

Evaluation des agents IA

Nassim Bennouar

La recrudescence des agents IA

- Assister les développeurs
- Aider aux tâches administratives
- Analyser des données

La recrudescence des agents IA

- Assister les développeurs
- Aider aux tâches administratives
- Analyser des données
- Organiser des voyages
- Aider à planifier des mariages

La recrudescence des agents IA

- Assister les développeurs
- Aider aux tâches administratives
- Analyser des données
- Organiser des voyages
- Aider à planifier des mariages
- (Bien) jouer à Minecraft

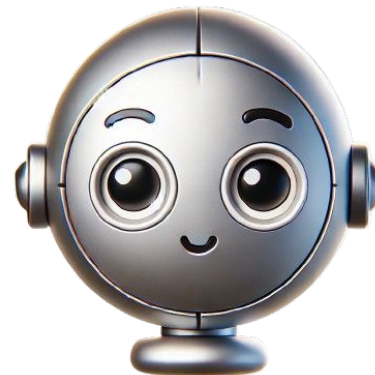
Sommaire

1. Qu'est-ce qu'un agent IA
2. Évaluer un agent IA
3. Agent-as-a-Judge
4. Perspectives d'avenir

Qu'est-ce qu'un agent IA ?

Fredo, le majordome virtuel

Voici Fredo



Fredo, le majordome virtuel

Voici Fredo

Sa mission : Mettre de l'ordre dans vos fichiers



Fredo, le majordome virtuel

Voici Fredo

Sa mission : Mettre de l'ordre dans vos fichiers

Comment ?

- **Il observe** 👁️ : le nom des fichiers, les types et les dates
- **Il réfléchit** 🧠 : aux schémas et aux catégories
- **Il agit** 🎯 : en classant, en renommant, en archivant intelligemment





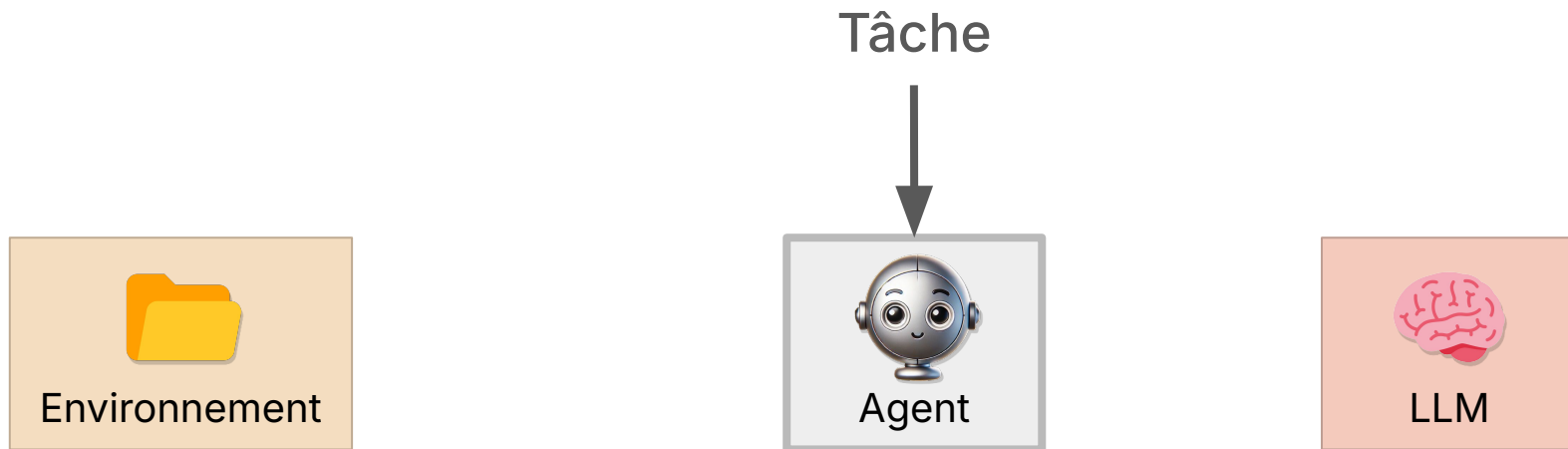
Environnement

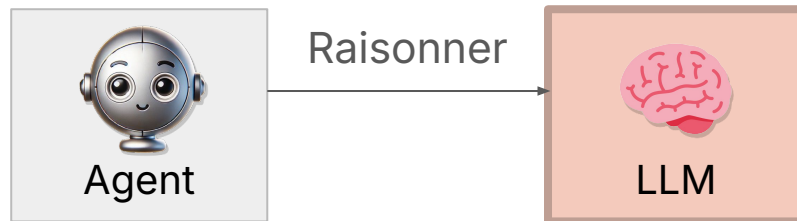
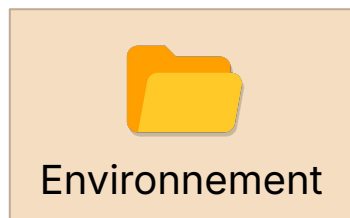


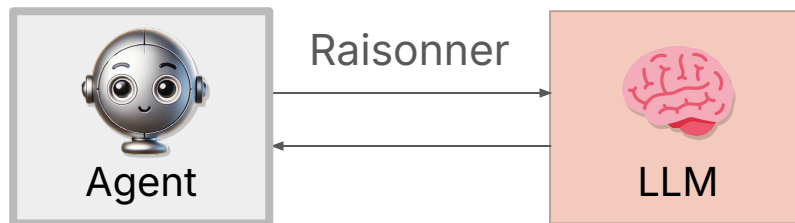
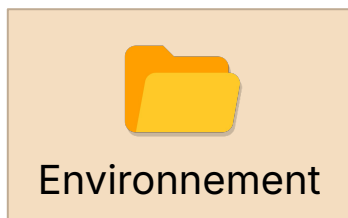
Agent

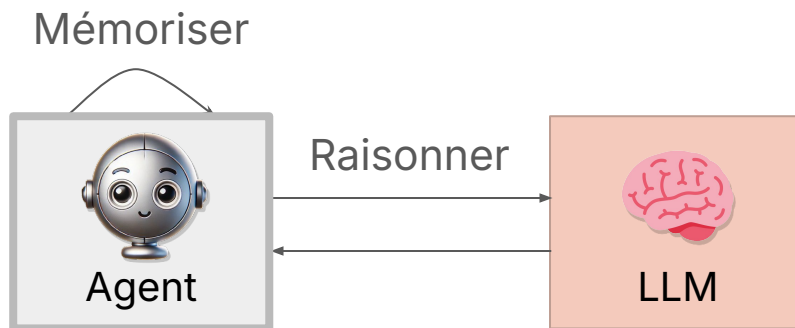
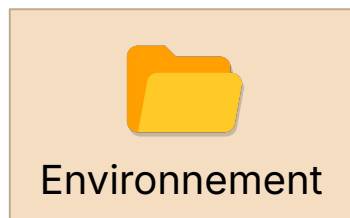


