

intel

Petits modèles, grandes ambitions

*L'opportunité des IA génératives
redéfinie grâce à l'essor des petits
modèles de langage*



Découvrez comment les Petits Modèles de Langage (SLM) révolutionnent le paysage de l'IA

Les Petits Modèles de Langage, ou Small Language Models (SLM) en anglais, sont des modèles d'IA capables de traiter, comprendre et produire du contenu en langage naturel. Leur taille est bien plus réduite que celle des Grands Modèles de Langage, ou Large Language Models (LLM en anglais). Les SLM comportent généralement moins de 3 milliards de paramètres, contre des centaines de milliards pour les LLM. Ils sont parfois qualifiés de « modèles compacts ».

Le terme « paramètres » désigne les variables internes qu'un SLM ajuste pendant son apprentissage pour prédire et générer du texte. Il s'agit principalement des poids des connexions entre les neurones artificiels.

Caractéristiques techniques :

Tout comme les modèles de langage de grande taille, les SLM s'appuient sur l'architecture Transformer.

Cette dernière permet un traitement parallèle des séquences d'entrée, source considérable d'amélioration des performances par rapport aux réseaux neuronaux traditionnels. Les SLM tirent également profit de techniques de compression de modèle (distillation des connaissances, élagage, quantification...) pour atteindre une taille allégée, sans que leurs performances ne soient fondamentalement affectées. Par ailleurs, les SLM sont souvent entraînés sur des jeux de données ciblés, propres à un domaine spécifique plutôt que sur des ensembles massifs de données textuelles (livres, articles, forums, réseaux sociaux...).

Caractéristiques des SLM	Bénéfices
Fonctionnement sur du matériel standard (CPU)	Investissements maîtrisés
Réponses en temps quasi réel	Latence réduite
Excellents résultats sur des tâches bien définies (classification de texte, résumé de documents, traduction, génération de code...)	Précision / ciblage
Possible déploiement on-premise	Confidentialité

Avantages :



Ciblage et spécialisation : entraînés uniquement sur des données spécifiques à leur domaine d'expertise, les SLM excellent dans leur domaine de spécialisation (par exemple le domaine médical ou le secteur juridique).



Rapidité et latence améliorée : les SLM offrent des temps de réponse plus rapides grâce à leur complexité réduite. Leur petite taille permet une inférence quasi immédiate sur du matériel standard, avec une latence nettement inférieure à celle des LLM équivalents.



Réduction des coûts : du fait de leur taille réduite, les SLM requièrent moins de ressources pour être entraînés et ajustés. Cela se traduit par une baisse significative des coûts en termes de calcul, notamment pour des déploiements sur des infrastructures internes ou à l'edge.



Empreinte carbone optimisée : un SLM consomme souvent moins d'énergie pour son entraînement et son utilisation qu'un LLM, ce qui diminue d'autant son impact environnemental. Ces petits modèles sont ainsi considérés comme des options à faibles émissions pour les organisations cherchant à réduire leur empreinte carbone liée au numérique.

Les SLM offrent une multitude de cas d'usage qui peuvent transformer votre entreprise

Les cas d'usage permis par les SLM sont multiples. En voici quelques-uns ci-dessous.



Chatbots et assistants virtuels :

Dans le cadre d'un service client, les chatbots alimentés par des SLM peuvent répondre aux questions fréquemment posées, diriger les utilisateurs vers les ressources appropriées et leur fournir une assistance. Un chatbot médical peut ainsi expliquer des termes simples tandis qu'un chatbot Ressources Humaines peut renseigner les collaborateurs sur leurs congés payés. Les assistants virtuels vont quant à eux plus loin en intégrant la compréhension du langage naturel (NLP), l'exécution de commandes complexes et l'adaptation aux préférences individuelles.

