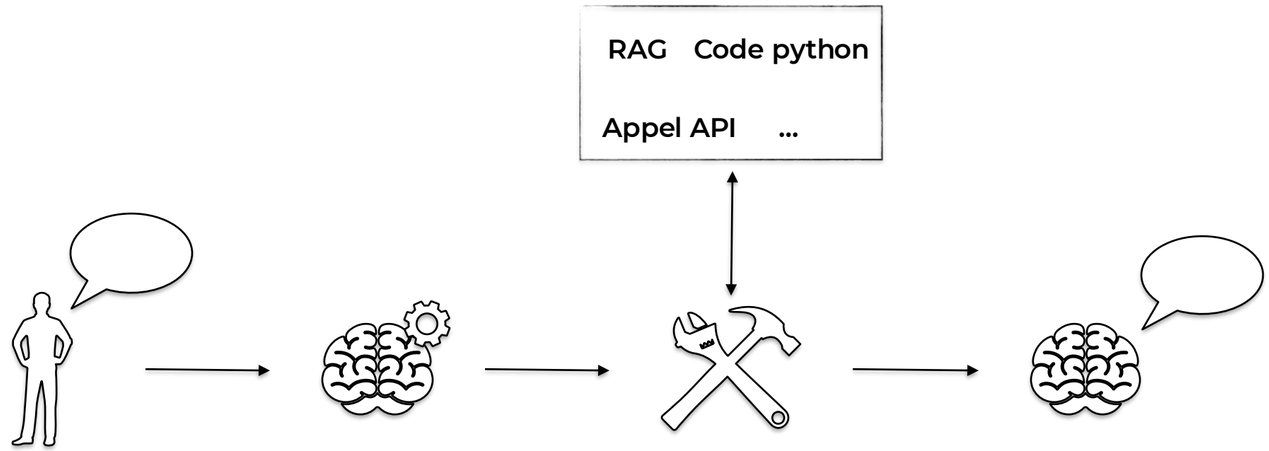


Source : haystack.deepset.ai

IA sous forme d'agent



IA sous forme d'agent



Pourquoi ?

- Donner des « mains » au LLM*
- Utilisation d’outils appropriés en fonction des situations
- Flexible et puissant

LLM: Large Language Model

Comment intégrer tout ça dans mes projets ?

Appels API directement vers le fournisseur

- Par exemple OpenAI, Anthropic, Mistral...
- Communication directe avec le fournisseur
- Pratique pour des intégrations simples et rapides
- Devient rapidement complexe

Utilisation de bibliothèques

- Création de workflows puissants et flexibles
- Pas de gestion API
- Sécurité et contrôle
- Dépendance additionnelle

