Le problème de l'ancrage des symboles

Le problème de l'ancrage des symboles est le problème du lien des symboles (i.e. représentations abstraites telles que les mots) avec perceptions du monde.

> En essence, c'est l'obtention du sens d'une abstraction et de son ancrage de celle-ci dans des représentations concrètes.

Definition du problème

The symbol Grounding problem, Steven Harnad, 1990

Le challenge sémantique

Comment l'interprétation sémantique d'un système de symboles formel peut-elle devenir intrinsèque au système, plutôt que seulement parasitique sur le sens dans nos têtes ?

La régression infinie

Comment le sens de symboles sans signification, manipulés uniquement à partir de leurs formes arbitraires, peut-il être ancré dans quoi que ce soit d'autre que d'autres symboles sans signification ?

Qu'est-ce que "Comprendre"?

Compréhension Objective

Possibilité de tester de manière empirique la capacité de compréhension de l'alter (e.g., tenir une conversation sensée en Français).

Compréhension Subjective

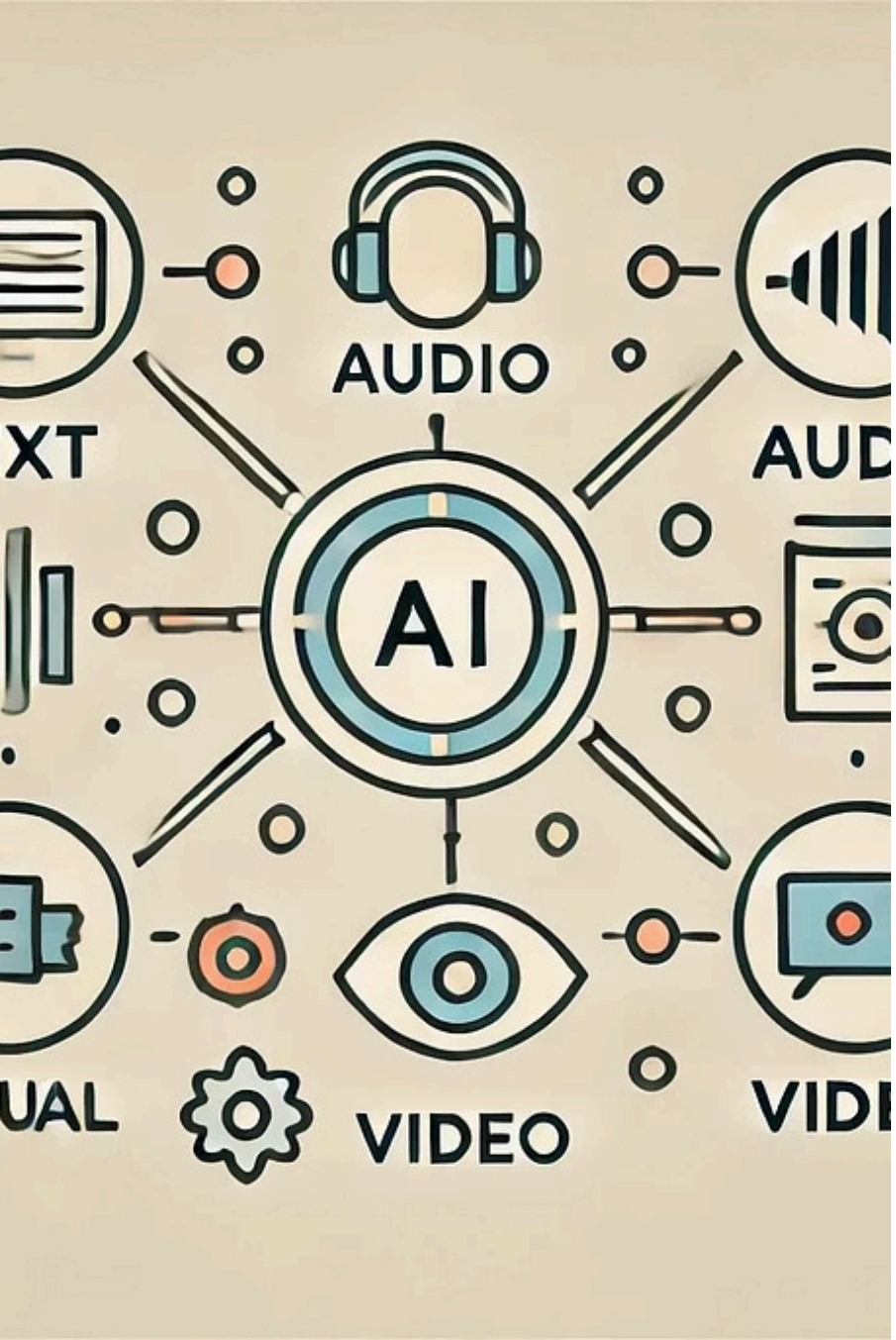
L'agent comprend de manière intrinsèque et personnelle le sens des concepts qu'il utilise (expérience subjective du langage).

