Agents

Introduction à l'intelligence artificielle générative

Théo Lopes Quintas

BPCE Payment Services,
Université Paris Dauphine

2025-2026

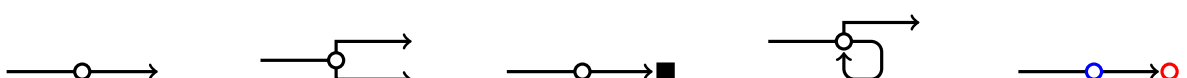
Introduction

Her - Spike Jonze

Introduction

Qu'est-ce qu'un agent ?

Un **Agent IA** est un programme où la sortie d'un modèle de langage contrôle la suite du programme.



(a) Processeur **(b)** *Router* **(c)** Appel de fonction **(d)** Multi Step Agent **(e)** Multi-Agent

Figure – Différents type d'agents

Introduction

Comment mesurer la performance ?

[Mialon et al., 2023] propose le benchmark GAIA [1]. Il est composé d'environs 450 questions réparti en trois niveaux représentant la difficulté. Au moment de la publication, les testeurs humains obtenaient 92% quand GPT-4 avec plugins obtenait 15%, allant à l'encontre de la tendance en 2023 d'avoir des modèles de langage qui dépasse les capacités humaine.

1. On recommande l'excellente vidéo "L'IA a discrètement franchi une étape historique" du podcast Underscore

Introduction

Comment mesurer la performance ?

[Mialon et al., 2023] propose le benchmark GAIA [1]. Il est composé d'environs 450 questions réparti en trois niveaux représentant la difficulté. Au moment de la publication, les testeurs humains obtenaient 92% quand GPT-4 avec plugins obtenait 15%, allant à l'encontre de la tendance en 2023 d'avoir des modèles de langage qui dépasse les capacités humaine.

Exemple de question de niveau 3

Question : In NASA's Astronomy Picture of the Day on 2006 January 21, two astronauts are visible, with one appearing much smaller than the other. As of August 2023, out of the astronauts in the NASA Astronaut Group that the smaller astronaut was a member of, which one spent the least time in space, and how many minutes did he spend in space, rounded to the nearest minute ? Exclude any astronauts who did not spend any time in space. Give the last name of the astronaut, separated from the number of minutes by a semicolon. Use commas as thousands separators in the number of minutes. Ground truth : White ; 5876

Figure – Charles Conrad documentant la collecte de roche lunaire de Alan Bean

1. On recommande l'excellente vidéo "L'IA a discrètement franchi une étape historique" du podcast Underscore

En résumé

Comment interagir ?

Model Context Protocol

Le **Model Context Protocol** est un protocole open-source proposé par Anthropic en novembre 2024. Il s'appuie sur un modèle client-server, qu'on peut simplifier :

Le **Host** correspond à un programme comme un IDE, et un **client MCP** est un protocol qui maintien les **connexions** uniques aux serveurs.

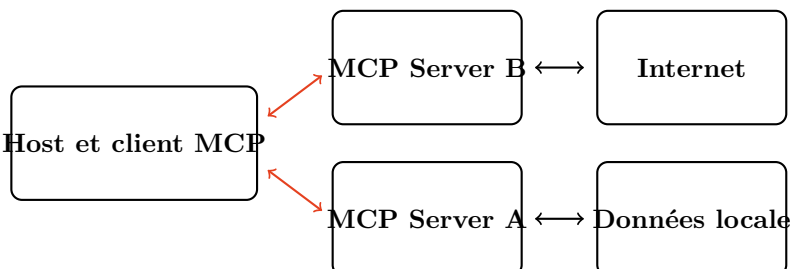


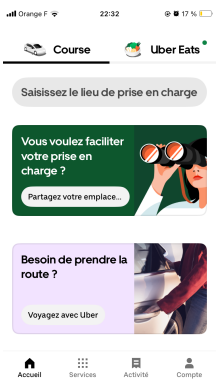
Figure – MCP simplifié

Ainsi, si on a M application différentes et N outils alors nous n'avons besoin de construire que $M+N$ intégration au lieu de $M \times N$.

