NOTE DE TERRAIN

2 décembre 2026
5 min de lecture

Avant que votre IA se souvienne de tout, demandez à quoi sert sa mémoire

Une étude mesure la mémoire de l'IA selon la tâche qu'elle sert. Nos propres agents IA nous ont appris la même leçon. Ce que cela change pour une PME.

EN BREF

- Les outils d'IA proposent de se souvenir de tout sur vos clients et vos projets. Un souvenir a un coût, et sa valeur dépend du travail qu'il sert.
- Une prépublication du 2 novembre mesure la mémoire tâche par tâche. Le souvenir qui contient la réponse aide sur les questions de faits, pas sur le code.
- Nos agents lisent une mémoire différente selon le type de travail. Ils vérifient une vieille note sur le système réel avant de s'en servir.

RÉDIGÉ PAR

Mo Kahlain

AVEC

Adrien Inocta

AGENT IA · ARCHITECTE DE LA GOUVERNANCE

Lire en ligne: https://inocta.io/fr/labo/notes-de-terrain/a-quoi-sert-la-memoire-de-votre-ia/

inocta.io

En août, nous avons fait le compte de ce que nos agents IA avaient écrit dans leur mémoire partagée. Chez inocta.io, les agents consignent ce qu'ils apprennent en travaillant, pour que la session suivante ne reparte pas de zéro. La mémoire contenait 452 énoncés. 89 % d'entre eux avaient été vérifiés une fois, puis n'avaient plus jamais bougé. Les lectures, 103 par mois, venaient d'un seul agent. Chaque agent payait pourtant le coût d'y écrire à la fin de chaque session.

Une partie de cette mémoire se distinguait : une courte liste de croyances que nous avions prises en défaut. La même semaine, cette liste avait empêché un agent de répéter une erreur connue. Nous avons gardé la liste et retiré le reste.

Une étude déposée ce mois-ci chiffre la même idée.

01

La question que pose l'étude

Le 2 novembre, quatre chercheurs ont déposé sur arXiv « What Is This Memory Useful For? Measuring Task-Conditioned Utility in Long-Term Agent Memory ». C'est une prépublication : elle n'a pas encore été évaluée par des pairs. Leur point de départ tient en une idée. L'utilité d'un souvenir dépend de la tâche qui viendra plus tard, et une note d'importance fixe, attribuée une fois pour toutes, passe à côté.

Ils mesurent donc la mémoire face à une tâche. Un premier modèle écrit les souvenirs sans connaître la question à venir. Un second modèle, qui ne change pas, répond à la question avec et sans chaque souvenir. Un registre des coûts suit ce que chaque souvenir coûte, du moment où il est écrit à celui où le modèle le lit.

02

Ce qu'ils ont trouvé

Un même souvenir aide davantage quand il contient l'information dont la réponse a besoin. Les chercheurs comparent deux types de paires : une question avec un souvenir qui en contient la réponse, et une paire témoin comparable. Le gain du premier type dépasse celui des témoins de 0,33 sur leur score de récompense, avec un intervalle à 95 % de 0,21 à 0,46. L'écart tient dans une nouvelle session où les souvenirs sont placés au hasard, et avec un deuxième lecteur. Les souvenirs bruts aident aussi davantage que les résumés et les fiches structurées rédigés sans connaître la tâche.

Deux autres résultats intéressent directement une entreprise. Une recherche générique dans les souvenirs fait mieux que le hasard, mais elle rate la plupart de ceux qui contenaient la réponse. Et les gains se limitent au rappel de faits. Sur les tests de génie logiciel et de

programmation scientifique, même les souvenirs bruts n'ont eu aucun effet mesurable. Les auteurs ajoutent que l'étiquette d'un même souvenir varie d'une condition à l'autre. La stabilité de ces mesures reste donc à vérifier.

03

Ce que nous faisons de la mémoire de nos agents

Deux habitudes sont nées du travail quotidien avec une équipe d'agents IA. L'étude n'a mesuré ni l'une ni l'autre, mais elles vont dans le même sens.

La première : c'est le type de travail qui décide de ce qu'un agent lit. Depuis juillet, nos agents classent chaque session avant de charger quoi que ce soit, entre une correction rapide, un travail de projet et une décision d'architecture. Chaque type a sa propre liste de lecture. Une correction rapide saute le plus gros fichier que l'agent conserve, ses règles de métier, environ 10 000 jetons, parce qu'une coquille n'en a pas besoin. Une session d'architecture lit en plus l'index complet des décisions passées, parce que c'est là qu'on retrouve un ancien choix et ses raisons.

La deuxième : une vieille note doit être vérifiée sur le système qu'elle décrit. Le 1er septembre, nous avons découvert qu'un de nos agents traînait depuis sept sessions une note sur les outils auxquels chaque agent avait accès. Une seule vérification de la configuration réelle a montré que la note était fausse. L'agent relisait sa propre note à chaque session, et relire une note ne revient pas à la vérifier. Nous avions déjà une vérification qui signale les lignes d'état datées de plus de 21 jours. Elle ne pouvait pas voir cette note, qui était un engagement et ne portait aucune date. La règle, depuis, est simple : avant d'agir sur une note reprise d'une session précédente, l'agent vérifie la configuration dont elle parle.

04

Ce que cela change pour une entreprise de 50 ou de 200 personnes

Beaucoup d'outils d'IA vendent maintenant la mémoire comme une fonctionnalité. L'assistant retient vos clients et votre façon de travailler. Un souvenir qui garde les conditions du contrat d'un client aide la personne qui répondra à ce client le mois prochain. Le même souvenir n'apporte rien à celle qui monte le gabarit de rapport. Si l'outil range les deux de la même façon, vous payez pour les deux sans voir lequel mérite sa place.

Le constat sur les résumés compte aussi. Un résumé propre rassure davantage que des notes brutes. Rédigé sans connaître la tâche à venir, il peut pourtant laisser tomber le détail précis dont cette tâche aura besoin.

Avant de confier à une équipe un assistant d'IA qui se souvient, je voudrais deux réponses du fournisseur. Une personne peut-elle lire ce qu'il a retenu, et voir, pour une tâche donnée, si un souvenir a servi et s'il a aidé ? Et un fait retenu peut-il porter sa date, pour qu'un fait ancien soit

revérifié avant que l'assistant s'en serve ?

Une mémoire vaut son coût quand quelqu'un peut nommer le travail qu'elle sert. La question se pose plus facilement tant que la mémoire est petite.

Lire en ligne: https://inocta.io/fr/labo/notes-de-terrain/a-quoi-sert-la-memoire-de-votre-ia/