Les bibliothèques

LlamaIndex

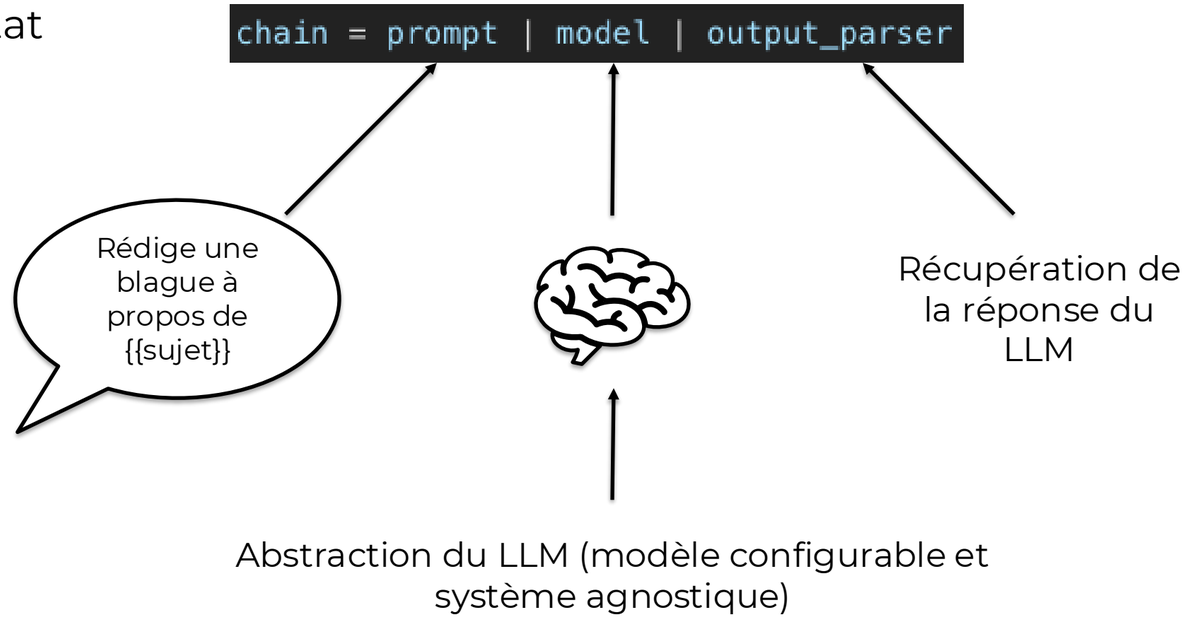
haystack
by deepset

LangChain

LangChain

LCEL (LangChain Expression Language)

- Configurer une chaîne (prompt -> LLM -> résultat par exemple)
- Accéder aux étapes intermédiaires
- Gérer des exécutions parallèles
- Support asynchrone
- Possibilité d'exposer une chaîne sous forme d'API (LangServe)

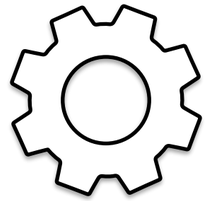


LangChain

{“sujet” : “football”}

Rédige une blague à propos de **football**

ChatMessage

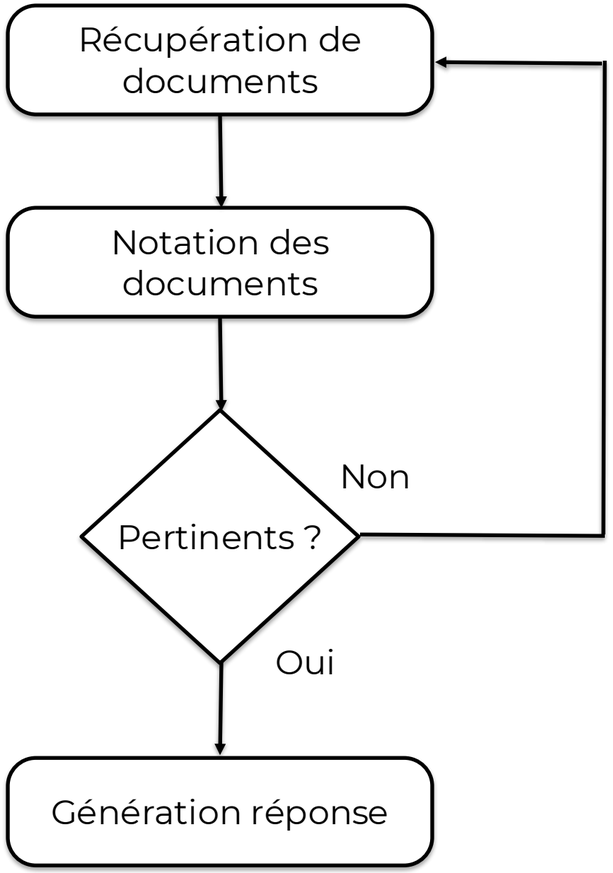


Texte

LangChain

LangGraph

- Construire des applications « stateful » et « agentique »
- Créer des cycles et des branches
- Human-in-the-loop
- Persistence de l’état du graph



Comment estimer les coûts efficacement ?

