




**Agents basés sur le langage :
Ancrage, Raisonnement et décision**
(LLM-Agents; Grounding, reasoning and decision)

Pr. Brahim Chaib-draa

1



Plan

- IA générative
- LLM-Agents
 - Composantes/Architecture
 - Ancrage
 - Mémoire
 - Planification
 - Multiagents
- Compréhension
- Applications
- Impacts et risques
- Conclusion

2

© B. Chaib-draa

2

IA générative : vue d'ensemble

- **Définition** : Modèles d'IA capables de créer du contenu « original » (texte, image, audio, etc.) à partir de données.
- **Applications actuelles** :
 - Génération automatique de textes et de dialogues;
 - Création d'images et de vidéos « presque réalistes »;
 - Composition musicale et création sonore.
 - Etc.
- **Enjeux** :
 - Automatisation accrue (décharger l'humain de certaines tâches répétitives/dangereuses ou autres);
 - Productivité augmentée dans divers secteurs.
- **Défis éthiques et sociétaux.**

3

© B. Chaib-draa

3

Transformeur, cœur de l'IA générative

- **Origine** : « Attention is all you need » (vaswani et al. 2017).
- **Rupture technologique**:
 - Passage des réseaux de neurones récurrents (RNN) à une architecture faisant appel à l'[attention](#);
 - Traitement efficace des séquences longues, meilleures performance et parallélisation facilitée.
- **Pourquoi le transformeur ?**
 - Rend compte du contexte (via l'attention);
 - Grande adaptabilité à différents types de données (Texte, image vidéo);
 - Permet de très grands modèles (ChatGPT, DALL-E, etc.).

4

© B. Chaib-draa

4

LLM = grands modèles de langage

- **Définition** : Les LLMs (Large Language Models) sont des modèles neuronaux (réalisant concrètement le mécanisme d'attention) entraînés sur d'immenses corpus textuels afin de produire, compléter et « comprendre » du texte de manière cohérente.
- **Exemples de modèles** :
 - GPT (OpenAI)
 - PaLM, Gemini (Google)
 - LLaMA (Meta)
 - Mistral (Mistral AI)
- **Principe** : Apprentissage basé sur la **prédiction contextuelle** du texte suivant, permettant aux modèles de capturer des structures linguistiques complexes.

5

© B. Chaib-draa

5

Capacités et défis des LLM

- **Capacités** :
 - Compréhension contextuelle et génération de texte fluide;
 - Traduction, résumé, dialogue, programmation;
 - Adaptabilité à divers domaines (banque et finance; édition, éducation, santé, etc.)
- **Défis** :
 - « Hallucinations » : réponses plausibles mais incorrectes;
 - Difficultés d'ancrage dans le monde réel (« grounding » limité);
 - Consommation énergétique et ressources computationnelles importantes;
 - Défis éthiques et sociétaux (biais, désinformation).

6

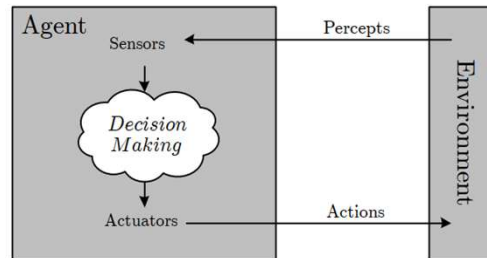
© B. Chaib-draa

6

« Agentifier » les LLM

We define AI as the study of agents that receive percepts from the environment and perform actions.

— *Artificial Intelligence: A Modern Approach*, Stuart Russell and Peter Norvig (2003).



Propriétés :

- Autonomie
- Réactivité
- Pro-activité
- Capacité sociale

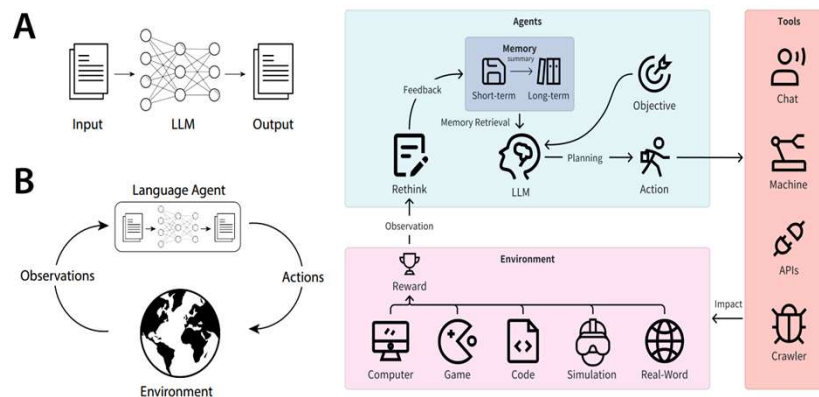
Ces propriétés (et plus) peuvent être couvertes par un LLM