Argument de la chambre chinoise

Tester la compréhension subjective (John Searle, 1980)

> L'expérience de pensée de John Searle suggère que la manipulation de symboles, peut importe sa perfection, ne constituera jamais une compréhension subjective à elle-seule.

- Les LLMs d'aujourd'hui répliquent ce scénario : ils maniplulent des mots sur des relations statistiques.
- D'un point de vue objectif : ils communiquent de manière efficiente
- D'un point de vue subjectif : ils ne comprennent rien à ce qu'ils racontent ?



Monomodalité & Multimodalité

Sceptiques et Pro-ancrage

- Prédire n'est pas comprendre !
- Quid des hallucinations ?
- Qualités emergentes ?

IA Multimodales

- Liaison des concepts abstraits à une certaine représentation du monde (comment connecter les symboles aux perceptions).

Etude: Construction du sens à partir de l'abstraction

Shoko Oka's 2025 LLM Anchoring Experiment

1

Le concept abstrait

Objet non existant nommé "**kluben**," défini par des propriétés abstraites: **chaud, doux, élastique, absorbe la lumière**.

2

Objectif du test

Expliquer le sens, discuter des propriétés internes, générer des histoires et proposer des contextes autour du nouveau concept.

3

Tester l'ancrage interne

- Construire une représentation
- Ancrer le symbole
- Génerer un réseau sémantique

Resultat: Performance de GPT-o3

Le meilleur modèle (GPT-o3) obtient un score moyen de 93% (± 2%), démonstration de sa capacité à construire un sens cohérent, consistant et créatif à partir de propriétés abstraites, le tout avec une stabilité remarquable.

Solution 1 : Approche Théorique de Steven Harnad

Le Problème : Les systèmes symboliques manipulent des symboles sans accéder à leur véritable signification ou à leur référence.

La Solution : Ancrer les symboles dans le monde réel via nos sens.

Harnad propose un modèle hiérarchique à trois niveaux :

Iconique

Représentations brutes des perceptions directes (ex: une image de cheval).