Les points importants

- Nombre de tokens* en entrée → 1000
- Nombre de tokens* en sortie → 500
- Nombre de requêtes → 100
- Performance requise (modèle requis) → GPT-4o

Exemple de calcul

Model	Pricing
gpt-4o	$5.00 / 1M input tokens
	$15.00 / 1M output tokens

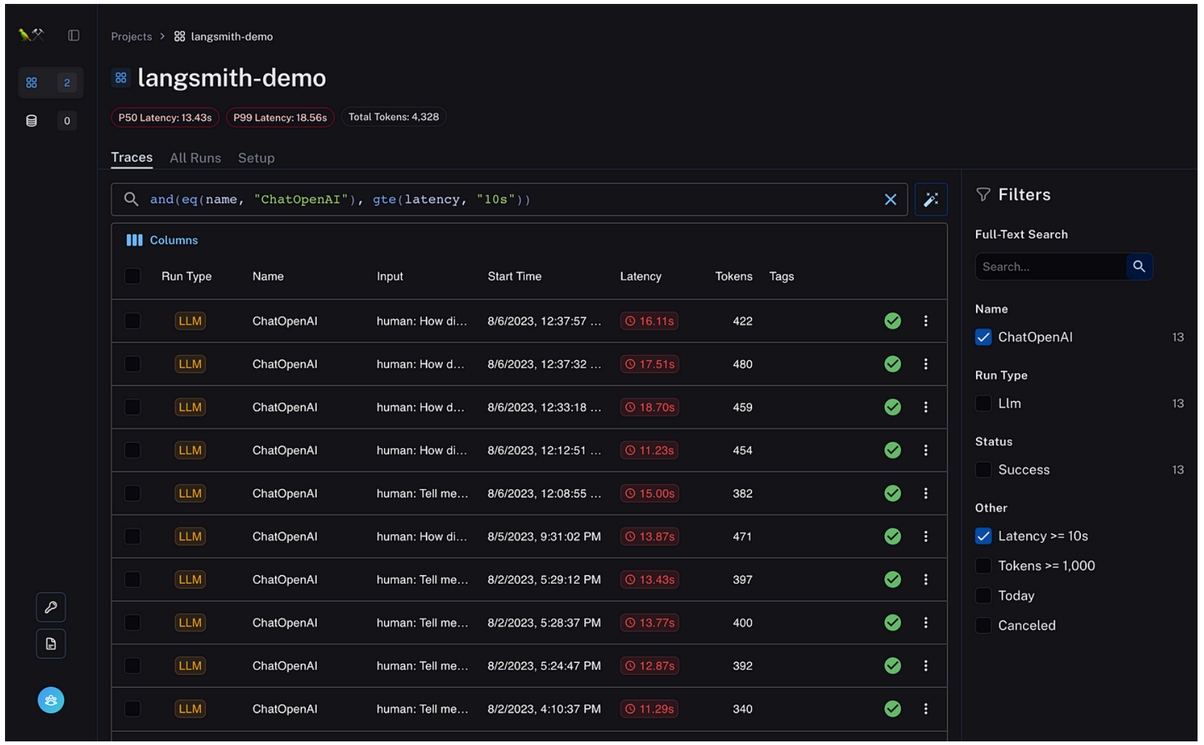
$$\left(\frac{1000 \times 5 + 500 \times 15}{10^6}\right) \times 100 = 1.25$$