Systèmes de recommandation : traditionnellement, les systèmes de recommandation reposent sur des algorithmes classiques, des modèles vectoriels ou des graphes. Les SLM permettent d'apporter une couche de compréhension linguistique (préférence, contexte utilisateur) et d'expliquer ou justifier une recommandation (via du texte généré). Ils peuvent être utilisés pour réaliser des recommandations de contenus dans une application métier, suggérer des options dans une interface utilisateur, ou bien encore proposer des articles similaires ou complémentaires aux clients d'un site e-commerce.

Computer vision : bien que les SLM soient principalement conçus pour le traitement du langage, ils peuvent jouer un rôle complémentaire important dans les systèmes de « computer vision », notamment pour les tâches combinant analyse d'image et traitement de texte. Ils peuvent notamment analyser et générer du texte à partir d'images (lecture de formulaires, tickets, notes manuscrites, affiches...) et interpréter les anomalies détectées par un modèle de vision et fournir une explication textuelle claire pour aider les opérateurs humains (maintenance prédictive). En combinant un SLM avec un système d'indexation visuelle, il est possible de rechercher des images dans une base de données en utilisant des requêtes en langage naturel.

Génération de code : certains SLM comme Microsoft Phi-2, StarCoder2 3B, DeepSeek-Coder 1.3B sont entraînés sur des corpus spécifiques de langages de programmation (Python*, JavaScript*, etc.) et de bibliothèques, ce qui leur permet de générer des extraits de code pertinents. Concrètement, ils peuvent procéder aux tâches suivantes : complétion de code, correction d'erreurs et refactoring, génération de tests automatisés, production de documentation technique.

Automatisation de tâches spécifiques (résumé, traduction...) : le déploiement d'un SLM permet d'automatiser certaines tâches ciblées de traitement du langage naturel (NLP), telles que la synthèse de contenu (résumé), la traduction, la reformulation ou la génération de textes courts (mails, réponses, suggestions, etc.). Les SLM peuvent aussi être utilisés pour la classification de contenu (tri automatique de tickets support, catégorisation d'articles...).

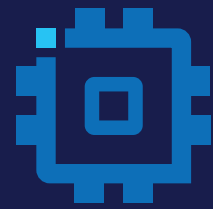


Intel rend la technologie des SLM accessible au plus grand nombre et ouvre la voie à de nouveaux usages dans la santé, le droit, les médias, entre autres.

Adel Chaïbi
(Ingénieur Machine Learning, Intel)

Explorons quelques-uns des SLM les plus innovants déjà disponibles

Voici un panorama des principaux SLM disponibles sur le marché. Il met en exergue leurs caractéristiques techniques et l'importance que revêt l'open source dans ce domaine.



L'open source dans les SLM est synonyme d'ouverture et d'innovation accélérée

Adel Chaibi
(Ingénieur Machine Learning, Intel)

Mistral et ses « Ministraux » :

Mistral AI* a annoncé en octobre 2024 la sortie de ses « Ministraux » : Ministral 3B et Ministral 8B. Ces SLM ont été conçus pour fournir une solution en matière de calcul et de faible latence, pour des applications critiques telles que la traduction sur appareil, les assistants intelligents sans internet, l'analytique locale et la robotique autonome.



Gemma 2-2.6B : après avoir sorti des versions de son SLM open source Gemma 2 dotées de 9 et 27 milliards de paramètres, Google* a lancé mi-2024 une déclinaison à 2,6 milliards de paramètres. Ce SLM a bénéficié de la technologie de distillation des connaissances qui lui permet de tirer parti des enseignements de ses prédécesseurs plus volumineux tout en conservant une taille réduite.

Claude 3.5 Haiku : Claude 3.5 Haiku est un SLM développé par Anthropic*. Il peut traiter jusqu'à 200 000 tokens en entrée et a été optimisé pour fournir des réponses quasi instantanées. Cela le rend particulièrement adapté aux applications nécessitant des interactions rapides, comme les chatbots ou les systèmes de service client.