LLM



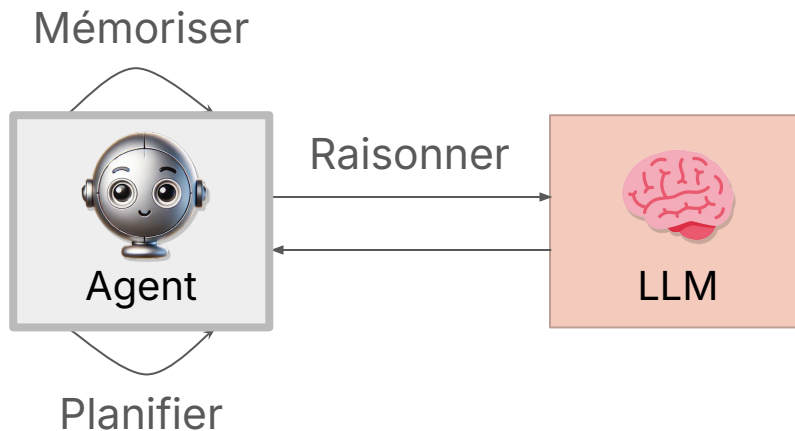
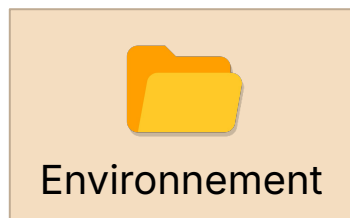






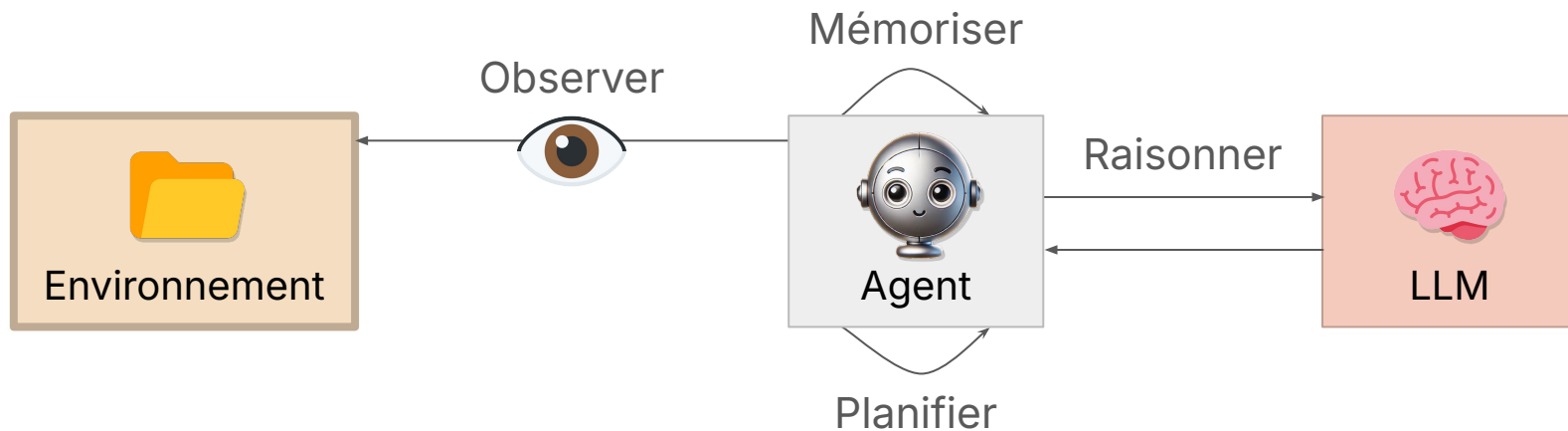
Liste d'outils

- Récupérer l'arborescence de fichiers (observation)



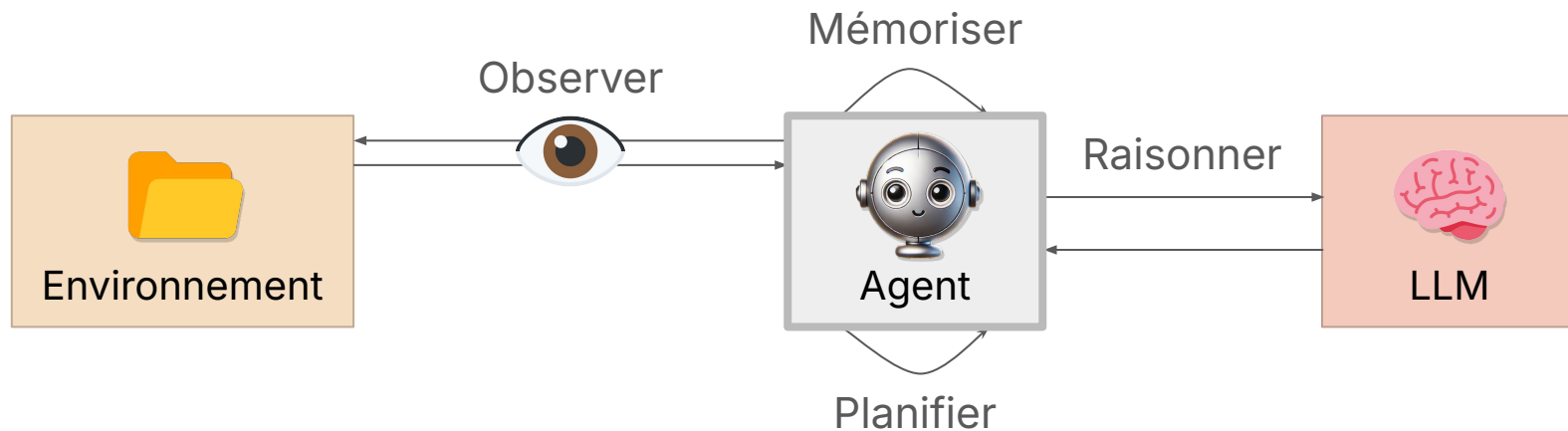
Liste d'outils

- Récupérer l'arborescence
de fichiers (observation)