*100 tokens = 75 mots

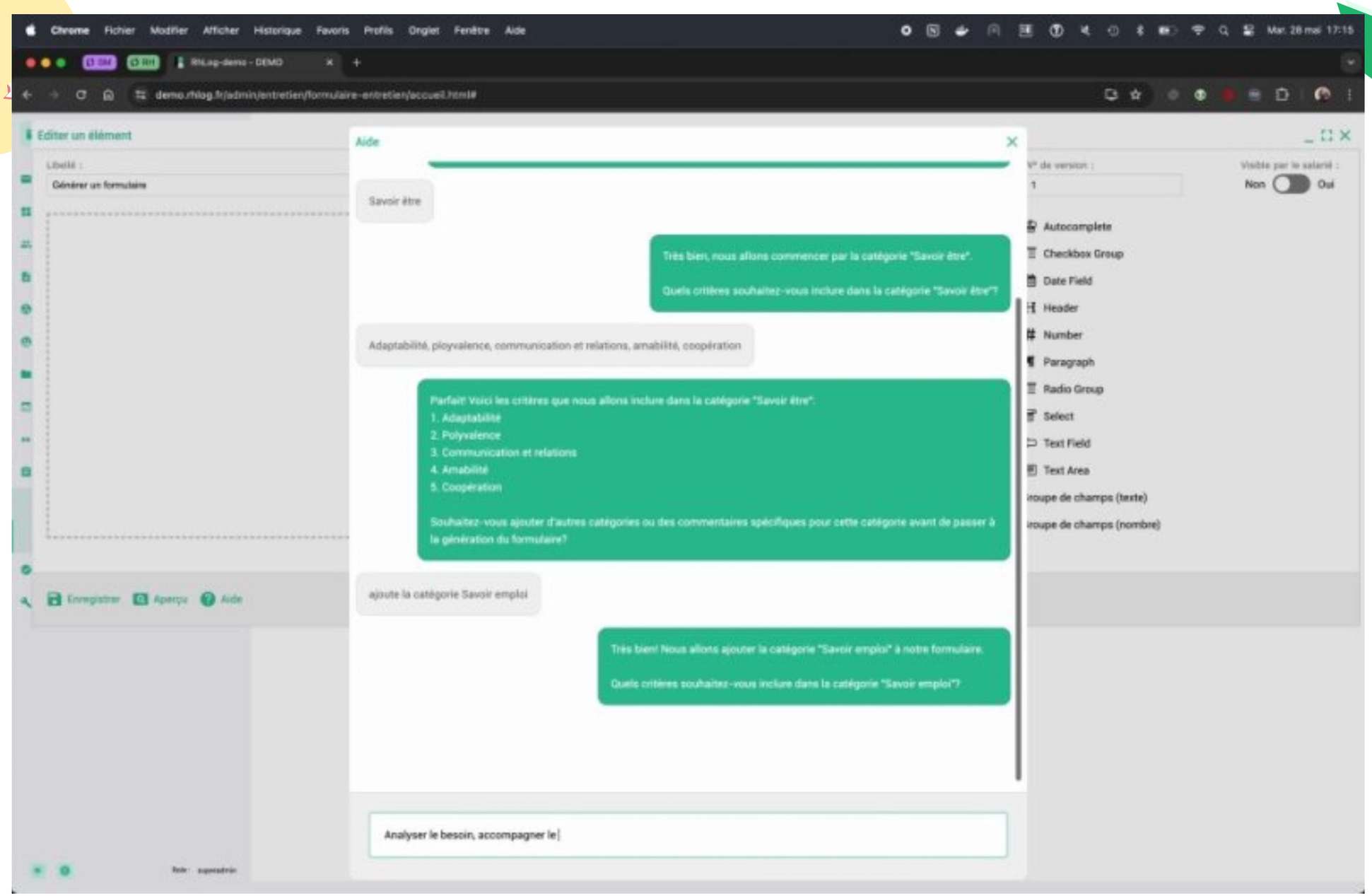
Comment superviser mon application ?

LangSmith

- Affiche les requêtes effectuées
- Intégration automatique depuis LangChain
- Estimation des coûts
- Possibilité de voir le détail des exécutions
- Évaluation de la performance
- 1000 traces/mois gratuites

