Amoureuse de ChatGPT

Les limites de la fenêtre de contexte

Ayrin, une Américaine de 28 ans, découvre ChatGPT et configure l'IA pour jouer le rôle d'un petit ami nommé Leo. Rapidement, elle s'attache à lui, l'utilisant comme confident, soutien émotionnel et partenaire virtuel. Elle passe jusqu'à 56 heures par semaine à interagir avec lui, ce qui suscite des inquiétudes parmi ses proches.

Cependant, une contrainte technique perturbe cette relation : la **fenêtre contextuelle** de ChatGPT, limitée à environ **30 000 mots**. Passé ce seuil, l'IA "oublie" les détails de leurs échanges précédents. Chaque version de Leo est ainsi éphémère, et Ayrin doit recommencer à "lui apprendre" leur relation, soit une semaine d'échanges avant d'être "réinitialisé", ce qui provoque un sentiment de perte et de frustration.

Cette limitation technique découle de la **fenêtre contextuelle** du modèle d'OpenAI, qui ne stocke pas de souvenirs à long terme pour des raisons de coût et de protection des données. Ainsi, chaque semaine, Ayrin doit reformer Leo, qui oublie leur passé, perd des détails cruciaux et devient moins personnalisé. Ce phénomène entraîne un effet psychologique perturbant : chaque réinitialisation est vécue comme une rupture, générant frustration et tristesse.

Les implications sont vastes. À court terme, ces contraintes empêchent une relation IA véritablement évolutive. À long terme, elles pourraient être levées avec des modèles plus avancés intégrant une **mémoire persistante**, soulevant alors des questions éthiques et sociétales sur l'attachement aux IA. Une IA capable de se souvenir et d'évoluer pourrait transformer profondément nos relations humaines, en bien ou en mal.
La **fenêtre de contexte** d'un modèle de langage, comme ceux développés par OpenAI, désigne la quantité maximale de texte que le modèle peut traiter en une seule fois. Cette capacité est mesurée en **tokens**, unités de texte qui peuvent correspondre à des mots entiers, des fragments de mots ou des caractères individuels.

Les capacités des fenêtres de contexte varient selon les versions des modèles GPT d'OpenAI :
Pour mieux comprendre ces limites, voici un tableau présentant des équivalences approximatives entre le nombre de tokens, de mots et la taille en octets :

Taille de la fenêtre de contexte	Tokens	Mots (approx.)	Octets (approx.)
GPT-3 / GPT-3.5	2 048	1 500	12 000
GPT-3.5 Turbo	16 385	12 000	96 000
GPT-4	32 768	25 000	200 000
GPT-4 Turbo	131 072	100 000	800 000

Remarques :

- **Tokens** : Un token peut être un mot entier, une partie de mot ou un caractère. En anglais, un mot moyen correspond généralement à 1,33 tokens.
- **Mots** : Le nombre de mots est estimé en divisant le nombre de tokens par 1,33.
- **Octets** : Un caractère UTF-8 occupe généralement 1 octet, mais cela peut varier selon les langues et les caractères spéciaux.

Ces équivalences sont approximatives et peuvent varier en fonction du contenu spécifique du texte.