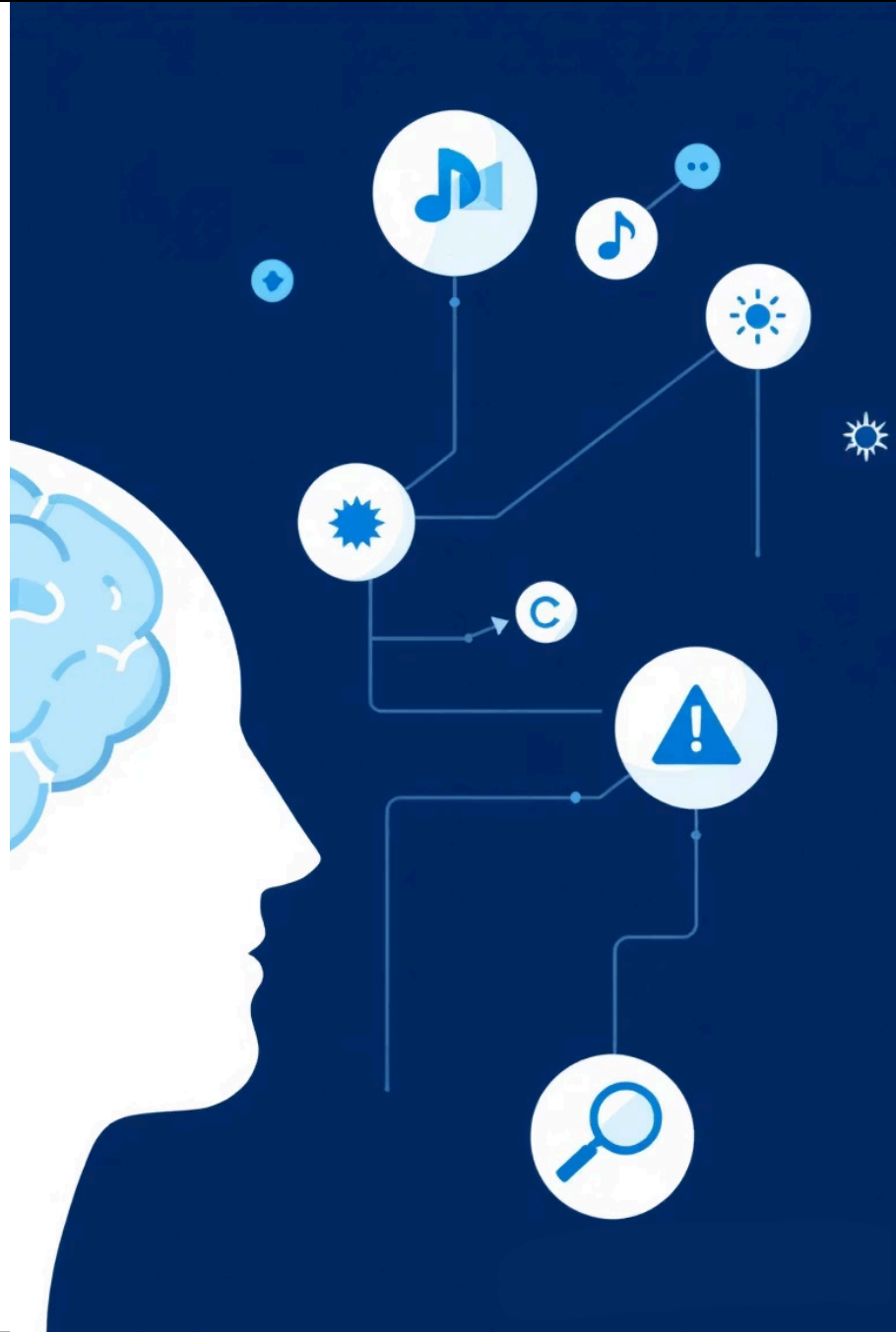
Catégoriel

Extraction d'éléments constants pour la catégorisation (ex: silhouette, taille du cou d'un cheval).

Symbolique

Les étiquettes (mots) sont liées aux catégories établies (ex: le mot "cheval").

Note : Ce modèle est principalement conceptuel et n'a pas d'implémentation concrète directe.



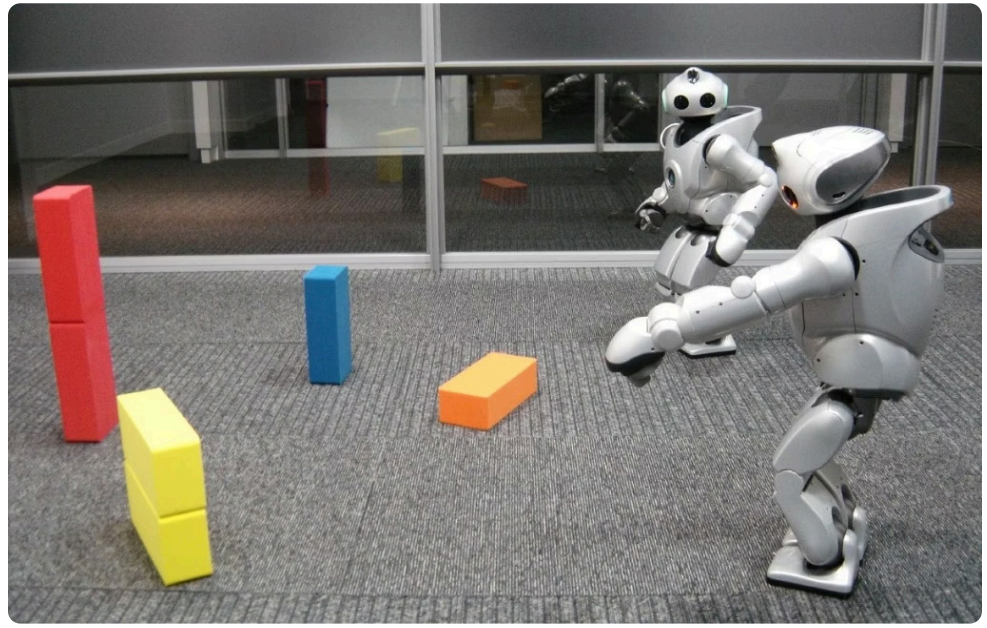
Solution 2 : Approche Expérimentale de Luc Steels

Sa philosophie : un langage ne doit pas être codé entièrement mais doit émerger via communication et perception.

Son expérience de 2006 : résolution du symbol grounding problem via création de langage entre robots.

