

Les données, matière première de l'économie des « tokens » et enjeu majeur pour l'intelligence artificielle

Pas d'IA générative sans « jetons », ces unités – fragment de mot, caractère ou ponctuation – utilisées par les grands modèles de langage pour traiter un texte et générer des réponses. Une nouvelle économie est née.

Par Charles de Laubier

Publié le 21 septembre 2026 à 15h00

A l'ère de l'intelligence artificielle (IA), les grands modèles de langage (dits « LLM », pour large language model) sont entraînés sur d'immenses volumes de données textuelles afin d'apprendre à comprendre et à générer du langage naturel, avec l'objectif de produire des réponses aussi pertinentes, cohérentes et précises que possible. L'entraînement des IA génératives de l'écrit – ChatGPT d'OpenAI, Claude d'Anthropic, Gemini de Google ou encore Vibe (anciennement Le Chat) de Mistral AI – sur ces corpus de données nécessite de découper ces dernières (issues d'articles, de pages Web, de livres ou encore de documents) en jetons, tokens en anglais.

« Un token est un fragment de mot, l'unité élémentaire à partir de laquelle les modèles d'IA générative lisent et produisent du texte. En ce sens, il est fondamental et constitue littéralement la brique de base : un segment plus court que les mots, qui sert de base aux calculs », explique Thierry Poibeau, directeur de recherche au CNRS (laboratoire Lattice), rattaché à l'Ecole normale supérieure-Paris sciences et lettres (ENS-PSL). Egalement titulaire d'une chaire à l'institut Prairie-PSAI (Paris School of AI) en traitement automatique des langues et humanités numériques, il participera à [La Nuit des données Festival sciences et lettres, organisée par l'ENS, à Paris](#).

Un modèle d'IA textuelle (le fameux LLM) ne « lit » pas de lettres, de mots ou de chiffres comme nous, les humains ; il travaille uniquement avec des nombres, chacun d'eux représentant un identifiant unique de token. Cette conversion des textes en jetons dotés chacun de son propre identifiant numérique (ID) se fait automatiquement grâce à un « tokéniseur » (tel que BPE, WordPiece ou SentencePiece), lequel est un algorithme de segmentation statistique qui travaille à la découpe de textes. « Très souvent, un token égale un mot. Mais certains mots longs sont décomposés en plusieurs tokens. Par exemple, “livre” pourra être un seul token, là où “habitable” sera peut-être décomposé en “habitat” et “able” », explique David Louapre, chercheur et pédagogue en IA chez Hugging Face, et créateur de la chaîne Science étonnante sur YouTube (plus de 1,5 million d'abonnés), également participant à La Nuit des données Festival sciences et lettres.

Pas d'interopérabilité

Ainsi, le « vocabulaire » d'un modèle de langage d'IA textuelle est la liste complète de tous les tokens que le modèle connaît, autrement dit la table de correspondance entre tokens et nombres. Par exemple : pour l'intitulé « la tokénisation du monde », la découpe est « la », « token », « isation », « du », « monde », ce qui donnerait pour ChatGPT les ID 5547, 43900, 4824, 1146, 17290 ou, pour Le Chat, 4821, 18230, 4201, 712,

12903. Car pour un même découpage, les identifiants diffèrent d'un modèle à l'autre : il n'y a pas d'identifiant universel pour un même token, donc pas d'interopérabilité entre les LLM.

Sans tokénisation, le modèle ne peut pas « lire » ; sans détokénisation, il ne peut pas « écrire ». Et pour bien entraîner un modèle de langage, il lui faut beaucoup de données textuelles, la taille des corpus se mesurant en nombre de tokens. « Les chiffres exacts ne sont pas publics, mais on peut retenir comme ordre de grandeur plusieurs milliers de milliards de tokens, donc plusieurs milliers de milliards de mots », évalue David Louapre. Et cela peut coûter cher dans cette nouvelle économie agentique mondiale que constitue désormais la « token economy » (l'économie du jeton). « Le token sert d'unité de compte pour mesurer la taille des textes en entrée et sortie des modèles. De fait, c'est aussi devenu une unité de facturation pour les fournisseurs. Dans certaines formules, on paie "au token" produit, par exemple : 2 euros par million de tokens », précise David Louapre.