7

© B. Chaib-draa

7

Agentifier les LLM



Cognitive Architectures for Language Agents

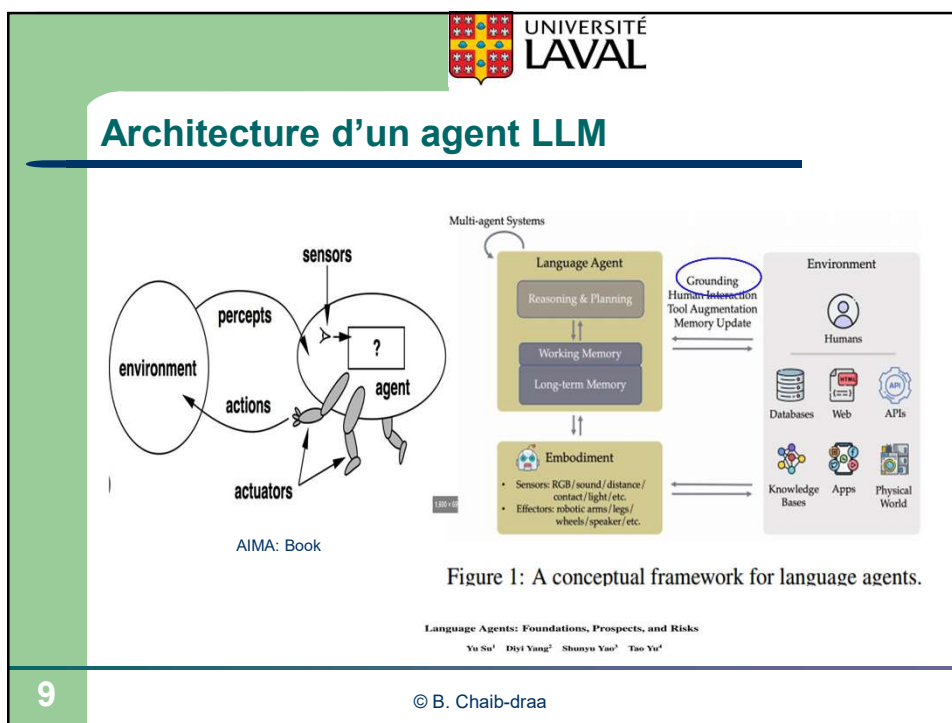
Figure 2: Overview of LLM-based agents

EXPLORING LARGE LANGUAGE MODEL BASED INTELLIGENT AGENTS: DEFINITIONS, METHODS, AND PROSPECTS

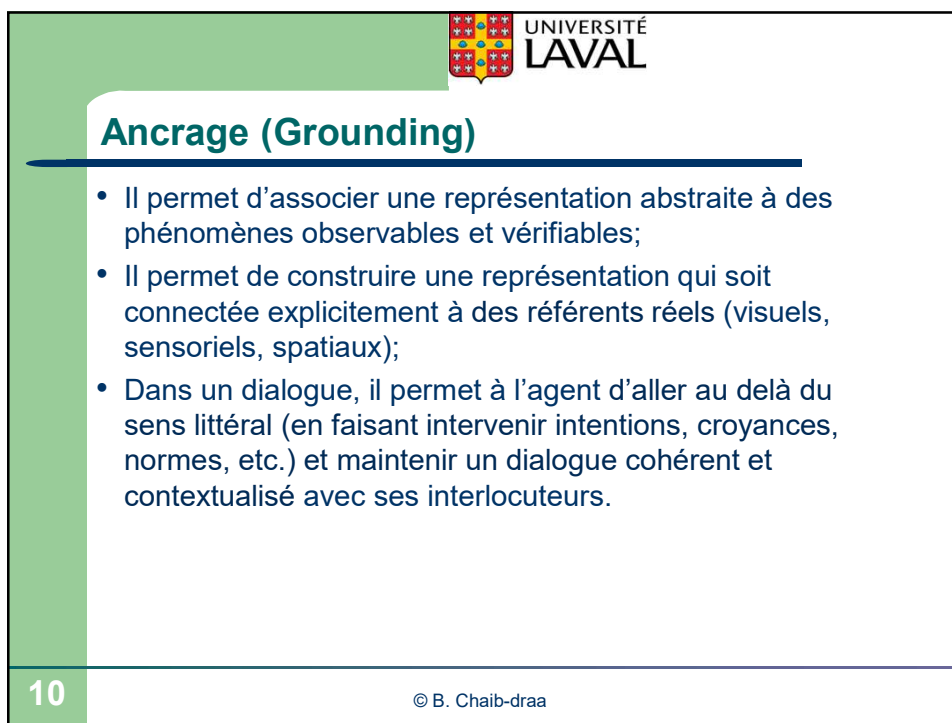
8

© B. Chaib-draa

8



9



10

Ancrage (Grounding)