Comment interagir ?

Problématiques

(a) AliPay



(b) Uber

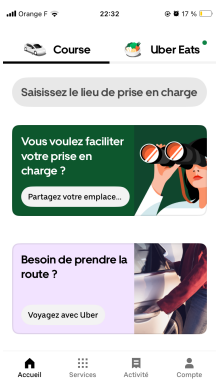
Figure – Deux exemples d'UI pour des applications mobile

Une autre approche pour rendre les agents autonome serait qu'il puisse *comprendre* une interface utilisateur (UI).

Comment interagir ?

Problématiques

(a) AliPay



(b) Uber

Figure – Deux exemples d'UI pour des applications mobile

Une autre approche pour rendre les agents autonome serait qu'il puisse *comprendre* une interface utilisateur (UI).Cependant, de grandes différences peuvent exister :

- Selon les **cultures**, asiatique et occidentale par exemple
- Pour chaque élément il faut être capable d'**inférer l'action** produite par son interaction
- En prenant en compte les différences de version de téléphone, donc de **résolution**

Comment interagir ?

Modèle Ferret

Apple publie en 2023 le modèle de langage multimodal **Ferret** [You et al., 2023] capable de comprendre les références dans une image de n'importe quelle forme à n'importe quelle échelle.

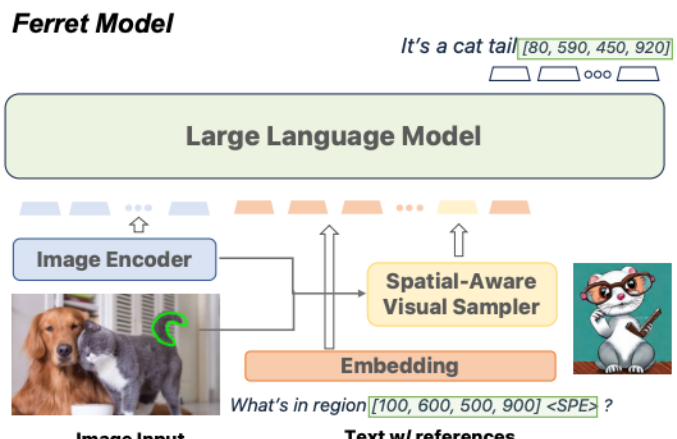


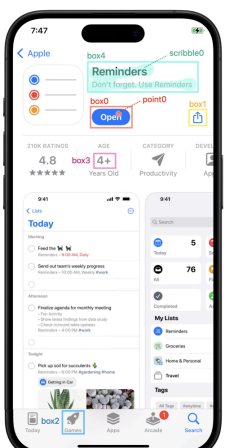
Figure – Schéma du modèle Ferret [You et al., 2023]

Pour encoder l'image CLIP-ViT-L/14 [Radford et al., 2021] est utilisé et pour le texte c'est le tokenizer du modèle concerné. Le *Spatial-Aware Visual Sampler* correspond lui à une méthode pour encoder l'information de la forme mise en évidence.

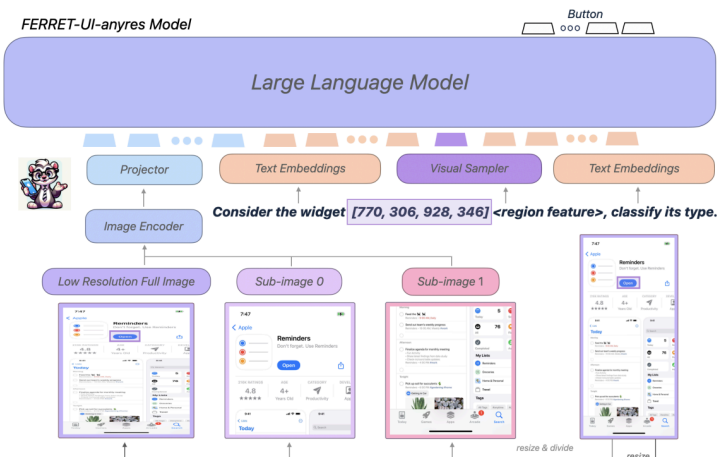
Comment interagir ?

