



DIGITAL
SUMM'R

Jeudi 29 août

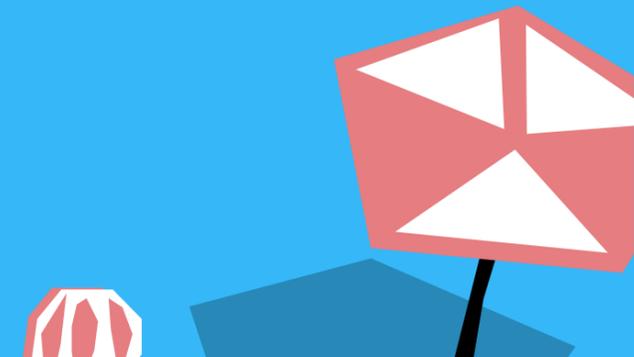
avec



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



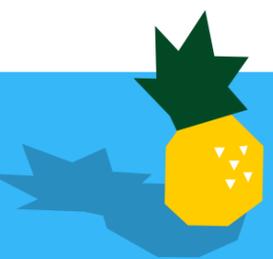


Connaissez-vous ?...

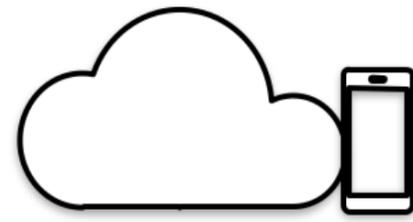


Gemini

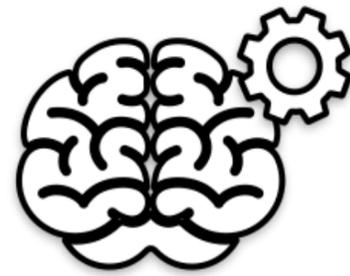
Claude



Qui suis-je ?