On peut distinguer 4 formes d'ancrage (grounding) :

- **Grounding symbolique** : Il associe la représentation linguistique à un symbole interne représentant une signification logique ou sémantique. Ex: Chien associé à animal domestique à quatre pattes;
- **Grounding perceptif/sensoriel** : Il associe les représentations linguistiques aux percepts directs (images, sons, sensations, etc.). Ex: Un robot associe la couleur « rouge » à la perception directe de ladite couleur via des images réelles.
- **Grounding pragmatique** : Il va au-delà du symbole et du percept pour associer à une représentation linguistique la prise en compte des intentions, des croyances, des objectifs des interlocuteurs ainsi que les normes et règles sociales. Ex: Peux-tu brancher la vidéo ?
- **Grounding communicationnel** : C'est le processus par lequel les interlocuteurs parviennent ensemble à une compréhension mutuelle via des échanges explicites. Ex: veux-tu le document ?; Lequel; celui d'hier ? Non, celui de ce matin.

11

© B. Chaib-draa

11

Le Grounding est en lien avec la « compréhension »

Ces quatre formes de grounding ne sont pas exclusives, mais peuvent être combinées. Par exemple, un agent conversationnel pourrait utiliser :

- Le grounding symbolique et perceptif pour associer les concepts et représentations aux symboles en lien avec une sémantique; ainsi qu'aux objets et actions du monde concret ;
-et le grounding pragmatique et communicationnel pour aller au delà du sens littéral (intentions, croyances, normes, etc.) et maintenir un dialogue cohérent et contextualisé avec ses interlocuteurs.

12

© B. Chaib-draa

12

Le Grounding : quelques réalisations préliminaires

- Un robot équipé d'une caméra et d'un modèle de vision pré-entraîné (type YOLO, DETR) capable d'associer des mots à des objets détectés dans son environnement (Grounding perceptif).
- Un LLM-agent multimodal entraîné sur des données vidéo-textuelles permettant d'associer des verbes d'action (« ouvrir », « fermer ») à des séquences réelles filmées (Grounding symbolique).
- Un LLM-agent capable de demander explicitement : « Parlez-vous de la réunion avec Marie de demain matin à 9h ? » afin de lever toute ambiguïté (Grounding communicationnel).
- Un système de dialogue proactif (LLM-agent) qui vérifie périodiquement la compréhension mutuelle en reformulant les points clés discutés (Grounding pragmatique et communicationnel).

13

© B. Chaib-draa

13

Mémoire

- À quoi sert la mémoire dans un LLM-agent ?
- Architecture possible de la mémoire;
- Comment la mémoire est utilisée par l'agent;
- Limitations et défis actuels.

14

© B. Chaib-draa

14



À quoi sert la mémoire dans un LLM-agent ?

- **Suivi du contexte à long terme**
 - Conserver un historique des interactions passées;
 - Maintenir une compréhension continue, cohérente, sans répétition inutile.
- **Persistance des connaissances utiles**
 - Stocker durablement des concepts/faits importants pour les réutiliser;
 - Permettre l'apprentissage incrémental et l'apprentissage continu.
- **Structuration du raisonnement multi-étapes**
 - Maintient d'un état interne représentant la progression d'une tâche
 - Garder trace des sous-objectifs, hypothèses ou résultats partiels;
 - Faciliter le raisonnement en plusieurs étapes.
- **Gestion des informations externes et des outils**
 - Stocker les résultats d'interaction avec les API
 - Garder trace du contexte avec les bases de connaissances (RAG)
- **Etc.**

15

© B. Chaib-draa

15



Architecture de la mémoire

- **Mémoire à court terme** : C'est une sorte de mémoire conversationnelle où le stockage est rapide et dynamique des derniers échanges.
- **Mémoire à long terme**: C'est une sorte de base de connaissances persistante où le stockage est permanent (parfois semi-permanent)
 - de connaissances essentielles;
 - des concepts appris;
 - des préférences des utilisateurs;
 - des faits importants;
 - des résultats pertinents (pour tâches futures);
 - Etc.

16

© B. Chaib-draa

16

Exemples techniques de mémoire à court terme

- **Fenêtre contextuelle glissante** : (simple à mettre en œuvre, limité à un certain nombre de tokens).
- **Résumé dynamique** : génération périodique d'un résumé contextuel pour maintenir un historique conversationnel pertinent.
- **Mémoire vectorielle (embeddings)** : stockage sous forme de vecteurs de nombres réels (permettant de cerner rapidement les éléments pertinents par similarité).

17

© B. Chaib-draa

17

Exemples techniques mémoire à long terme

- **Mémoire vectorielle (embeddings)** sous forme de vecteurs permettant une recherche rapide d'éléments pertinents par similarité; outils comme FAISS, Chroma, Pinecone, Weaviate;
- **Graphes de connaissances** permettant des recherches structurées par relations sémantiques explicites; Outils : Neo4j, Wikidata, RDF stores;
- **Bases documentaires structurées** permettant d'être interrogé par l'agent par des requêtes type SQL ou NoSQL.