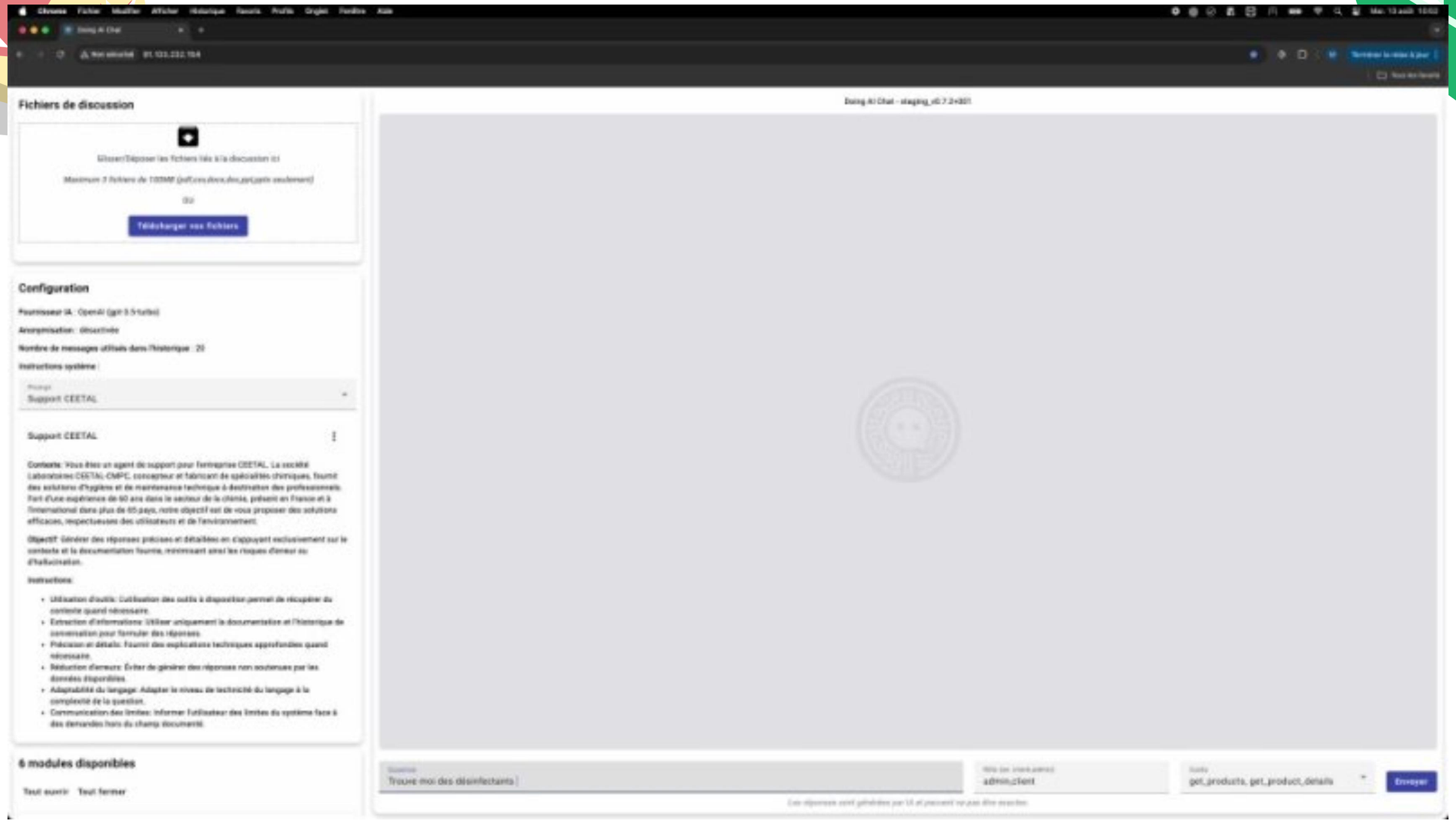


L'architecture chez DOING



DIGITAL SUMM'R

made with ♥ by DIGITAL LEAGUE Auvergne-Rhône-Alpes


Fichiers de discussion
Télécharger vos fichiers
Configuration
Support CEETAL
6 modules disponibles
Tout ouvrir
Tout fermer
Trouve moi des désinfectants
admin_client
get_products, get_product_details
Envoyer

DIGITAL SUMM'R

Edition 2024

Pour ne rien oublier ...

votre anti-sèche !

Les 5 choses à retenir du talk : Intégration de l'IA générative dans les applications métiers

1. Pérennité

Il est possible de passer d'un modèle à un autre facilement

2. Sécurité

On a le contrôle sur toutes les étapes intermédiaires, ce qui permet de ne rien laisser au hasard

3. Simplicité d'intégration

Avec LangChain et d'autres outils, il est possible de créer une API en quelques minutes. Cela permet d'intégrer de l'IA dans n'importe quelle application métier de façon indépendante

4. Contrôle des données

Les données complètes restent dans l'infrastructure locale et on contrôle de façon précise les entrées et sorties

5. Gestion des coûts

Définir des indicateurs pertinents et réfléchir à l'usage de l'IA permet d'anticiper les coûts et de faire un meilleur pricing

made with ♥ by DIGITAL LEAGUE Auvergne-Rhône-Alpes