IIoT



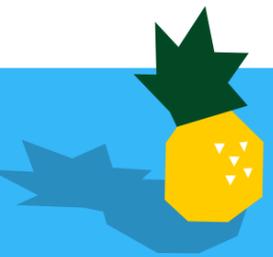
IA



Maxime PERRIN
Ingénieur projets innovants
DOING



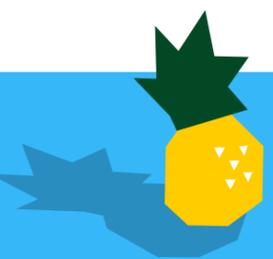
Web



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 ?**



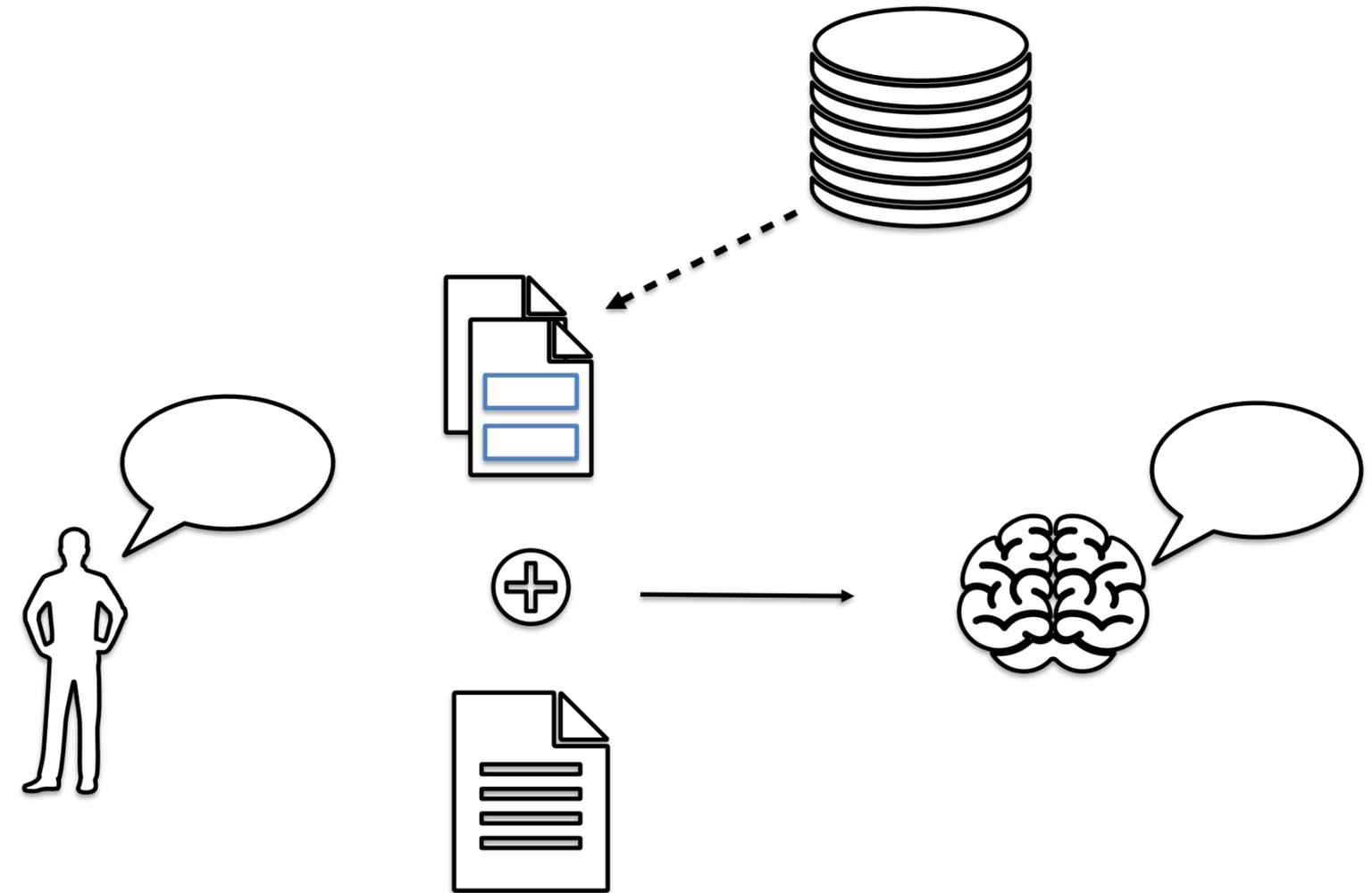
**DIGITAL
SUMM'R**

made with ♥ by
**DIGITAL
LEAGUE**
Auvergne-Rhône-Alpes

IA contextuelle : RAG (Retrieval-Augmented Generation)

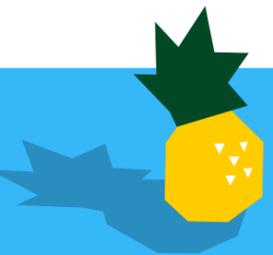
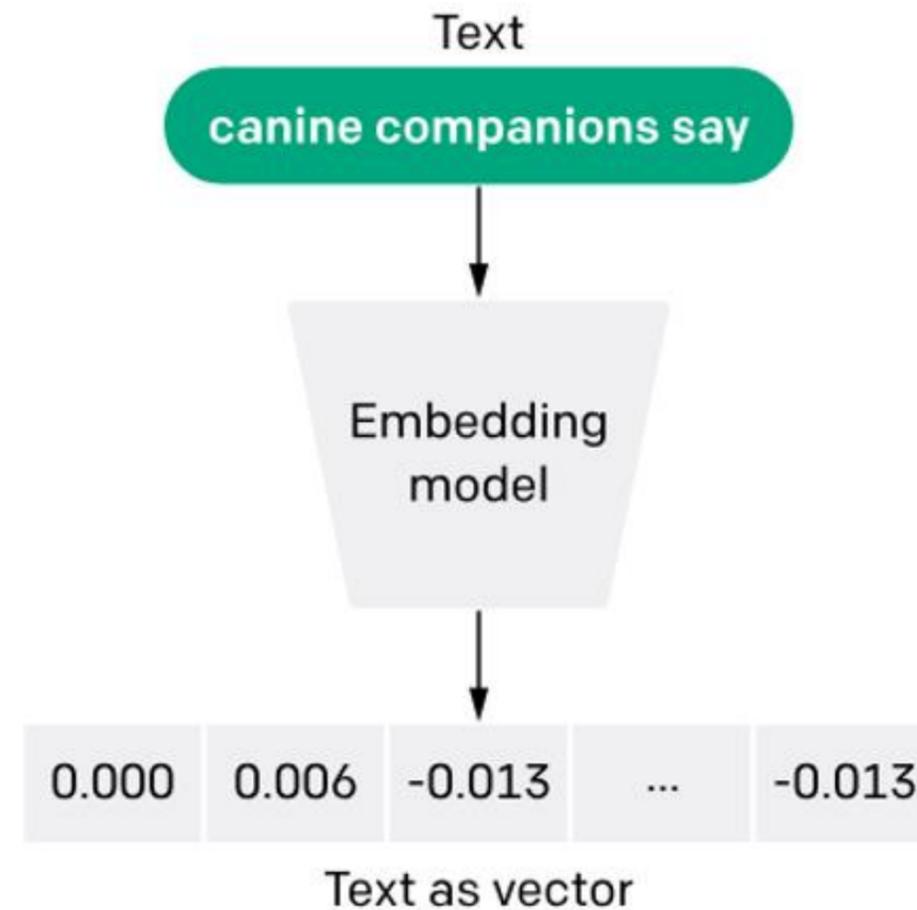
Pourquoi ?