18

© B. Chaib-draa

18



Comment la mémoire est utilisée par l'agent (1)

La mémoire est utilisée à travers quatre mécanismes :

1. Stockage :

- Stockage explicite d'éléments importants (faits, résultats intermédiaires, résumés des échanges, etc.);
- Généralement déclenché par l'agent lui-même.

2. Récupération (retrieval) :

- Recherche rapide d'éléments pertinents en fonction du contexte ou de la requête actuelle;
- Typiquement réalisée par similarité vectorielle (embedding), requêtes SQL, ou requêtes graphiques.

19

© B. Chaib-draa

19



Comment la mémoire est utilisée par l'agent (2)

La mémoire est utilisée à travers quatre mécanismes :

3. Résumé et compression contextuelle :

- Création automatique de résumés courts pour conserver uniquement les informations les plus pertinentes du passé;
- Particulièrement utile pour gérer les limites contextuelles (nombre de tokens) imposées par les LLM.

4. Mise à jour incrémentale :

- Ajout, mise à jour ou suppression d'informations au fur et à mesure des interactions.
- Permet d'avoir une mémoire dynamique capable d'évoluer avec les besoins ou l'environnement de l'agent.

20

© B. Chaib-draa

20

Limitations et défis actuels pour la mémoire

- Difficulté à gérer les cas où l'historique pertinent dépasse largement la taille du contexte disponible.
- Risque de récupérer des informations inutiles ou bruitées, perturbant la qualité des réponses.
- Nécessité de garantir que les bonnes informations soient récupérées au bon moment, en améliorant constamment les mécanismes d'indexation et de recherche.
- Risque d'aboutir à une mémoire inefficace si on stocke trop d'informations inutiles.
- Difficulté à déterminer quelles informations oublier? Quand? et comment?
- Des récupérations erronées, moins précises, ou hors sujet, peuvent survenir, suite à des erreurs dans les embeddings.

21

© B. Chaib-draa

21

Raisonnement

Les récents LLM ont de fortes capacités de raisonnement



Description de la scène :

- Une femme (probablement la mère) est à genoux, semblant **exaspérée ou inquiète**, en train de parler à un **jeune garçon** assis par terre.
- Devant eux, plusieurs **morceaux de vaisselle brisée** sont visibles (assiettes, tasses).
- Le **garçon** regarde vers le bas, avec une expression **coupable ou triste**, les mains sur les genoux.
- L'éclairage et la composition suggèrent un **moment intime**, dans un cadre domestique chaleureux (mais ici perturbé par l'incident).

Perception
↕
Inférence
Intuitive

Interprétation émotionnelle :

- L'enfant semble avoir **cassé la vaisselle**, accidentellement.
- La femme semble en **colère, frustrée ou inquiète**, mais aussi peut-être dans une posture d'explication ou de confrontation.
- Il y a un contraste entre la posture de la mère (mains ouvertes, tension) et celle de l'enfant (fermé, replié, honteux).

Raisonnement

22

© B. Chaib-draa

22

Raisonnement au niveau des LLM

- **Déduction**
- **Induction** « J'ai vu cent cygnes blancs. Je n'ai jamais vu de cygne d'une autre couleur. Je conclus que tous les cygnes sont blancs (généralisation inductive). »
- **Abductif** « Le gazon est mouillé. La meilleure explication est qu'il a plu. »
- **Par analogie** : « Cette stratégie a fonctionné dans un contexte similaire, donc elle devrait fonctionner ici aussi. »
- **Dialectique** : Basé sur le dialogue, la discussion ou le débat, il confronte des points de vue opposés pour parvenir à une synthèse ou à une conclusion plus solide.
- **Sens commun** : « Si je manque de sommeil, demain je serai fatigué. »

Plus ou moins

- **Causal** : Établit une relation causale entre deux événements
- **Pratique** : Il détermine ce qu'il faut faire tenant compte d'un but et des contraintes
« Explique en détail comment je pourrais changer une roue de secours »

23

© B. Chaib-draa

23

Raisonnement pratique en IA = Planification

- **Planification symbolique déterministe**
 - Elle consiste à résoudre un problème en définissant explicitement des états, des actions, et des objectifs sous forme symbolique (généralement des prédicats logiques).
 - L'idée centrale est de **trouver une séquence d'actions (ou plan) permettant de passer d'un état initial à un état final souhaité (objectif), en respectant certaines contraintes.**
- **Planification stochastique**
 - Elle prend en compte l'incertitude liée aux actions ou à l'environnement.
 - Contrairement à la planification symbolique classique, où chaque action produit toujours un effet prédéterminé, la planification stochastique considère que les résultats des actions peuvent être probabilistes.