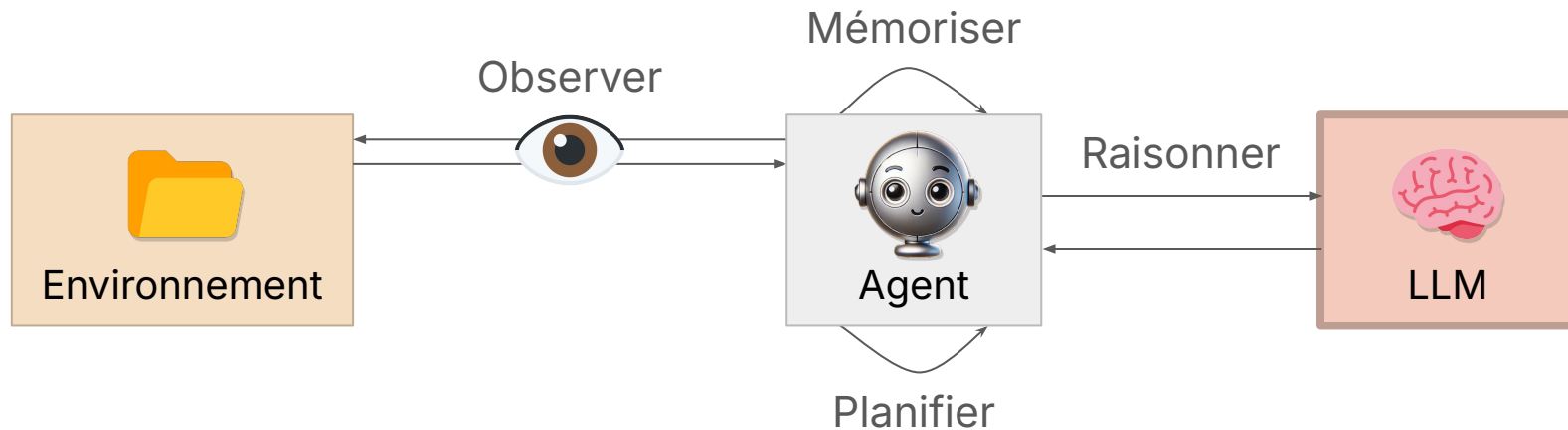
Liste d'outils

- Récupérer l'arborescence
de fichiers (observation)



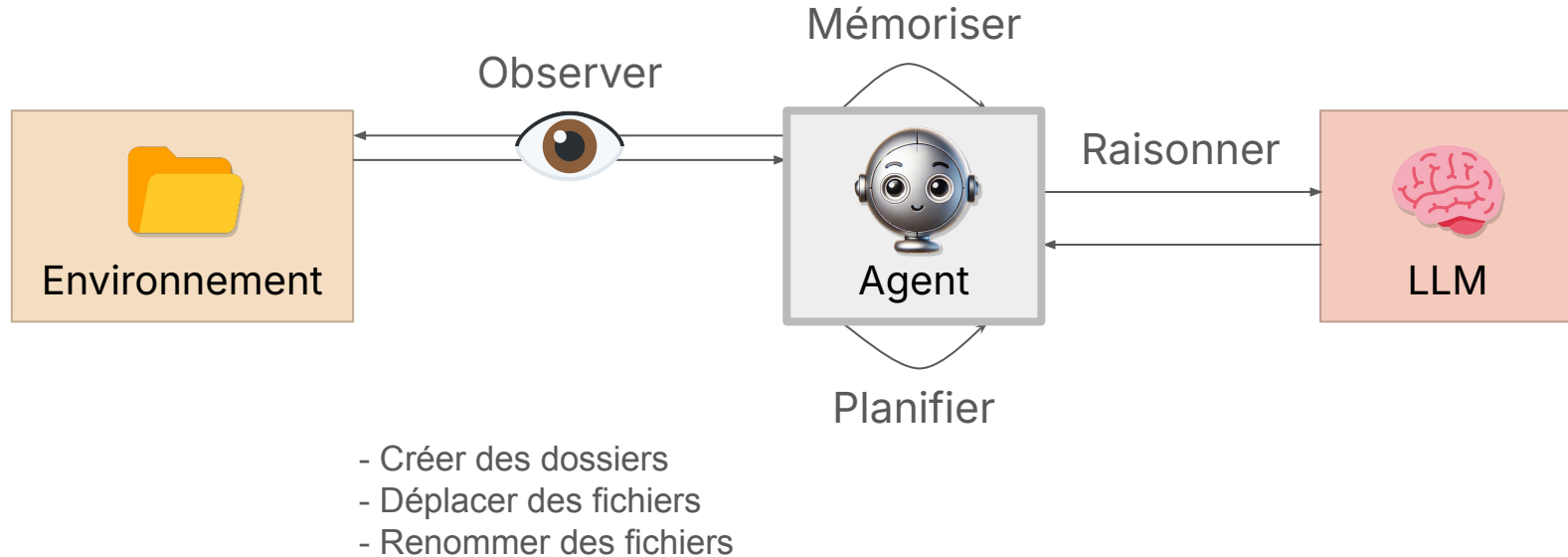
Liste d'outils

- Récupérer l'arborescence
de fichiers (observation)