- 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



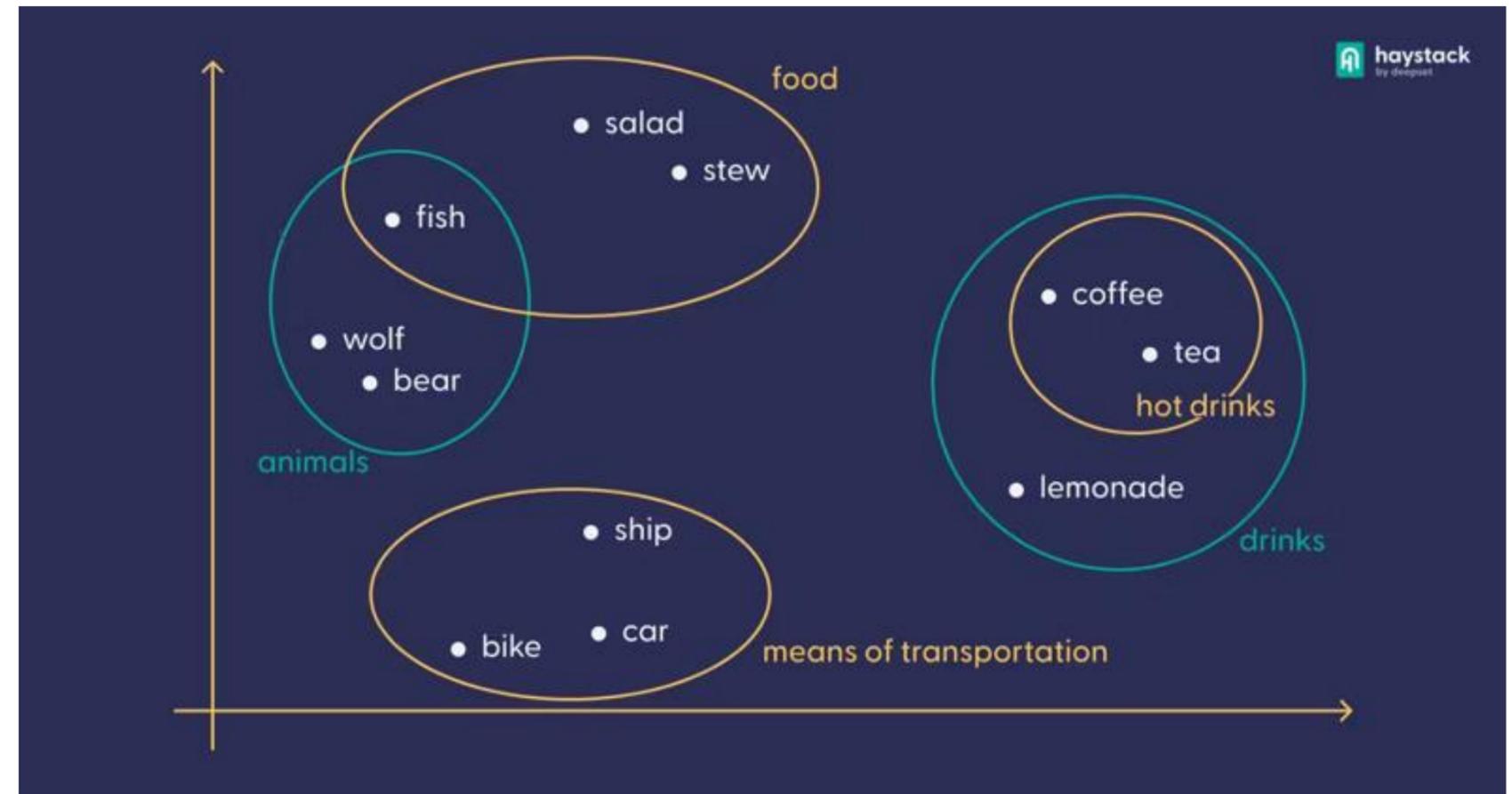
Text embeddings : Qu'est-ce que c'est ?

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

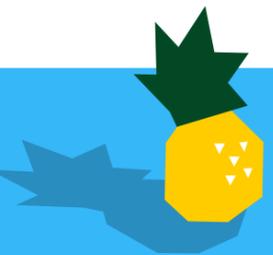


Text embeddings : Qu'est-ce que c'est ?