Ainsi, d'un modèle de langage à l'autre (de la famille GPT-5.6 chez OpenAI aux dernières générations de Gemini, Claude ou Mistral), le prix du token peut varier dans des proportions considérables. Facturés au million, les tokens consommés en entrée comme ceux qui sont générés en sortie peuvent rapidement alourdir la facture, particulièrement pour les entreprises qui traitent de très gros volumes de requêtes.

Texte caché

A cette tokénisation de l'économie s'ajoute une autre variable : le nombre de tokens nécessaires pour traiter un même texte dépend du tokéniseur utilisé, et de la langue. En général, un token représente quelques caractères en anglais ; en français, le découpage peut être plus coûteux en tokens, notamment en raison de la structure de la langue et de ses caractères accentués. Autrement dit, derrière chaque requête adressée à une IA générative se cache un compteur de tokens – et donc, potentiellement, une facture.

« Tous les grands acteurs du domaine facturent au nombre de tokens produits. Ce qui est moins visible, c'est que les modèles génèrent aujourd'hui beaucoup de texte caché, ou intermédiaire : pour analyser une question et construire une réponse, ils passent par de très nombreuses étapes (de "raisonnement"), qui consomment elles aussi des tokens, et qui font donc monter la facture. C'est une manière ingénieuse de facturer le raisonnement complet, et pas seulement la réponse finale, infiniment plus courte », souligne Thierry Poibeau, qui vient de publier *Les Mots sans les choses. Philosophie, éthique et impact des grands modèles de langue* (Editions Rue d'Ulm, 258 pages, 22 euros). Ces textes cachés rendent la « tokenomics » (étude économique des tokens) très incertaine : « Il suffirait que d'autres acteurs facturent moins, ou autrement, pour ébranler le modèle des principaux fournisseurs, dont les besoins financiers sont gigantesques et toujours croissants », prévient Thierry Poibeau.

Mais la tokénisation s'arrête là où commencent les IA non textuelles, comme Stable Diffusion de Stability AI pour la génération d'images, Sora d'OpenAI pour la vidéo ou encore Suno de la société éponyme pour la musique. Ces modèles utilisent d'autres méthodes que les jetons. Des recherches sont aussi en cours afin de se passer des tokens – même pour les IA textuelles – et de travailler directement sur les octets, ou encore des unités conceptuelles ou latentes. L'unité de facturation des IA pourrait s'en trouver profondément chamboulée.

Cet article est réalisé par Le Monde dans le cadre d'un partenariat avec [La Nuit des données Festival sciences et lettres, organisée par l'Ecole normale supérieure \(ENS\)](#), le 25 septembre 2026.

La Nuit des données Festival sciences et lettres

Intelligence artificielle (IA), santé, climat, géopolitique, physique, philosophie... C'est un dialogue inédit entre les disciplines pour explorer les données et l'IA sous toutes leurs formes, qui est proposé au grand public lors de La Nuit des données Festival sciences et lettres, organisée par l'Ecole normale supérieure (ENS). Lors de cet événement, dont Le Monde est partenaire, sont prévus plus de 70 conférences, débats, tables rondes,

ateliers, expositions et spectacles. Parmi les nombreux intervenants figurent Philippe Aghion (Prix Nobel d'économie), Joëlle Barral (Google DeepMind), Anne Bouverot (Conseil de l'IA et du numérique), Stéphane Mallat (Collège de France), Marc Mézard (physicien, ancien directeur de l'ENS), Joëlle Pineau (Cohere), Frédéric Worms (directeur de l'ENS).

Les thèmes abordés lors de cette Nuit des données seront multiples : « De la donnée à la découverte en sciences de la Terre », « Découvrir les molécules et matériaux de demain grâce à l'intelligence artificielle », « La naissance de la robopsychologie », « Deviner l'Univers grâce aux données qu'il nous envoie », « L'éducation nationale a des données : un trésor pour la recherche et la société », « Big Data, Big Brother. Comment les Etats gouvernent par les données »...

Vendredi 25 septembre, de 17 heures à minuit, à l'Ecole normale supérieure, 45, rue d'Ulm, Paris 5e. Gratuit [sur inscription](#).

Charles de Laubier