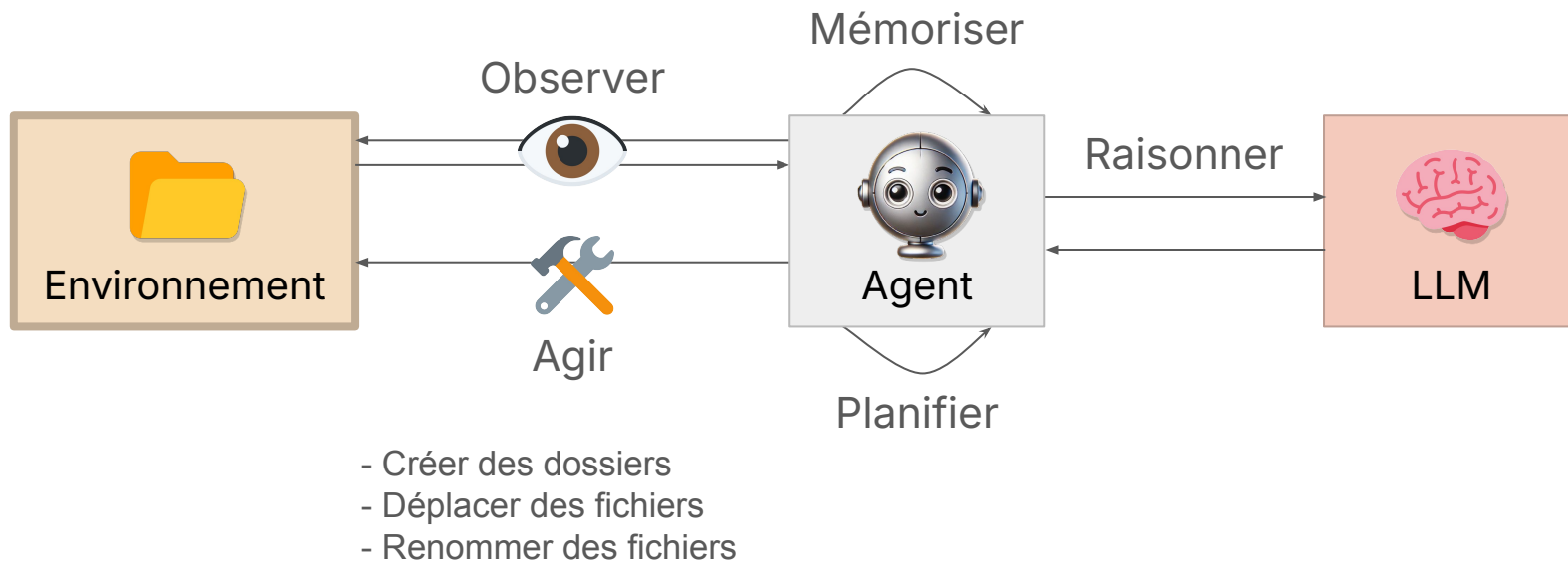
Liste d'outils

- Récupérer l'arborescence de fichiers (observation)



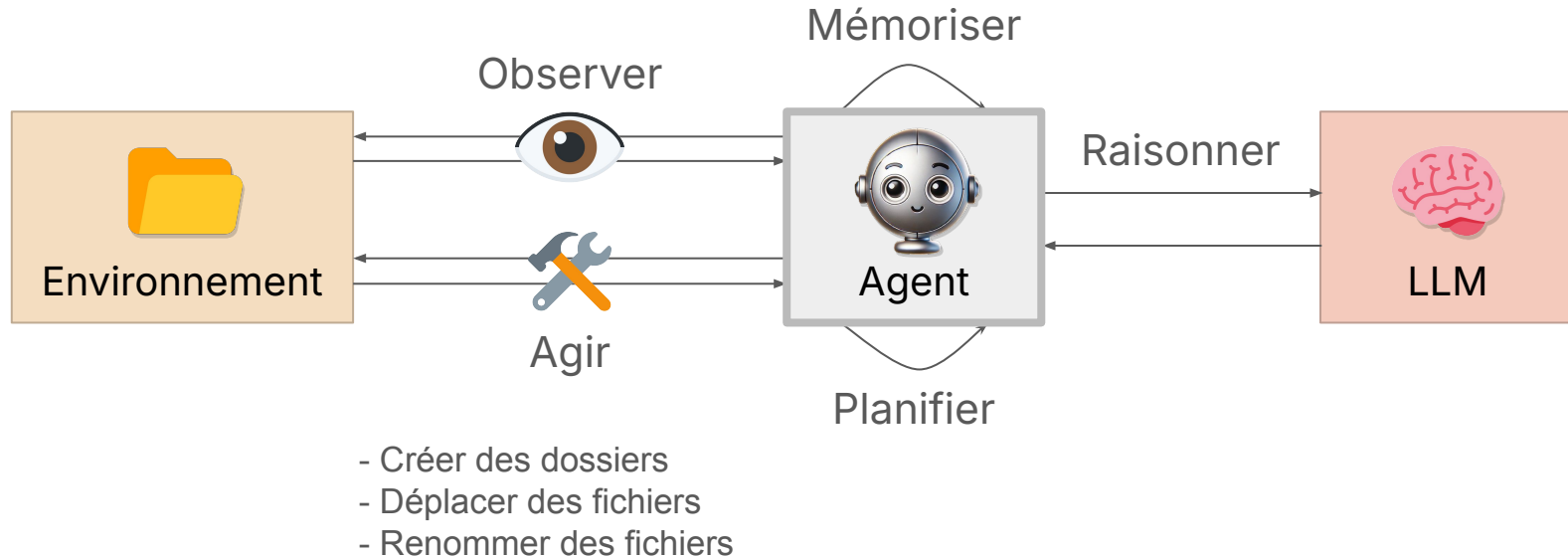
Liste d'outils

- Récupérer l'arborescence
de fichiers (observation)



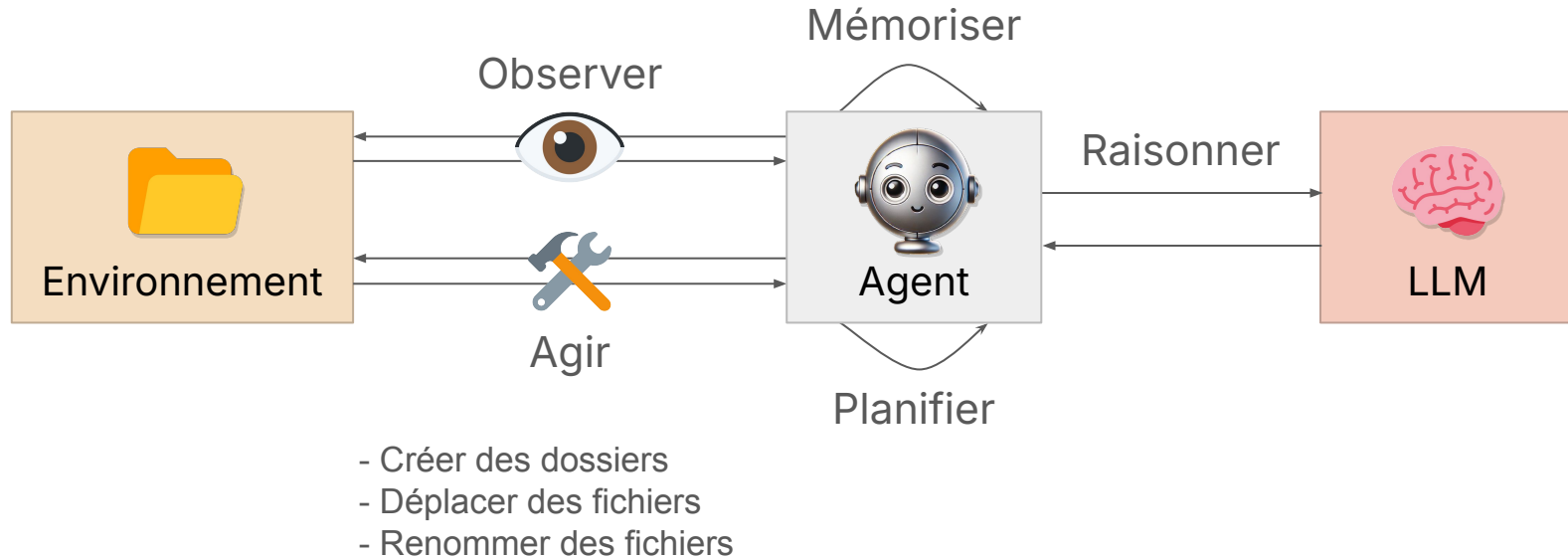
Liste d'outils

- Récupérer l'arborescence
de fichiers (observation)