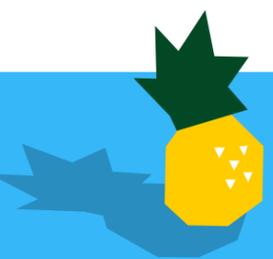
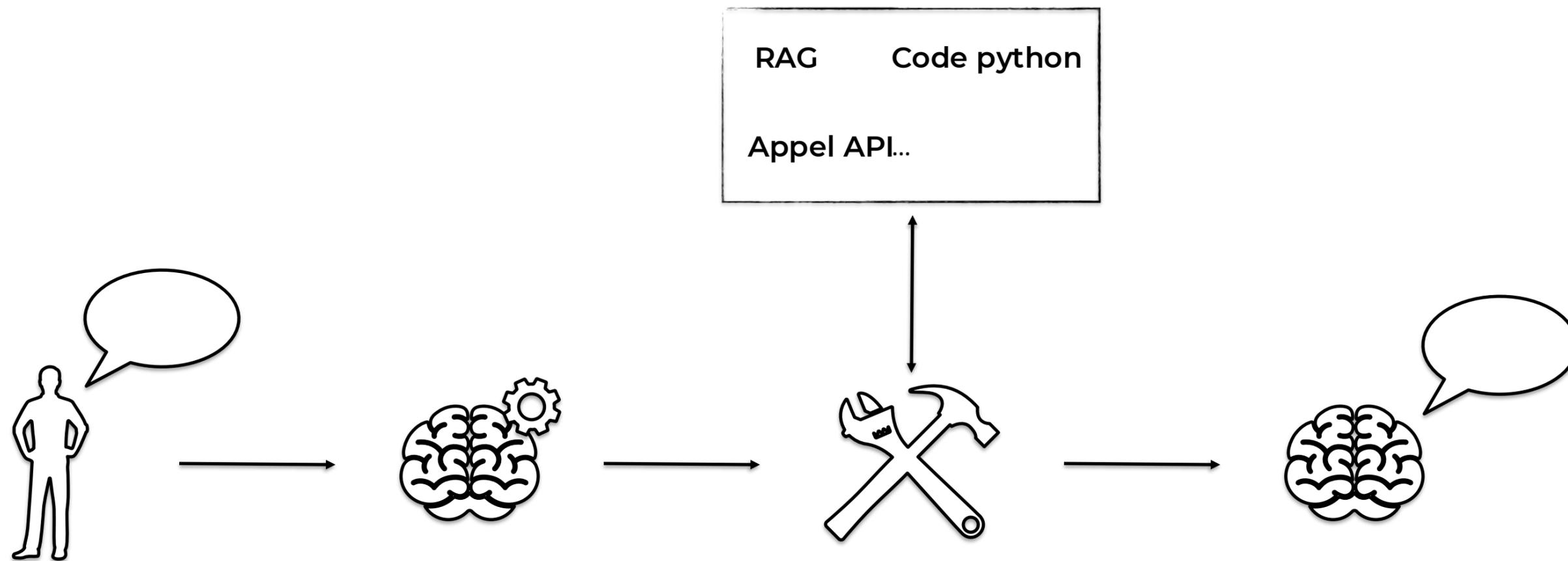
- Pour faire une recherche par similarité, des calculs de distance sont effectués entre les vecteurs
- Exemple : si on cherche *hot chocolate* (chocolat chaud), on le trouvera entre café et thé