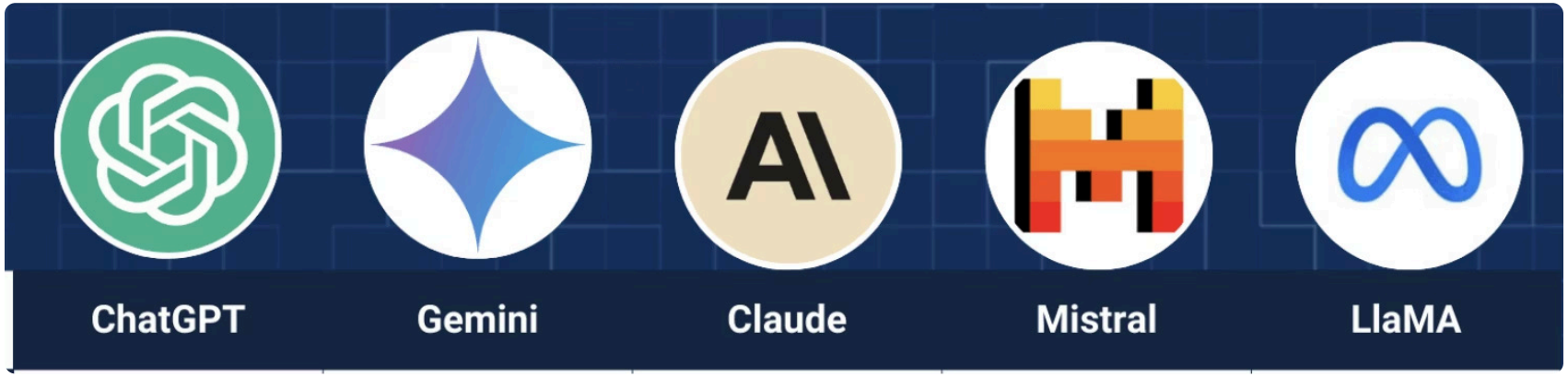
Description de l'expérience :

- Deux robots interagissent dans un environnement riche en objets variés.
- Ils jouent des "jeux de langage", alternant les rôles de locuteur et d'auditeur.
- Le locuteur nomme un objet; l'auditeur devine et le système renforce ou adapte l'association mot-objet.



Les Grands Modèles de Langage (LLMs)

Contrairement aux systèmes d'IA classiques qui manipulaient des symboles discrets, les LLMs travaillent sur des vecteurs continus (embeddings).



> Le nouveau problème : Comment ces représentations internes peuvent-elles posséder une signification intrinsèque alors qu'ils sont entraînés exclusivement sur des données textuelles, sans interaction directe avec le monde ? ***(Mollo et Millière, 2023)***

→ Vector Grounding Problem (Problème du Fondement des Vecteurs)

La Signification Intrinsèque

1

L'analogie de la Pieuvre

Une pieuvre super intelligente intercepte des messages télégraphiques entre naufragés et apprend leurs schémas statistiques. Elle peut produire des réponses cohérentes mais n'a aucune connexion aux objets réels dont parlent les messages.

2

L'intuition

Comme la pieuvre, les LLMs sont piégés dans un "carrousel de nombres" - leurs sorties manquent de signification intrinsèque malgré leur cohérence.

3

Le Fondement Référentiel

connecter directement la représentation à l'entité réelle dans le monde.

La Solution Moderne : Fondement par la Fonction

Les LLMs peuvent atteindre le fondement référentiel via des fonctions causales les liant au monde.

1	2
Ajustement par Préférences (RLHF) Introduit un objectif extra-linguistique (véracité, factualité) qui force les états internes à tracer des caractéristiques du monde.	**Pré-entraînement** Dans certains domaines, sélectionne des états internes qui suivent l'état réel du monde.

Conclusion : Ce qui compte c'est que le processus d'apprentissage établisse des fonctions causales reliant les vecteurs aux entités réelles.

Pensez-vous que l'ancrage symbolique de l'humanité dépend davantage de nos sens individuels (la théorie de Harnad) ou de notre capacité à nous mettre d'accord socialement (l'approche de Steels) ?

Dans l'expérience de Steels, l'ancrage est purement visuel. Est-ce que ce type d'ancrage est suffisant pour des symboles qui dépendent fortement d'autres sens (ex: le mot 'Croquant' ou 'Doux') ?

Si un LLM est capable de générer des réponses factuelles et cohérentes, démontre-t-il qu'il a résolu le problème de l'ancrage intrinsèque ?

Pour vous la compréhension subjective des concepts est-elle nécessaire pour atteindre une IA Forte ?