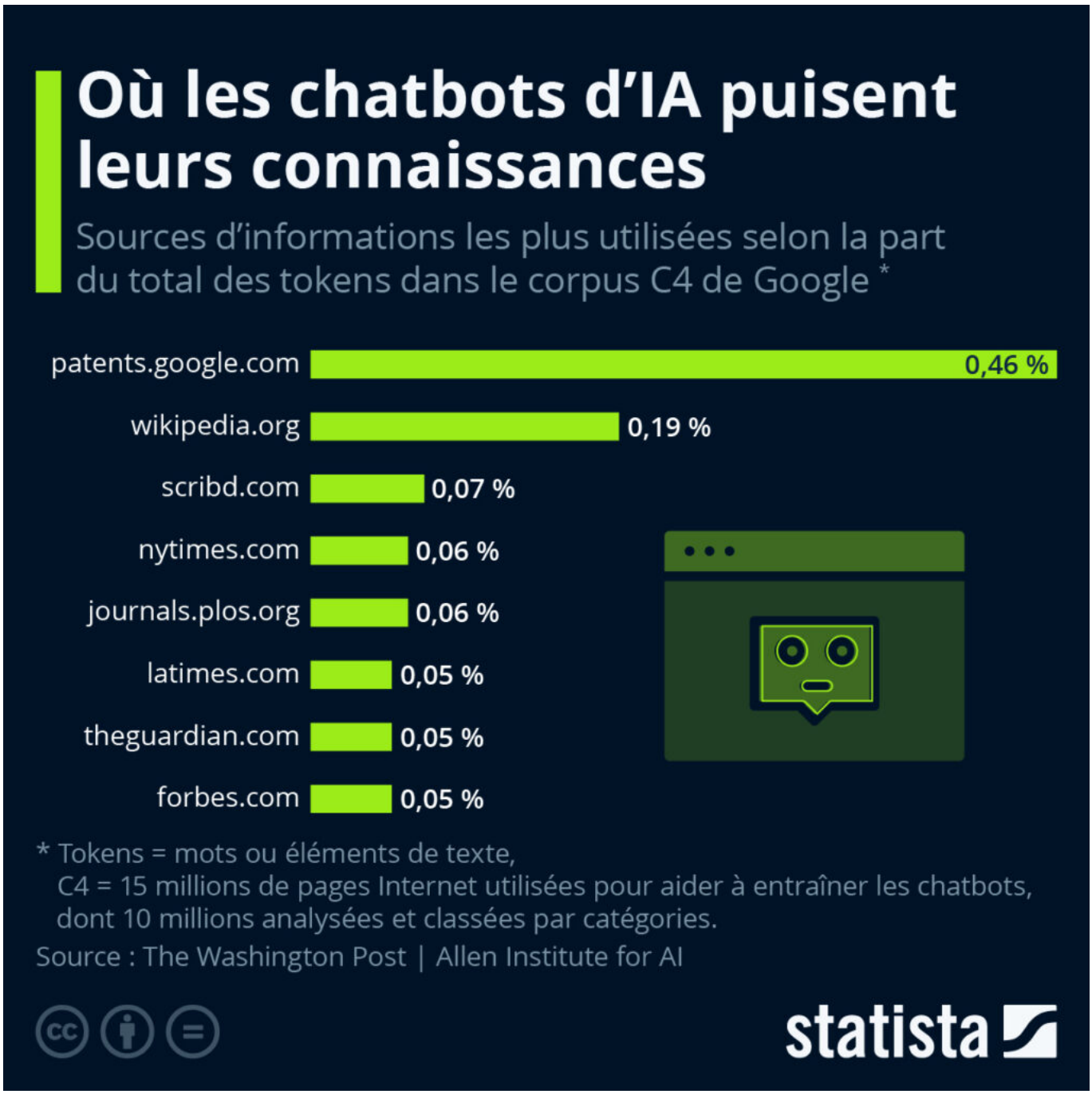
Les Petites Affiches de Vaucluse depuis 1839

Ecrit par le 1 février 2026

Où les chatbots d'IA puisent-ils leurs connaissances ?



Les Petites Affiches de Vaucluse depuis 1839

Ecrit par le 1 février 2026

Contrairement à la perception que l'on pourrait avoir, les chatbots (ou robots conversationnels) d'IA actuellement disponibles, comme ChatGPT d'OpenAI ou Bard de Google (dont l'intégration dans les services Google a été annoncée à la conférence I/O 2023), ne sont pas à proprement parler intelligents et ne possèdent pas de conscience propre. Les grands modèles de langage (LLM) sur lesquels ils s'appuient sont entraînés à partir d'informations déjà disponibles sur Internet. Ces connaissances sont ensuite restituées de façon à ce que le résultat résiste à un test de probabilité considérant tous les codes du langage naturel (orthographe, syntaxe, grammaire, etc.). Notre graphique, basé sur une étude publiée par le Washington Post, montre les sources d'informations qui sont les plus utilisées.

Le journal américain a analysé, en collaboration avec l'Allen Institute for AI, le corpus C4 publié par Google, une immense base de données regroupant 15 millions de sites web qui ont été utilisés pour entraîner des IA. Ils ont ensuite pu déterminer la répartition des « tokens » par source, c'est-à-dire la provenance des éléments de texte contenus dans le corpus. Avec 0,46 % du contenu, le moteur de recherche de brevets de Google, « patents.google.com », représente de loin la plus grande part. Cette plateforme indexe les brevets et demandes de brevet provenant du monde entier depuis 2006 et en regroupe aujourd'hui plus de 120 millions.

En deuxième position, on trouve « wikipedia.org » avec une part de 0,19 % du contenu, suivi de « scribd.com » avec 0,07 %. Ce dernier interpelle notamment en ce qui concerne le respect des droits d'auteur pour les textes générés par l'IA. Alors que les contenus de Wikipédia sont placés sous licences Creative Commons et sont diffusables librement, Scribd est un site de partage de documents en ligne sur lequel de nombreuses œuvres protégées ont été téléchargées. Plusieurs organes de presse tels que le New York Times, le Guardian et Forbes figurent également dans le top 8. Il est important de souligner que l'analyse du Washington Post ne prétend pas à l'exhaustivité ou à une exacte représentativité, car aucun modèle d'IA n'est entraîné sur la base d'un seul et unique corpus de données.

Alors que la réglementation et la législation en matière d'IA est plutôt à la traîne jusqu'à présent, certaines autorités nationales et internationales ont commencé à s'activer dans cette direction. L'Italie a été la première à agir : estimant qu'OpenAI avait enfreint le RGPD avec ChatGPT, le pays a décidé de bloquer son accès fin mars jusqu'à ce que la société se remette en règle. Dans l'Union européenne, les États membres discutent actuellement de l'introduction de l'AI Act, qui doit créer un cadre juridique transnational pour l'utilisation de l'IA dans l'UE.

De Tristan Gaudiaut pour Statista