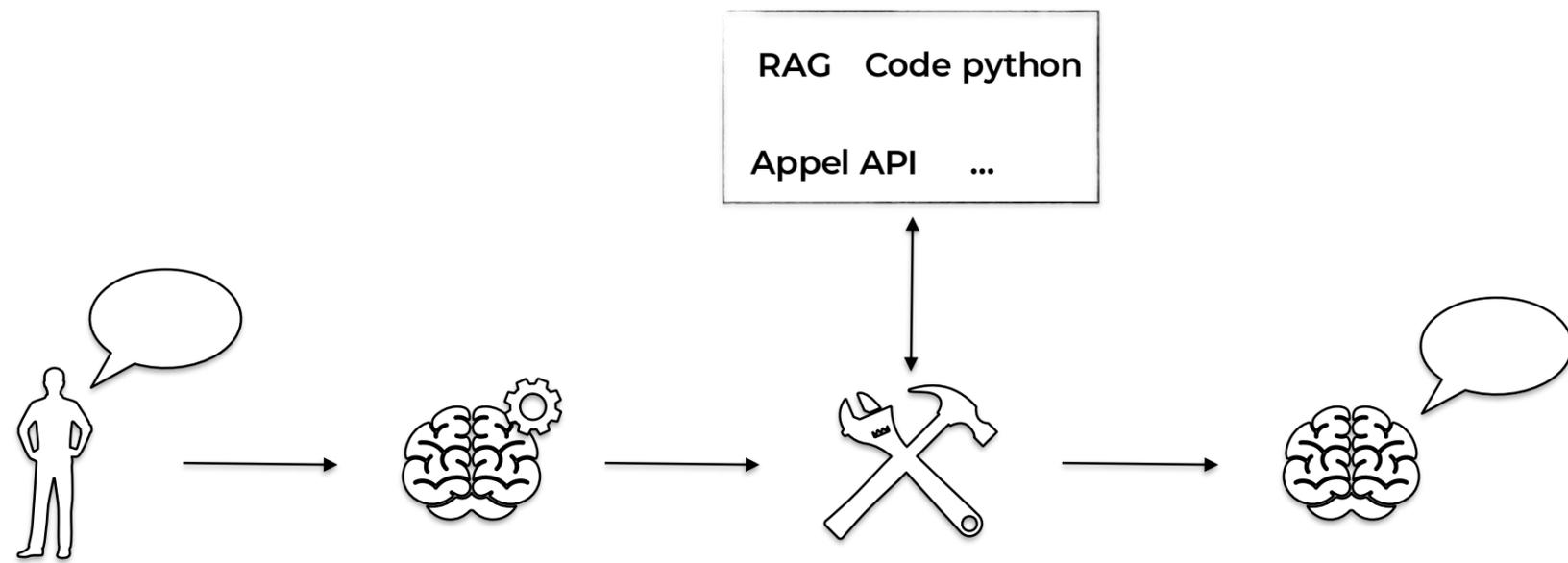
Source : haystack.deepset.ai



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



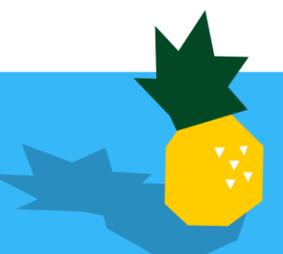
LlamaIndex



haystack
by deepset



LangChain

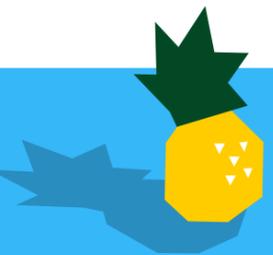
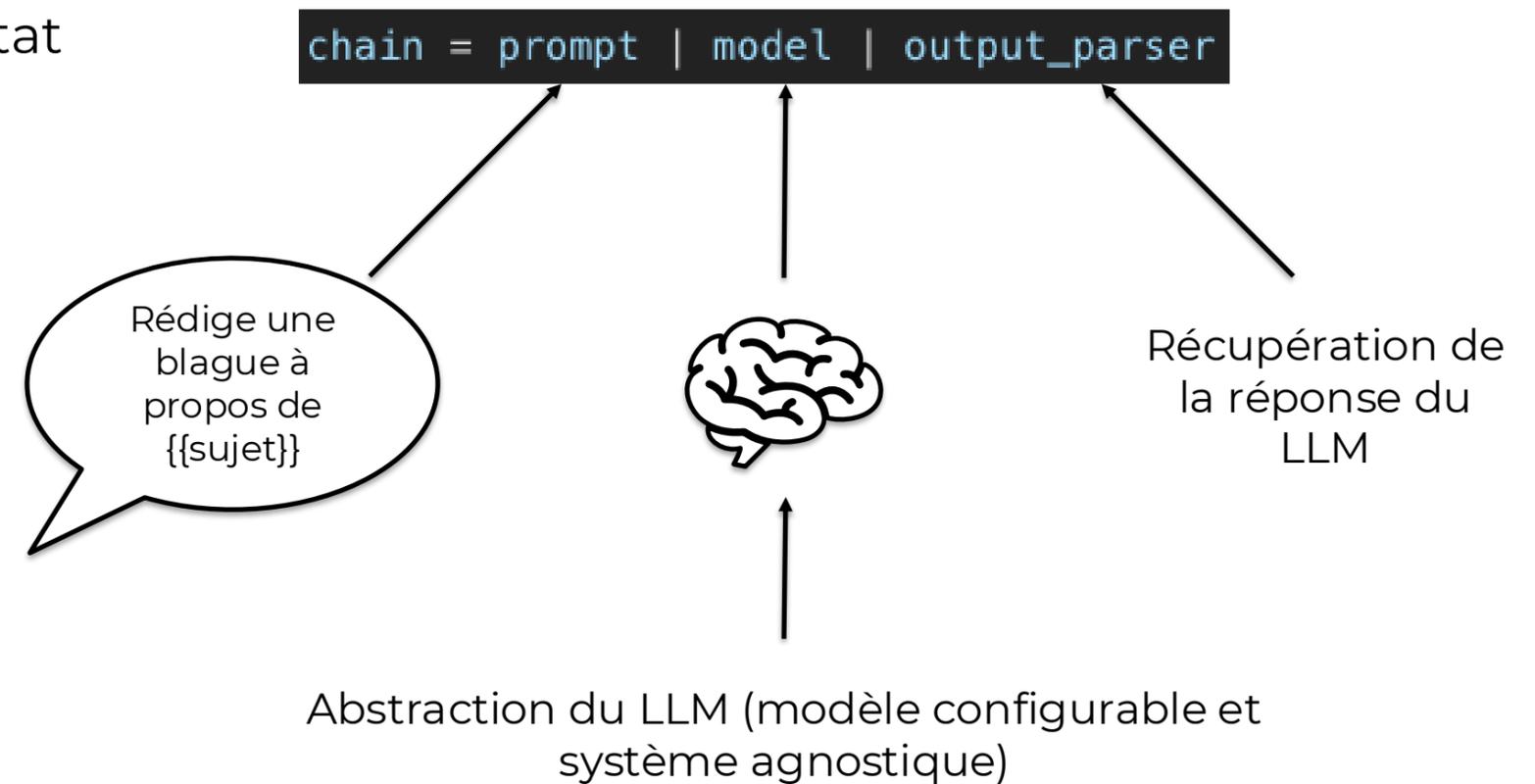




LangChain

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)