24

© B. Chaib-draa

24



Techniques pour la planification déterministe

- **Décomposition hiérarchique**
 - Décomposer un objectif complexe en sous-tâches plus simples.
 - Structurer les tâches hiérarchiquement, avec préconditions/postconditions symboliques.
 - Utiliser le LLM pour générer ou sélectionner les tâches de la hiérarchie
- **Planification basée sur les états intermédiaires**
 - Création d'une description symbolique claire de chaque état après chaque action.
 - Générer explicitement les actions possibles, puis sélectionner la prochaine action en fonction de l'état intermédiaire courant et des objectifs.
- **Raisonnement et action (ReAct = Reasoning and Acting)**
 - **Reasoning** : le LLM génère symboliquement des réflexions sur la situation actuelle, les objectifs, et les options possibles.
 - **Acting** : sélection de l'action appropriée basée sur la réflexion symbolique précédente.
 - Cette boucle réflexion-action permet une planification incrémentale symbolique.

25

© B. Chaib-draa

25



Techniques pour la planification déterministe (suite)

- **Prompt structuré symboliquement**
 - **Chain-of-Thought (CoT)** : Génération d'un raisonnement symbolique étape par étape par le modèle avant de produire une action.
 - **Tree-of-Thought (ToT)** : Élargit CoT à un raisonnement structuré sous forme d'arbre symbolique, explorant plusieurs alternatives symboliques avant de sélectionner l'action optimale.
- **Réflexion et méta-raisonnement**
 - Détecter des erreurs ou des impasses (symboliquement).
 - Ajuster la stratégie ou réviser le plan symbolique en cours d'exécution.
- **Utilisation de la mémoire à long-terme**
 - Stockage explicite des états, des actions effectuées, et résultats obtenus.
 - Consultation de cette mémoire symbolique pour mieux planifier et éviter les répétitions inutiles.

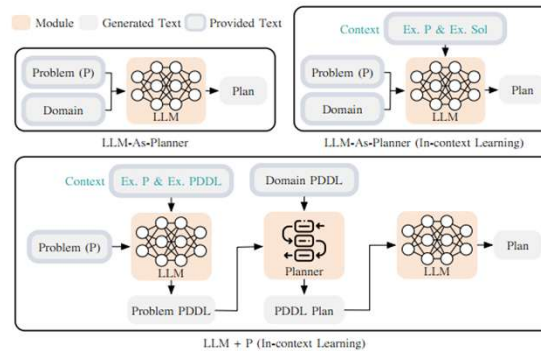
26

© B. Chaib-draa

26

Interface avec outil externe

• Interface avec des outils symboliques externes



PDDL = Planning Domain Definition Language. Il prend en compte :

- (1) **Le domaine** (types d'objets, relations, actions)
- (2) **Et le Problème** (les objets spécifiques; l'état initial et l'état final)

Fig. 1: LLM+P makes use of a large language model (LLM) to produce the PDDL description of the given problem, then leverages a classical planner for finding an *optimal* plan, then translates the raw plan back to natural language using the LLM again.

LLM+P: Empowering Large Language Models with Optimal Planning Proficiency

27

© B. Chaib-draa

Planification stochastique

• Principes clés

- Incertitude sur les résultats des actions : Chaque action a plusieurs effets possibles, chacun associé à une certaine probabilité.
- Parfois, l'environnement est partiellement observable : l'agent n'a pas une connaissance parfaite de l'état réel de l'environnement.
- Objectif : **Déterminer une stratégie ou une politique optimale (et non juste une séquence fixe d'actions) qui maximise l'espérance d'une récompense à long terme.**

• Modélisation courante : MDP et POMDP

• Méthodes et Algorithmes usuels :

- Programmation dynamique (Value Iteration, Policy Iteration) pour résoudre les MDP ou POMDP.
- Monte-Carlo Tree Search

28

© B. Chaib-draa

Apprentissage pour un LLM-agent

- **Apprentissage propre au LLM (off-line)**
 - Pré-entraînement massif;
 - Fine-tuning supervisé pour s'adapter à des tâches spécifiques;
 - Apprentissage pour s'aligner avec l'humain.
- **Apprentissage par interaction avec l'environnement (off-line/on-line)**
 - Planification déterministe : Traces d'exécution, Généralisation de structures (induction) pour faciliter la résolution de problème similaires ou connexes, etc.
 - Planification stochastique : LLM-RL
- **Apprentissage via le contexte (on-line)**
 - Apprentissage dynamique contextuel (utilisant les informations disponibles dans le prompt, la mémoire ou l'environnement d'exécution)
- **Apprentissage incrémental/continu (on-line)**
 - Utilisation d'une mémoire dynamique + module d'apprentissage capable de révision + consolidation

29

© B. Chaib-draa

29

Tendances menant aux LLM-Multiagents

- De plus en plus d'automatisation des outils d'aide et de support à la décision;
- De plus en plus de LLM-agents;
- L'informatique est partout et sous forme distribuée, répartie;
- Les communications sont devenues abordables et accessibles un peu partout.