Liste d'outils

- Récupérer l'arborescence
de fichiers (observation)

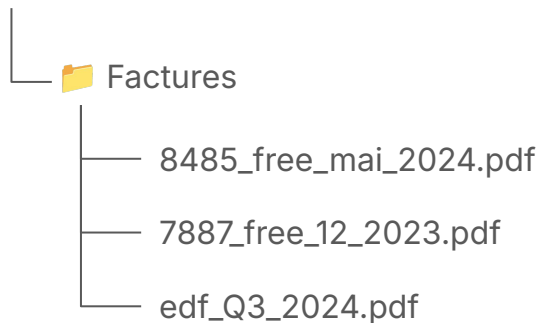


Qu'est-ce qu'un agent IA ?

Pour résumer, un agent IA est un programme capable d'agir sur un **environnement** de façon à atteindre les **objectifs** que l'utilisateur lui donne, en se reposant sur un moteur qui simule un **raisonnement** pour **planifier** des actions pertinentes et les **effectuer**.

Évaluer un agent IA

Évaluer les résultats de Fredo

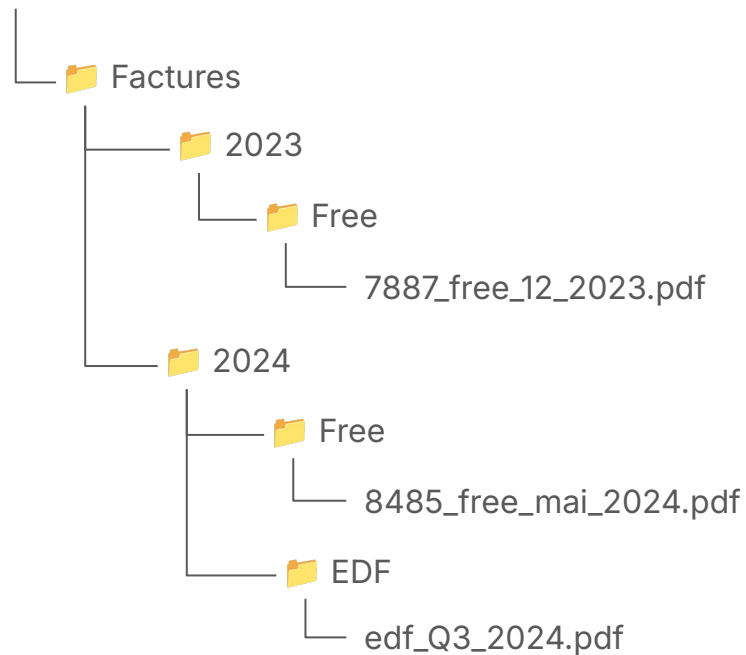
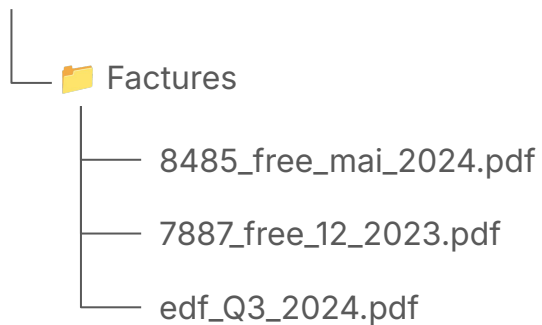


Objectif : Classer des factures par service puis par année

Critères d'évaluations :

- Les fichiers ont bien été classés
- Les fichiers ont été classés dans le bon sens
- De préférence, les fichiers ont été renommés