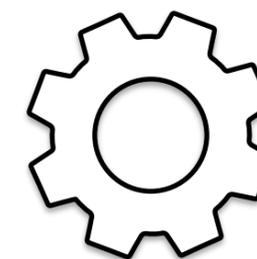
LangChain



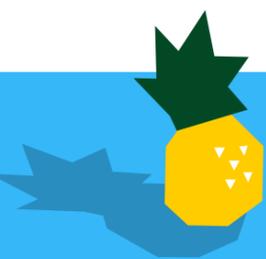
`{"sujet": "football"}`

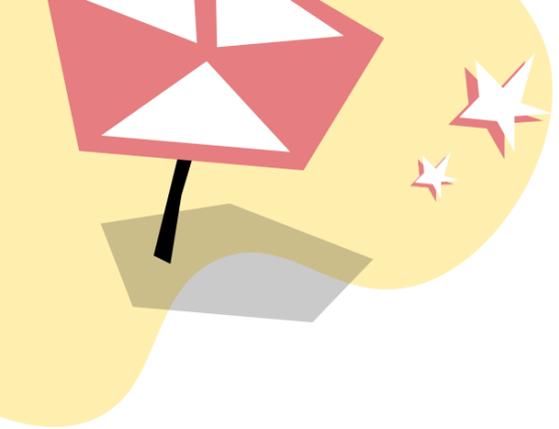


ChatMessage



Texte

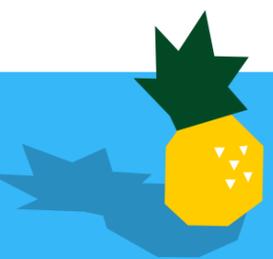
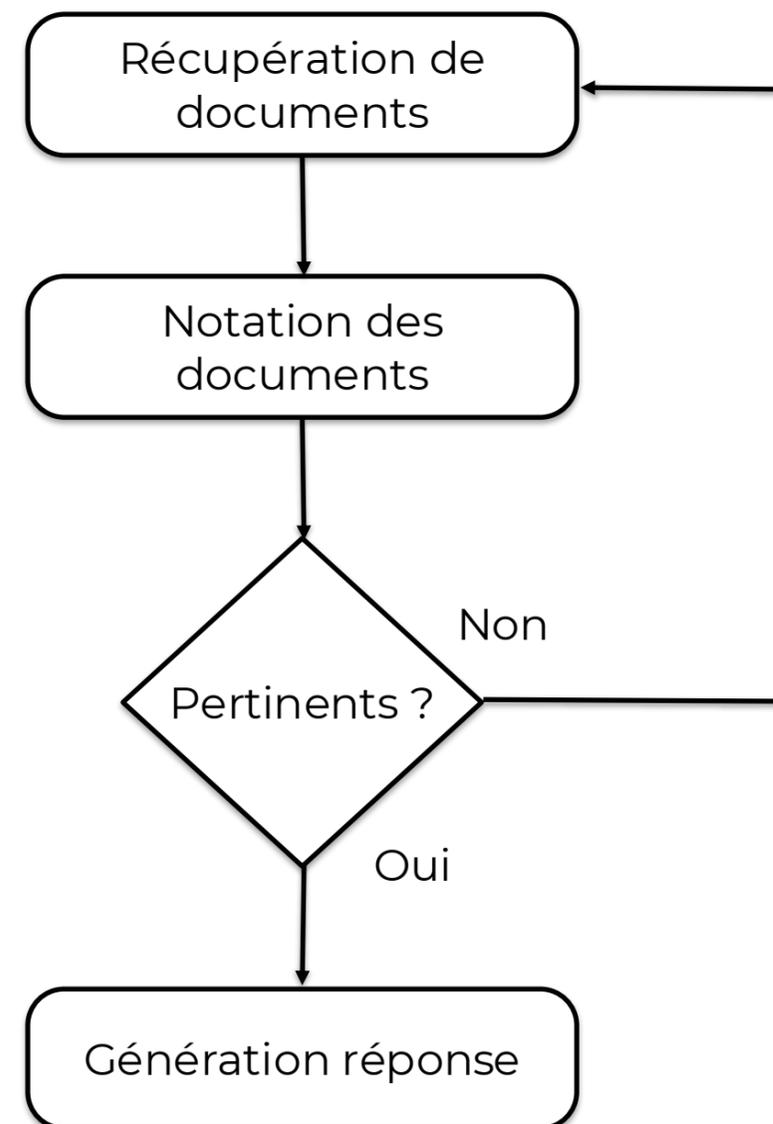




LangChain

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



Nombre de tokens* en sortie



Nombre de requêtes



Performance requise (modèle requis)



Exemple de calcul

1000

500

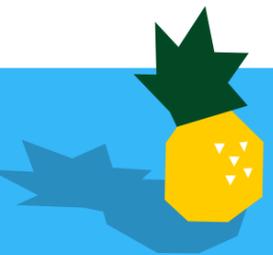
100

GPT-4o

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



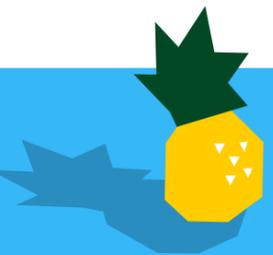
Comment superviser mon application ?