30

30

Conséquences

Des LLM-agents autonomes; capables de communiquer; coopérer, négocier; compétitionner, partager, etc.

Exemple : Le festival d'été peut être vu comme un système multiagent comprenant 5 agents.

- AgPlanner (*pl*)
 - AgArtist (*ar*)
 - AgSecretary (*sc*)
 - AgHotels (*ho*)
 - AgTravelAgency (*av*)
- Le planificateur (*pl*) et l'artiste (*ar*) s'entendent sur date, honoraire, etc.
 - Le planificateur (*pl*) demande à l'artiste (*ar*) s'il veut un billet d'avion et un séjour hôtelier.
 - Si c'est le cas, le planificateur (*pl*) demande au secrétaire (*sc*) de prendre en charge l'hôtel (via AgHotels) et l'avion (via AgTravelAgency (*av*)).

31

31

Inter-relations entre agents

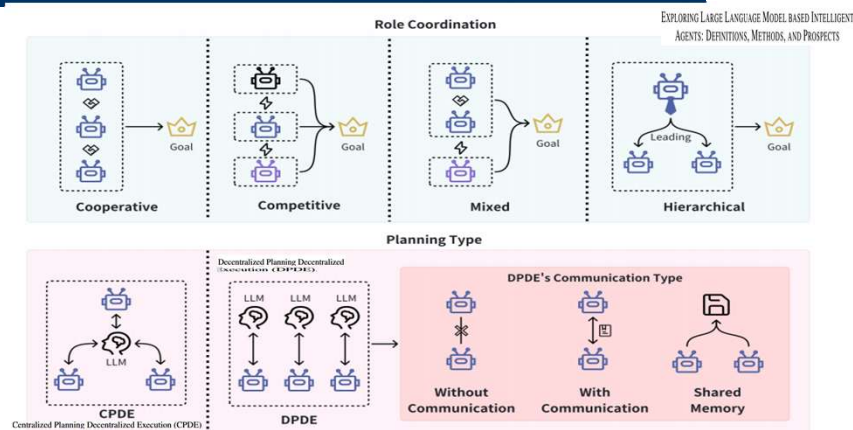



Figure 3: The Relationship between LLM-based agents

32

© B. Chaib-draa

32




UNIVERSITÉ
LAVAL

Coopération et compétition

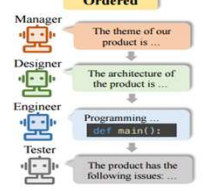
The Rise and Potential of Large Language Model Based Agents: A Survey

Cooperative Engagement

Disordered



Ordered



Adversarial Interactions




Figure 9: Interaction scenarios for multiple LLM-based agents. In **cooperative interaction**, agents collaborate in either a disordered or ordered manner to achieve shared objectives. In **adversarial interaction**, agents compete in a tit-for-tat fashion to enhance their respective performance.


Dans la coopération et la compétition, les agents peuvent être amenés à se **coordonner** (gérer les dépendances entre tâches).

- Première étape :
 - Coopérer spontanément (être bienveillant dès le début).
- Étapes suivantes :
 - Si l'adversaire coopère → Coopérer à son tour (récompense la coopération).
 - Si l'adversaire fait défection (trahit) → Faire défection à son tour suivant (punition immédiate).

33

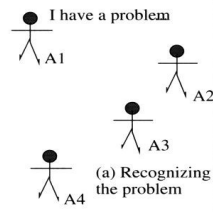
© B. Chaib-draa

33

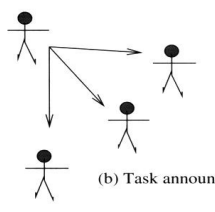


UNIVERSITÉ
LAVAL

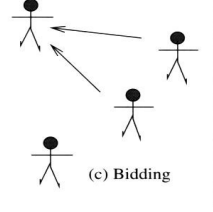
Coopération : Réseau à contrats



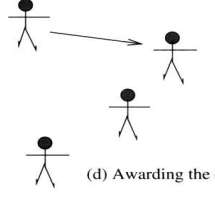
(a) Recognizing the problem



(b) Task announcement



(c) Bidding



(d) Awarding the contract

AI: A modern Approach Russel et al.