Premier essai

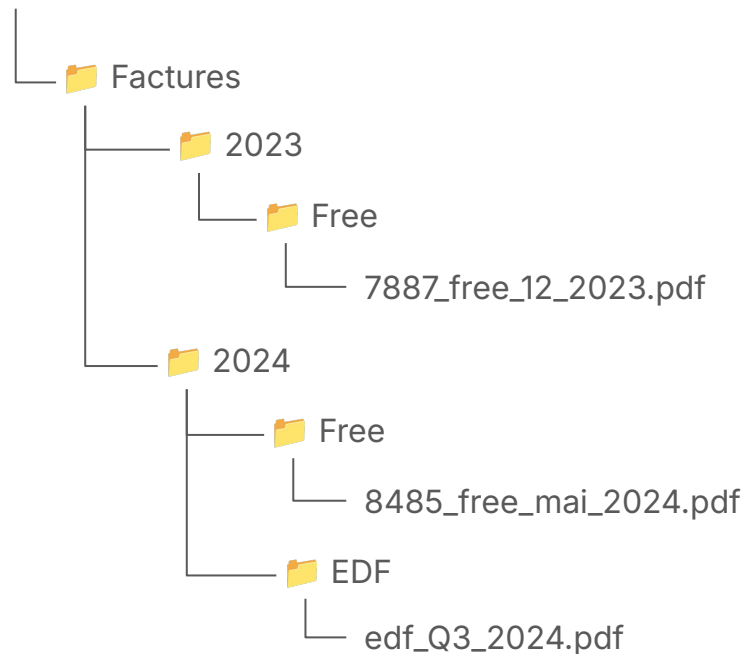


Premier essai

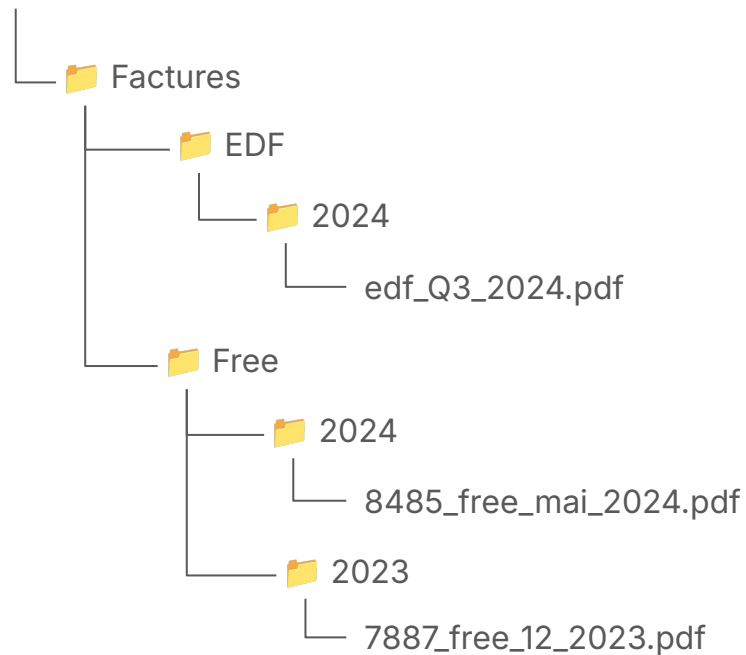
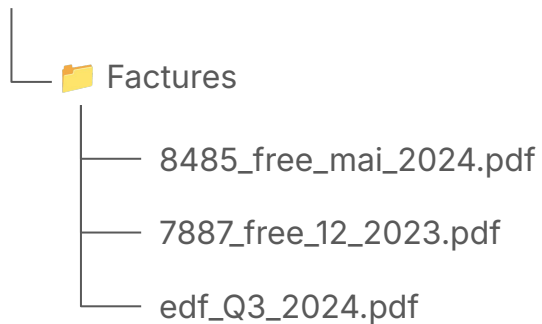
Fichiers classés de façon cohérente

Mauvais ordre de classement :
année ↔ service

Résultat : 4/10



Deuxième essai

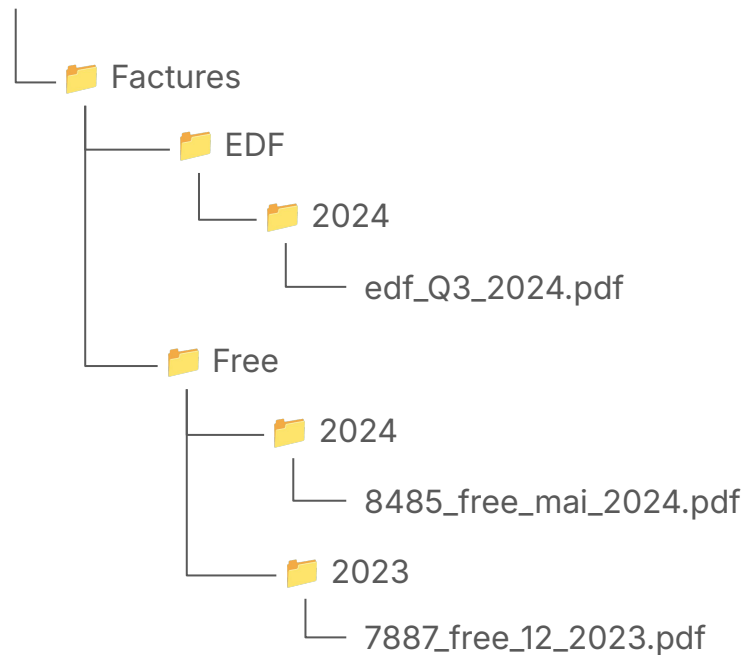


Deuxième essai

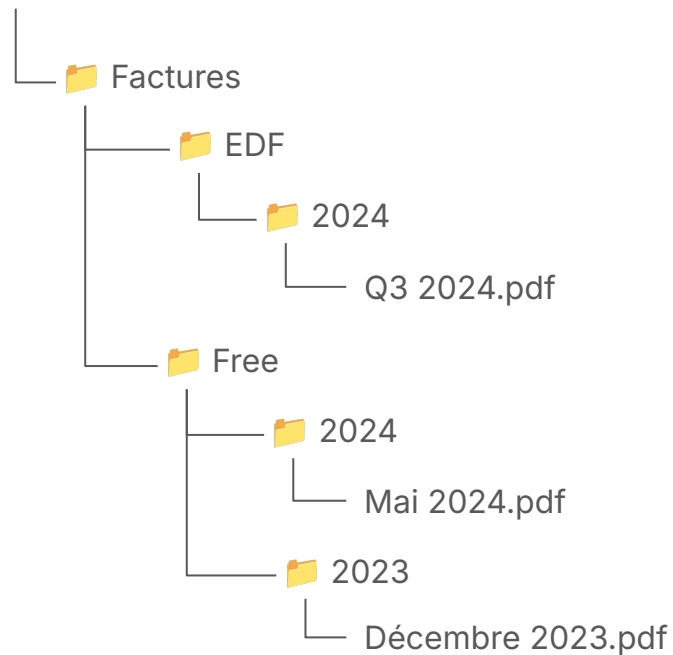
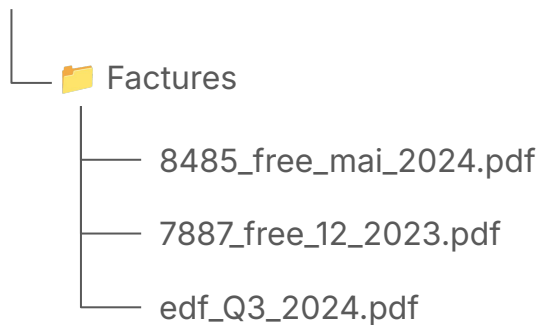
Fichiers bien classés

Fichiers non renommés

Résultat : 9/10 🖐️



Troisième essai

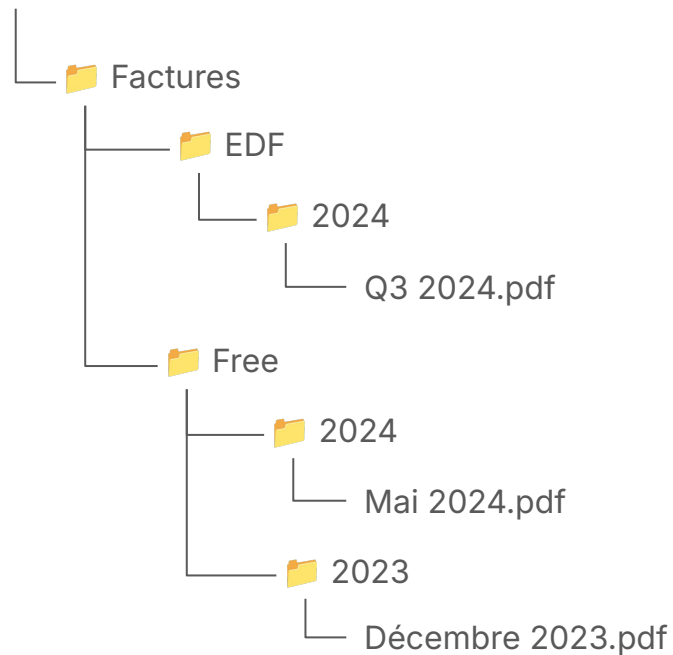


Troisième essai

Fichiers bien classés

Fichiers renommés

Résultat : 10/10 🎉



Est-ce suffisant de regarder le résultat ?