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

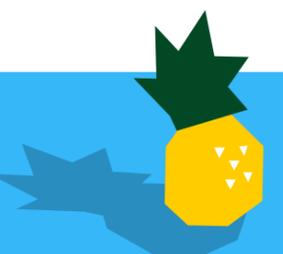
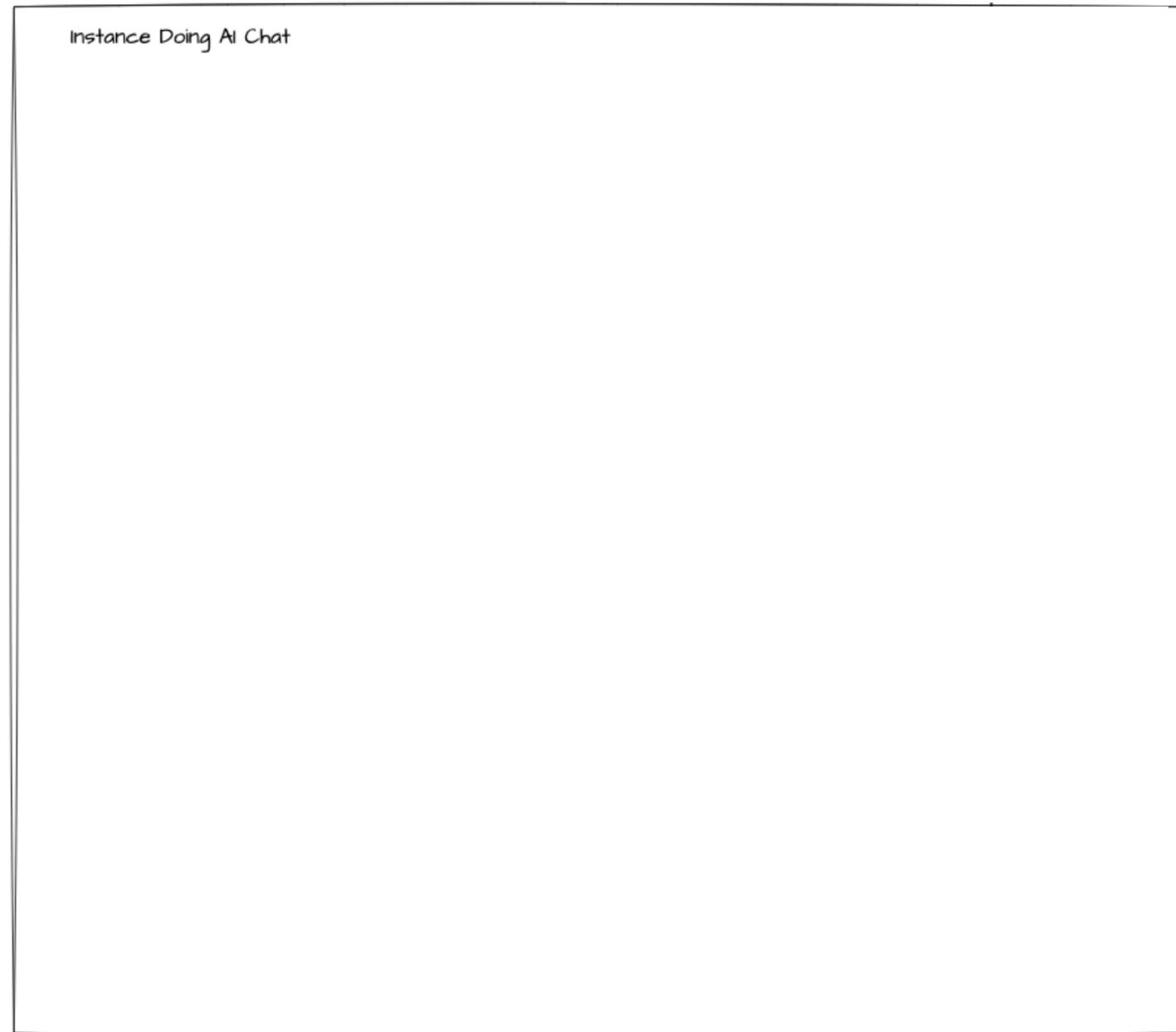
The screenshot displays the LangSmith interface for a project named 'langsmith-demo'. At the top, it shows performance metrics: P50 Latency: 13.43s, P99 Latency: 18.56s, and Total Tokens: 4,328. Below this, there are tabs for 'Traces', 'All Runs', and 'Setup'. A search bar contains the query: `and(eq(name, "ChatOpenAI"), gte(latency, "10s"))`. A table of traces is shown with the following columns: Run Type, Name, Input, Start Time, Latency, Tokens, and Tags. The table lists several traces for 'ChatOpenAI' with various latencies and token counts. On the right, a 'Filters' sidebar is visible, showing filters for Name (ChatOpenAI), Run Type (Llm), Status (Success), and Other (Latency >= 10s).