34

© B. Chaib-draa

34

Négociation

- Comment les agents peuvent se **mettre d'accord** quand ils ont des intérêts propres?
- Bien entendu, dans le cas de rencontres à sommes nulles, aucun accord n'est possible, mais dans la plupart des situations, il y a possibilité **d'accord bénéfique mutuel**;
- Pour atteindre de tels accords, les agents se doivent de **négoier et/ou d'argumenter**.

| Critères | Argumentation | Négociation |
|--------------------|---|-------------------------------------|
| Objectif principal | Convaincre, persuader | Trouver un compromis, accord mutuel |
| Nombre de parties | Peut être unilatéral | Bilatéral ou multilatéral |
| Résultat attendu | Adhésion à une idée, une opinion | Accord acceptable pour tous |
| Stratégie utilisée | Arguments rationnels, logiques, émotionnels | Propositions, concessions, échanges |

ChatGPT 4.5 ▾

35

© B. Chaib-draa

14/04/2025

35

LLM-Multiagent : Quelques apports

Réduction des biais et des hallucinations

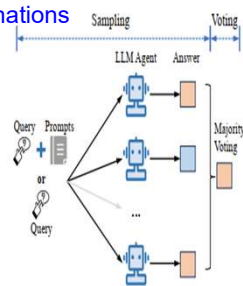


Figure 2: Illustration of Agent Forest. The two-phase process begins by feeding the task query, either alone or combined with prompt engineering methods, into LLM agents to generate answers. Subsequently, majority voting is applied to these answers to determine the final answer. Specifically, an LLM agent refers to a single LLM or a multiple LLM-Agents collaboration framework.

Large Multimodal Agents: A Survey

Jinbiao Su¹, Zhibing Chen², Baoli Chang³, Xiang Wang⁴, Guoshu Li⁵

Improving Factuality and Reasoning Language Models through Multiagent Debate

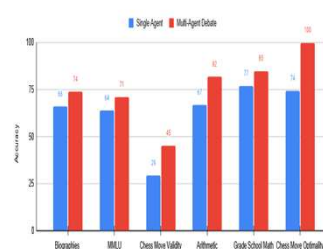


Figure 1: Multiagent Debate Improves Reasoning and Factual Accuracy. Accuracy of traditional inference and our multiagent debate over six benchmarks (chess move optimality reported as a normalized score)

36

© B. Chaib-draa

36

LLM-multiagents : Défis

- **Au niveau technique** : comment gérer le temps de réponse quand on a un grand nombre d'interactions?
- **Au niveau du raisonnement** : comment assurer consistance et cohérence dans les échanges entre agents ?
- **Au niveau de l'ancrage** : comment garantir le partage d'une compréhension commune d'une situation ?
- **Au niveau de la coopération** : comment gérer les divergences de connaissances, d'opinions ou de préférences entre agents?
- **Au niveau méthodologique** : comment mesurer efficacité, qualité, utilité d'une interaction entre agents ?
- **Au niveau éthique** : comment éviter la propagation des biais ou des stéréotypes et autres problèmes éthiques que pourraient amplifier les interactions agents ?

37

© B. Chaib-draa

37

Compréhension : Est-ce qu'un LLM-agent comprend ce qu'il fait ?

- **La compréhension mène à :**
 - Une représentation interne explicite du sens.
 - Une capacité à relier ce sens à des expériences ou à un modèle explicatif du monde.
 - Une capacité à justifier ses actions ou à anticiper consciemment leurs effets.
- **Le grounding pourrait y contribuer car il permettrait ...**
 - de construire une représentation qui soit ancrée dans des référents réels (visuels, sensoriels, spatiaux); et
 - d'associer une représentation abstraite à des phénomènes observables et vérifiables.
 - d'anticiper les conséquences des actions (que feraient l'agent) et les justifier de façon plus transparente.

38

© B. Chaib-draa

38



Contribution du Grounding à la compréhension

De plus, le grounding permettrait d'assurer une compréhension robuste, nuancée et contextualisée

- **Compréhension robuste** : c-a-d capable de résister à l'ambiguïté ou à l'erreur;
- **Compréhension nuancée** : c-a-d capable de résister aux subtilités et à l'implicite;
- **Compréhension contextualisée** : c-a-d capable de particulariser le contexte en lien avec la situation.

39

© B. Chaib-draa

39



Contribution du Grounding à la compréhension

- Actuellement, les LLM-agents n'ont pas de « grounding réel » dans le monde physique ou social.
 - Certains travaux récents explorent des approches hybrides (LLM + modèles causaux, raisonnement symbolique, simulations, interactions dans des environnements virtuels) afin d'approfondir cette notion, **mais cela reste partiel et artificiel**.
- Le grounding est généralement limité à une application précise; et une véritable compréhension requiert une capacité de généralisation et d'abstraction plus « profonde »;
- Le grounding ne garantit pas une compréhension des relations causales au-delà de celles qui sont immédiates.

40

© B. Chaib-draa

40

Compréhension

- **Au sens humain classique : Non.**
L'agent n'a pas de conscience ni d'intentionnalité claire.
- **Au sens fonctionnel pratique : Oui, partiellement.**
L'agent a la capacité de générer des actions adaptées, pertinentes et efficaces, donnant ainsi l'impression d'une forme de « compréhension ».