Phi-2 de Microsoft* : développé par Microsoft Research, Phi-2 est un SLM doté de 2,7 milliards de paramètres. Entraîné sur 1 400 milliards de tokens, il inclut des ensembles de données synthétiques et du contenu web axés sur le raisonnement logique, les connaissances générales et les activités quotidiennes. Il peut être intégré dans des pipelines NLP (Natural Language Processing) ou utilisé pour la génération de texte, la complétion de code ou la création de chatbots personnalisés.

Llama 3.2 1B et 3B : en septembre 2024, Meta AI a annoncé la sortie de la famille de LLM Llama 3.2. Parmi ces modèles figurent Llama 3.2 1B et 3B. Avec 1 milliard de paramètres, Llama 3.2 1B est conçu pour fonctionner en local sur des appareils très contraints (mobiles, microcontrôleurs avancés...). Quant à Llama 3.2 3B (3 milliards de paramètres), il est adapté au suivi d'instructions (prompts) et à la synthèse de textes ou à leur réécriture.

TinyLlama : SLM basé sur l'architecture de Meta LLaMA, TinyLLaMA a été développé par la communauté open source et non par Meta* directement. Il est optimisé pour être extrêmement compact (1,1 milliard de paramètres). Ses cas d'usage idéaux sont les chatbots légers (assistants intégrés dans des applications mobiles) ou le traitement de texte local (résumé, correction grammaticale...).

Qwen 2.5 (Alibaba) : dans la série de modèles Qwen 2.5, développée par Alibaba Cloud*, figurent des modèles de différentes tailles, allant de 0,5 à 72 milliards de paramètres. Les modèles à 0,5 et 1,5 milliard de paramètres sont conçus pour fonctionner localement ou sur des machines à faibles ressources.

StableLM-3B : SLM développé par Stability AI*, StableLM-3B comprend 3 milliards de paramètres. Il a été pré-entraîné sur 1 000 milliards de tokens issus de divers ensembles de données.

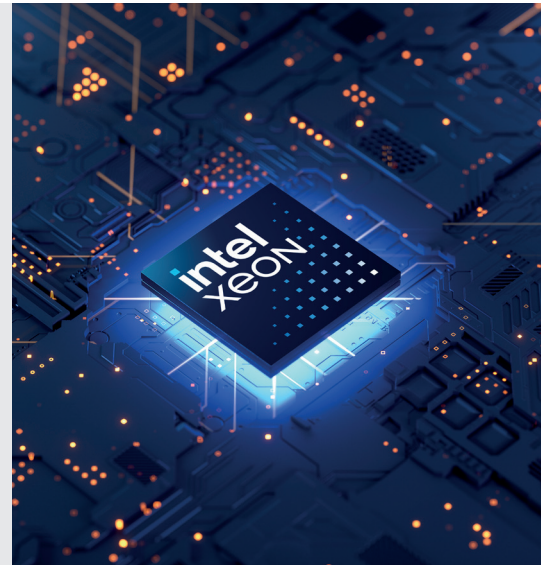
* Les autres noms et marques peuvent être revendiqués comme la propriété de tiers.

Découvrez comment déployer un SLM peut être simple et économique pour votre entreprise

L'un des principaux atouts des SLM réside dans leur faible empreinte matérielle. Contrairement aux LLM classiques qui nécessitent des clusters GPU et une infrastructure cloud coûteuse, les SLM peuvent fonctionner sur des appareils standards (PC portable, serveur local, smartphone ou microcontrôleur haut de gamme...).

À titre d'exemple, LLaMA 3.2 - 1B est conçu pour tourner sur des dispositifs très contraints, tels que des smartphones ou des objets connectés. Les modèles comme Mistral 3B, Gemma 2B ou Qwen-1.5B peuvent de leur côté être exécutés sur un CPU sans carte graphique dédiée, dès lors qu'ils sont quantifiés, c'est-à-dire transformés pour utiliser moins de bits par paramètre, dans le but de réduire leur taille mémoire et les besoins en calcul. Cela permet un déploiement sur site, mobile ou embarqué, dans des environnements industriels (véhicules, IoT, santé) sans dépendance au cloud.