Run Type	Name	Input	Start Time	Latency	Tokens	Tags
LLM	ChatOpenAI	human: How di...	8/6/2023, 12:37:57 ...	16.11s	422	✓
LLM	ChatOpenAI	human: How d...	8/6/2023, 12:37:32 ...	17.51s	480	✓
LLM	ChatOpenAI	human: How d...	8/6/2023, 12:33:18 ...	18.70s	459	✓
LLM	ChatOpenAI	human: How di...	8/6/2023, 12:12:51 ...	11.23s	454	✓
LLM	ChatOpenAI	human: Tell me...	8/6/2023, 12:08:55 ...	15.00s	382	✓
LLM	ChatOpenAI	human: How di...	8/5/2023, 9:31:02 PM	13.87s	471	✓
LLM	ChatOpenAI	human: Tell me...	8/2/2023, 5:29:12 PM	13.43s	397	✓
LLM	ChatOpenAI	human: Tell me...	8/2/2023, 5:28:37 PM	13.77s	400	✓
LLM	ChatOpenAI	human: Tell me...	8/2/2023, 5:24:47 PM	12.87s	392	✓
LLM	ChatOpenAI	human: Tell me...	8/2/2023, 4:10:37 PM	11.29s	340	✓





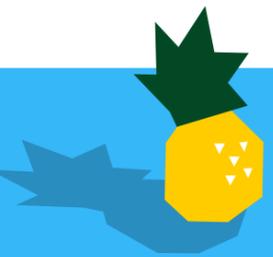
L'architecture chez DOING



The screenshot shows a web browser window with a URL `demo.rhag.fr/admin/entretien/formulaire-entretien/accueil.html#`. The main interface is titled "Editer un élément" and includes a "Générer un formulaire" button. A chatbot modal window is open, displaying a conversation:

- User: "Savoir être"
- Bot: "Très bien, nous allons commencer par la catégorie 'Savoir être'.
Quels critères souhaitez-vous inclure dans la catégorie 'Savoir être'?"
- User: "Adaptabilité, polyvalence, communication et relations, amabilité, coopération"
- Bot: "Parfait! Voici les critères que nous allons inclure dans la catégorie 'Savoir être':
1. Adaptabilité
2. Polyvalence
3. Communication et relations
4. Amabilité
5. Coopération
Souhaitez-vous ajouter d'autres catégories ou des commentaires spécifiques pour cette catégorie avant de passer à la génération du formulaire?"
- User: "ajoute la catégorie Savoir emploi"
- Bot: "Très bien! Nous allons ajouter la catégorie 'Savoir emploi' à notre formulaire.
Quels critères souhaitez-vous inclure dans la catégorie 'Savoir emploi'?"

At the bottom of the modal, there is a text input field containing the text: "Analyser le besoin, accompagner le".



Chrome Ficheur Mailto Afficher Historique Favoris Profils Onglets Fenêtres Aide

Long AI Chat

Non connecté 01.021.021.104

Long AI Chat - staging_v0.7.2+3201

Fichiers de discussion

Glissez/Déposez les fichiers liés à la discussion ici
 Maximum 3 fichiers de 100MB (autres documents, pdf, carte, seulement)

Télécharger vos fichiers

Configuration

Fournisseur IA: OpenAI (gpt-3.5-turbo)

Organisation: OpenAI

Nombre de messages utilisés dans l'historique: 20

Instructions système

Message

Support CEETAL

Support CEETAL

Contexte: Vous êtes un agent de support pour l'entreprise CEETAL. La société Laboratoire CEETAL CMPC, concepteur et fabricant de spécialités chimiques, fournit des solutions d'hygiène et de maintenance technique à destination des professionnels. Fort d'une expérience de 60 ans dans le secteur de la chimie, présent en France et à l'international dans plus de 65 pays, notre objectif est de vous proposer des solutions efficaces, respectueuses des utilisateurs et de l'environnement.

Objectif: Générer des réponses précises et détaillées en s'appuyant exclusivement sur le contexte et la documentation fournis, minimisant ainsi les risques d'erreur de hallucination.

Instructions:

- Utilisation d'outils: L'utilisation des outils à disposition permet de récupérer du contexte quand nécessaire.
- Extraction d'informations: Utiliser uniquement la documentation et l'historique de conversation pour formuler des réponses.
- Précision et détails: Fournir des explications techniques approfondies quand nécessaire.
- Séduction d'erreurs: Éviter de générer des réponses non soutenues par les données disponibles.
- Adaptabilité du langage: Adapter le niveau de technicité du langage à la complexité de la question.
- Communication des limites: Informer l'utilisateur des limites du système face à des demandes hors du champ documenté.

6 modules disponibles

Tout ouvrir Tout fermer

Message

Trouve moi des désinfectants

Info sur l'utilisateur

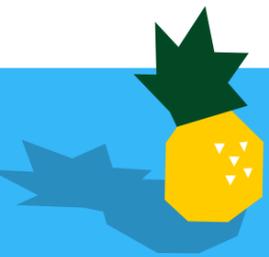
admin:chat

Tools

get_products, get_product_details

Envoyer

Ces réponses sont générées par IA et peuvent ne pas être exactes.



Pour ne rien oublier ...

vosre 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