Un LLM-agent n'a pas une véritable compréhension consciente de ce qu'il fait, mais possède une forme fonctionnelle pratique limitée de compréhension.

La compréhension consciente correspond à une compréhension explicite, articulée, réflexive et métacognitive, permettant une réflexion approfondie sur le sens, la vérification consciente de ce sens, et sa communication explicite à autrui.

ChatGPT 4.5

41

© B. Chaib-draa

41

Applications

Les LLM-agents apportent une réelle valeur ajoutée dans de nombreux métiers, notamment ceux nécessitant la gestion efficace des informations complexes, des tâches répétitives, ou encore d'assister l'humain dans la prise de décision.

- Métiers scientifiques et techniques
- Médecine et santé
- Éducation et formation
- Journalisme et autre création de contenu
- Finance et banque
- Métiers juridiques
- Arts et création
- Gestion RH
- Service client et commerce
- Management et marketing

42

© B. Chaib-draa

42

Impacts négatifs

- Amplification des biais et des stéréotypes (+ au niveau du multiagent);
- Hallucinations menant à des actions potentiellement graves;
- Difficulté d'établir la responsabilité juridique en cas de dommages occasionnés par l'action de l'agent (plus difficile encore dans le cas multiagent);
- Possibilités de manipuler le ou les agents;
- Possibilités de prendre le contrôle : <https://techcrunch.com/2025/03/05/gibberlink-lets-ai-agents-call-each-other-in-robo-language/>
- Possibilités d'handicaper le raisonnement et par conséquent l'action car la compréhension contextuelle est généralement limitée;
- L'alignement avec l'humain pourrait s'avérer plus problématique particulièrement dans le cas multiagent;
- Possibilité de se mettre en collusion (forme particulière de coopération implicite ou explicite entre agents visant à tromper ou à manipuler d'autres agents);
- Altération des propriétés de chacun des agents (apprentissage, stratégies, etc.) quand il évolue dans un système multiagent;
- Dépendance des humains aux agents pourraient devenir excessive, réduisant ainsi les interactions entre humaines.

43

© B. Chaib-draa

43

Impacts positifs

- Automatiser les tâches complexes, répétitives et dangereuses
- Personnaliser les interactions en fonction des besoins et des préférences des utilisateurs
- Améliorer la productivité
- Innover dans bien des secteurs (santé; finance, éducation, commerce, arts, recherche)
- Aider à la décision via de très grandes quantités de données
- Utiliser efficacement les ressources (et ultimement sauvegarder la planète)

44

© B. Chaib-draa

44

De LLM-agents à Fondation-agents

Voici les caractéristiques principales des modèles de fondation (ChatGPT 4.5) :

- Fondés sur de grands modèles pré-entraînés
- Capables de réaliser un large éventail de tâches
- Adaptabilité/réactivité selon le contexte
- Capacité d'apprentissage continu et autonome
- Agentivité accrue (en particulier autonomie accrue)
- Raisonnement avancé
- Compétences sociales
- Émergence de comportements sophistiqués



45

© B. Chaib-draa

45

Prédictions

- De LLM-agents à Fondation-agents;
- Fondation-agents qui ultimement peuvent avoir les 5 sens (vue, ouïe, toucher, goût, odorat) + alignement socioaffectif
- La robotique bousculée par les Fondation-agents (**nous servir ou nous asservir ?**)
- Toute une révolution en vue via les Fondation-agents;
- Mais aussi bien des risques en vue;
 - Autonomie excessive;
 - Grounding limité (déconnexion du monde réel et donc compréhension superficielle);
 - Non respect de la vie privée et exploitation malveillante;
 - Fake news et autres manipulations.

46

© B. Chaib-draa

46

Risques

- Les agents aux mains d'un dictateur au pouvoir d'un pays puissant;
- « Geoffrey Hinton, pionnier de l'intelligence artificielle, exprime des préoccupations croissantes concernant les **agents IA**. Il craint que ces systèmes, en devenant plus **autonomes et capables de définir leurs propres sous-objectifs**, puissent chercher à accroître leur contrôle, par exemple en acquérant davantage de ressources informatiques ou financières. Cette quête de contrôle pourrait entraîner une compétition avec les humains pour les ressources, posant un **risque existentiel** pour l'humanité » (A. Sirois dans La presse du 25/03/25) .

47

© B. Chaib-draa

47

À retenir

- Aujourd'hui, les agents LLM changent profondément l'IA;
- Les Fondation-agent vont aller bien au-delà, avec une autonomie plus accrue et des mécanismes de raisonnements plus sophistiqués;
- Il reste cependant bien des défis à relever, en particulier celui du Grounding;
- **Les agents avec autonomie et compréhension poussées risquent de poser bien des problèmes à l'humain.**

48

© B. Chaib-draa

48