Ferret UI

Les auteurs de Ferret le spécialise ensuite sur la reconnaissance d'interface utilisateur pour proposer le modèle **Ferret UI** [You et al., 2024].



(a) Exemple d'un écran



(b) Architecture du modèle Ferret UI

Figure – [You et al., 2024].

Comment interagir ?

Discussion entre agents

Un agent pouvant en appeler un autre [2], il faut trouver la meilleure manière pour qu'ils puissent échanger. Initialement, les agents transmettaient les informations via des JSON.
[Nguyen et al., 2024, Wang et al., 2024] montrent que la communication via une rédaction de code est plus efficace que par JSON pour les raisons suivantes :

- **Flexibilité** : une fonction en Python permet d'exprimer plus simplement des actions qu'une multitude de JSON, éventuellement imbriqués
- **Management** : il n'est pas aisé de stocker une image ou un son par exemple dans un JSON
- **Généralité** : les actions exprimées avec du code représente presque l'entiereté de ce que l'on peut faire avec un ordinateur
- **Représentation** : dans les données d'entraînement, une large part est donnée à la programmation, plus qu'à un JSON !

2. C'est le cadre multi-agent, voir figure 1e)

Bibliographie I

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022).
Constitutional ai : Harmlessness from ai feedback.
arXiv preprint arXiv :2212.08073.

Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., and Scialom, T. (2023).
Gaia : a benchmark for general ai assistants.
In *The Twelfth International Conference on Learning Representations*.

Nguyen, D., Lai, V. D., Yoon, S., Rossi, R. A., Zhao, H., Zhang, R., Mathur, P., Lipka, N., Wang, Y., Bui, T., et al. (2024).
Dynasaur : Large language agents beyond predefined actions.
arXiv preprint arXiv :2411.01747.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021).
Learning transferable visual models from natural language supervision.
In *International conference on machine learning*.

Wang, X., Chen, Y., Yuan, L., Zhang, Y., Li, Y., Peng, H., and Ji, H. (2024).
Executable code actions elicit better llm agents.
In *Forty-first International Conference on Machine Learning*.

Bibliographie II


You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.-F., and Yang, Y. (2023).
Ferret : Refer and ground anything anywhere at any granularity.
arXiv preprint arXiv :2310.07704.


You, K., Zhang, H., Schoop, E., Weers, F., Swearngin, A., Nichols, J., Yang, Y., and Gan, Z. (2024).
Ferret-ui : Grounded mobile ui understanding with multimodal llms.
In *European Conference on Computer Vision*. Springer.

Annexe

Anthropic

Fondée en 2021, Anthropic est une entreprise américaine qui a développé la série de modèle **Claude** ainsi que publié des résultats notable dans la sécurité autour de l'intelligence artificielle.

Figure – Dario Amodei, CEO d'Anthropic

L'approche **Constitutional AI** [Bai et al., 2022] a été développé par Anthropic dans le but de produire des modèles de langage sans danger sans recourir à des usages extensif de labellisateur humain. Dans une première phase le modèle génère une réponse puis la critique à la lumière d'un ensemble de principe et finalement révise sa réponse. Le modèle est ensuite fine-tuné avec de l'apprentissage par renforcement pour s'aligner avec les préférences.

L'entreprise est valorisée à 61.5 milliards de données en 2025 avec Amazon en premier investisseur en 2023 pour plus de 4 milliards de dollars. En contrepartie Anthropic utilisera AWS et les puces d'Amazon. L'entreprise a également un partenariat avec Palantir pour proposer à l'industrie militaire américaine des solutions d'IA.