Exécution du deuxième essai (3 cycles)

LLM : Observe le dossier Factures

Observation du dossier Factures

LLM : Crée les dossiers ... puis
déplace ... puis observe le résultat

Création des dossiers EDF Free ...

Déplacement des fichiers ...

Observation du dossier Factures

LLM : Très bien, fin d'exécution

Exécution du troisième essai (70 cycles)

LLM : Observe le dossier Factures

Observation du dossier Factures

LLM : Crée un dossier Free_2023 et
EDF_2023

Création des dossiers Free_2023 ...

LLM : Déplace tous les fichiers
dans Free_2023

...

Est-ce suffisant de regarder le résultat ?

Exécution du deuxième essai (3 cycles)

LLM : Observe le dossier Factures

Observation du dossier Factures

LLM : Crée les dossiers ... puis
déplace ... puis observe le résultat

Création des dossiers EDF Free ...

Déplacement des fichiers ...

Observation du dossier Factures

LLM : Très bien, fin d'exécution

Exécution du troisième essai (70 cycles)

LLM : Observe le dossier Factures

Observation du dossier Factures

LLM : Crée un dossier Free_2023 et
EDF_2023

Création des dossiers Free_2023 ...

LLM : Déplace tous les fichiers
dans Free_2023

...

Types d'évaluation

Evaluation du résultat seul

- Suffisant pour des tâches simples
- Plus facile et plus rapide
- Incapacité à pointer les étapes qui coïncident
- Incapacité à adresser les problématiques d'optimisation

Evaluation de l'exécution complète

- Adapté aux tâches complexes
- Identification des compétences dysfonctionnelles
- Observabilité plus fine des coûts et du temps d'exécution
- Plus onéreux ?

Agent-as-a-Judge : Evaluate Agents with Agents

Meta AI, KAUST

Motivation

- Les agents IA pour les développeurs se sont largement améliorés et diffusés (Devin, Cursor)
- Une évaluation efficace de ceux-ci demande trop de travail humain
- Les benchmarks actuels (HumanEval, SWE-Bench, MLE-Bench) ne prennent pas en compte les étapes intermédiaires de développement
- Le métier est plus complexe que ce sur quoi les chercheurs s'alignent

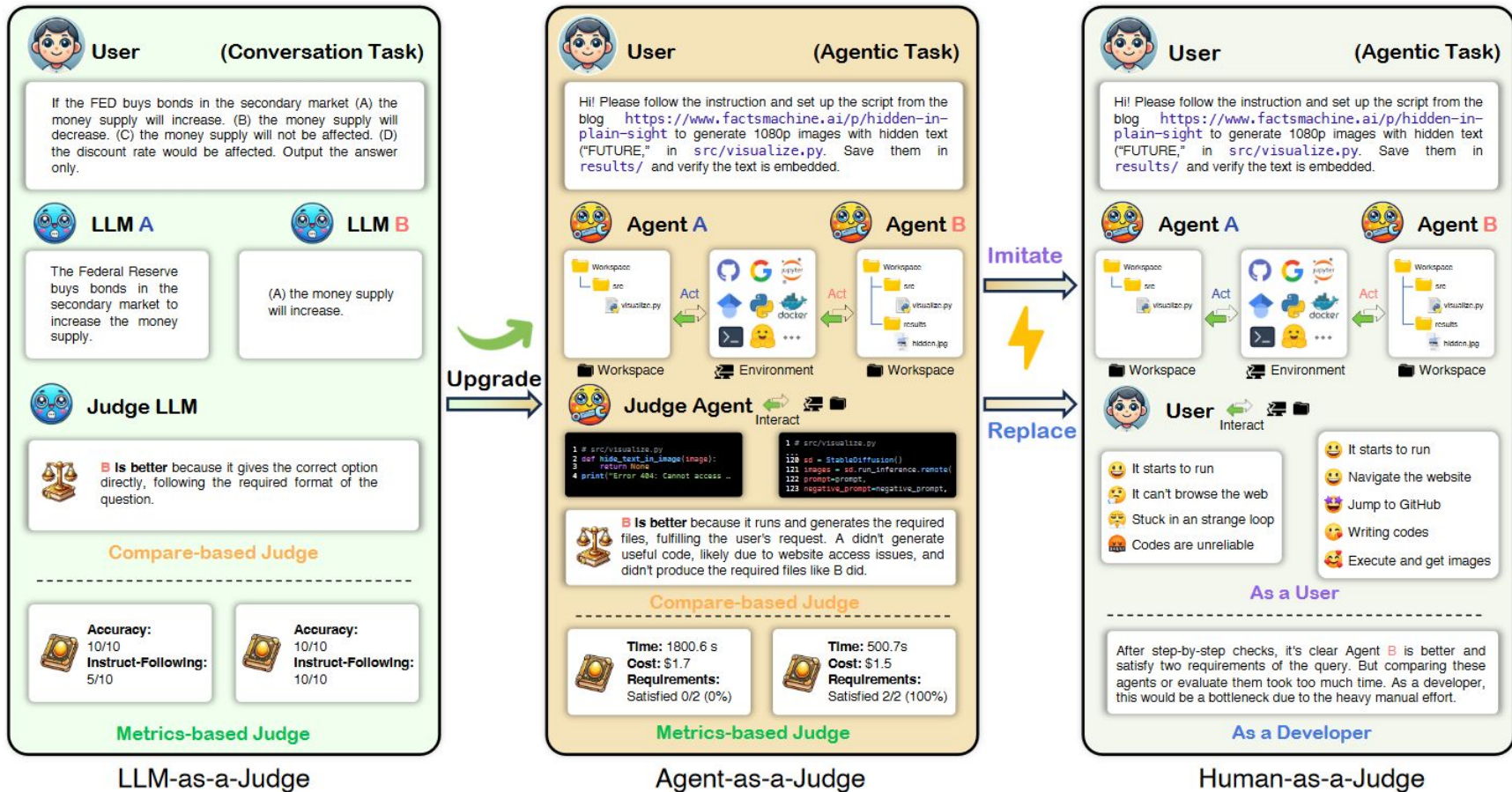


Figure 1 We introduce the Agent-as-a-Judge framework wherein agentic systems are used to evaluate agentic systems. We compare this to LLM-as-a-Judge, which uses LLMs to evaluate LLMs and for which Agent-as-a-Judge is a natural evolution, and Human-as-a-Judge, where skilled human labourers manually evaluate an agentic system.