Le déploiement d'un SLM est économiquement accessible. À l'inverse des LLM, il n'exige pas de serveurs haute performance ni d'infrastructure cloud. Côté licence, la majorité des modèles récents sont open source (sous licence Apache 2.0 ou équivalente). En termes de délais, un prototype fonctionnel peut être mis en place entre quelques heures et quelques jours, selon le niveau de personnalisation requis.



Le développement et le déploiement de modèles d'IA linguistique reposent traditionnellement sur des LLM et des serveurs équipés de GPU. Cependant, les coûts et l'infrastructure nécessaires à ces solutions peuvent être prohibitifs pour de nombreuses organisations. Pour déployer efficacement des applications comme les chatbots, certains architectes préfèrent ainsi les SLM.

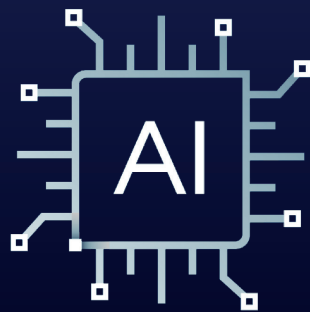
Pour optimiser la rentabilité, ces architectes explorent les possibilités d'exécution de charges de travail SLM sur des CPU, que ce soit dans un centre de données ou à l'edge. Combinés ensemble, les SLM et les CPU proposent une solution légère et économique pour l'implémentation de chatbots. L'utilisation de CPU plutôt que de GPU pour les petits modèles linguistiques aide en effet à réduire les coûts, la complexité et la consommation de ressources. Par exemple, les serveurs utilisant des processeurs Intel® Xeon® peuvent permettre d'exécuter des SLM sur une architecture CPU à un coût raisonnable et avec une latence limitée.



Les SLM facilitent l'adoption de l'IA en réduisant la dépendance aux infrastructures lourdes. Sur CPU, avec l'architecture Intel Xeon, ils s'exécutent localement, en datacenter ou dans le cloud.

Au-delà du matériel, la couche logicielle intermédiaire - bien que discrète - joue un rôle structurant : les accélérations sont intégrées via des bibliothèques standardisées portées par la Fondation UXL (Unified Acceleration Foundation).

Adel Chaïbi (Ingénieur Machine Learning, Intel)



**Vous
souhaitez
en savoir
plus sur les
technologies
Intel dédiées
à l'IA ?**

Adoptez les SLM pour une IA plus simple et efficace

En combinant spécialisation, performance ciblée et sobriété computationnelle, les Small Language Models (SLM) constituent une réelle alternative aux LLM.

Leur adoption permet aux organisations de déployer des solutions d'IA linguistique efficaces, économiques et compatibles avec des contraintes d'infrastructure locales ou embarquées.

Vous souhaitez en savoir plus sur la manière dont les Petits Modèles de Langage (SLM) peuvent transformer votre infrastructure IT ? Nos experts sont à votre disposition pour discuter de vos besoins spécifiques et vous accompagner dans l'intégration de ces solutions innovantes. Ensemble, faisons de l'intelligence artificielle un levier de performance et d'efficacité pour votre organisation.



intel

Avis et avertissements

Les technologies Intel peuvent nécessiter du matériel, des logiciels ou l'activation de services compatibles. Aucun produit ou composant ne peut être totalement sécurisé en toutes circonstances. Vos coûts et résultats peuvent varier. © Intel Corporation. Intel, le logo Intel, Xeon et les autres marques d'Intel sont des marques commerciales d'Intel Corporation ou de ses filiales. Les autres noms et marques peuvent être revendiqués comme la propriété de tiers.