Contributions

- Création du DevAI dataset, comprenant cinquante-cinq tâches réalistes :
 - Avec un certain nombre d'exigences intermédiaires
 - Des dépendances entre les exigences
 - Et quelques préférences
- Conception d'un framework d'évaluation agentique, Agent-as-a-Judge
- Evaluation de trois agents IA sur le DevAI dataset avec :
Agent-as-a-Judge, LLM-as-a-Judge, Human-as-a-Judge

Contributions

- Création du DevAI dataset, comprenant cinquante-cinq tâches réalistes :
 - Avec un certain nombre d'exigences intermédiaires
 - Des dépendances entre les exigences
 - Et quelques préférences
- ~~— Conception d'un framework d'évaluation agentique, Agent as a Judge~~
- Evaluation de trois agents IA sur le DevAI dataset avec :
Agent-as-a-Judge, LLM-as-a-Judge, Human-as-a-Judge



Texte “FUTURE” incrusté par Stable Diffusion et ControlNet



Exemple abrégé de tâche dans DevAI

Requête : Salut !

Suit les instructions de l'article [Hidden in Plain Sight](#) pour développer un script qui génère des images avec du texte caché dans `src/visualize.py`.

Assure-toi que l'image soit en 1080p et enregistrée dans `results/`.

Crée des images de contrôle en intégrant le texte "FUTURE" et enregistre les également dans `results/`.

Vérifie manuellement que le texte caché soit intégré à l'image.

Exemple abrégé de tâche dans DevAI

Exigence 1 : A suivi les instructions de l'article et développé le script pour générer des images avec du texte caché dans `src/visualize.py`

- **Dépendances** → {}

Exigence 2 : Les images sont générées en 1080p dans `results/`

- **Dépendances** → {E1}

Exigence 3 : A créé des images de contrôle en intégrant le texte "FUTURE" et en les enregistrant dans `results/`.

- **Dépendances** → {E2}

Exemple abrégé de tâche dans DevAI

Préférence 1 : Le système doit être capable d'apprendre et de s'adapter à des technologies et outils inattendus

Préférence 2 : Après avoir vu l'article de blog, ControlNet (*le modèle qui permet d'intégrer le texte*) doit tourner sur Modal (*une plateforme serverless recommandée dans l'article*) pour produire des images avec "FUTURE" caché

Evaluation de trois agents de développement stars

- Trois agents ont été testés sur les tâches du DevAI dataset : MetaGPT, GPT-Pilot, OpenHands

Evaluation de trois agents de développement stars

- Trois agents ont été testés sur les tâches du DevAI dataset : MetaGPT, GPT-Pilot, OpenHands
- Trois juges ont été convoqués pour évaluer leurs résultats
 - Un groupe de trois experts en IA (Human-as-a-Judge)
 - LLM-as-a-Judge (sur les résultats finaux)
 - Agent-as-a-Judge

Evaluation de trois agents de développement stars

- Trois agents ont été testés sur les tâches du DevAI dataset : MetaGPT, GPT-Pilot, OpenHands
- Trois juges ont été convoqués pour évaluer leurs résultats
 - Un groupe de trois experts en IA (Human-as-a-Judge)
 - LLM-as-a-Judge (sur les résultats finaux)
 - Agent-as-a-Judge
- Est mesuré l'alignement de Agent-as-a-Judge et de LLM-as-a-Judge avec le groupe d'experts
- Est mesurée la performance des agents







Evaluation de trois agents de développement stars

- Trois agents ont été testés sur les tâches du DevAI dataset : MetaGPT, GPT-Pilot, OpenHands
- Trois juges ont été convoqués pour évaluer leurs résultats
 - Un groupe de trois experts en IA (Human-as-a-Judge)
 - LLM-as-a-Judge (sur les résultats finaux)
 - Agent-as-a-Judge
- Est mesuré l'alignement de Agent-as-a-Judge et de LLM-as-a-Judge avec le groupe d'experts
- ~~— Est mesurée la performance des agents~~

Détails du protocole







- Graybox signifie l'accès à toutes les traces de l'environnement pendant l'exécution
- Blackbox signifie un accès plus restreint aux données intermédiaires
- Human-as-a-Judge est uniquement en graybox
- Agent-as-a-Judge et LLM-as-a-Judge sont exécutés des deux façons

Alignment rates

	Moyenne
LLM-as-a-Judge 	69.94%
Agent-as-a-Judge 	87.49%
LLM-as-Judge 	70.49%
Agent-as-a-Judge 	89.61%
Moyenne des experts 	86.64%
Human-as-a-Judge 	94.44%

- Le consensus des experts est le plus performant
- Agent-as-a-Judge surpasse l'expert individuel
- La blackbox est plus efficace que la graybox

Alignment rates

	MetaGPT	GPT-Pilot	OpenHands
LLM-as-a-Judge 	84.15%	65.30%	60.38%
Agent-as-a-Judge 	88.15%	83.88%	90.44%
LLM-as-Judge 	68.86%	71.85%	70.76%
Agent-as-a-Judge 	92.07%	86.61%	90.16%
Moyenne des experts 	89.34%	84.88%	85.70%
Human-as-a-Judge 	95.08%	93.98%	94.26%

Quelques performances

97.72%

Du temps est économisé en préférant
Agent-as-Judge à Human-as-a-Judge

14h d'exécution

97.64%

Du coût financier est économisé en préférant
Agent-as-a-Judge à Human-as-a-Judge

210.65\$ dépensés

Perspectives d'avenir

Perspectives d'avenir

Testing des agents

- Création de jeux de tests similaires à DevAI pour des projets spécifiques
- Avènement d'outils Agent-as-a-Judge généraux
⇒ - cher + rapide + efficace

Amélioration des agents IA

- R&D sur des standards plus élevés
- Création d'agents de plus en plus autonomes

Merci pour votre écoute !