Communiqué de presse
10 mars 2025

INTELLIGENCE ARTIFICIELLE ET NLP : LA SUITE D'ENCODEURS EUROBERT FRANCHIT UN NOUVEAU CAP

Entraînée sur 5000 milliards de tokens, cette suite propose des modèles souverains, open source et délivrant les meilleures performances de représentation textuelle sur les langues européennes ainsi que pour les tâches liées aux mathématiques et au code.

La collaboration entre le laboratoire MICS de CentraleSupélec, Diabolocom, Artefact et Unbabel, soutenue par l'expertise technologique d'AMD et du CINES, a abouti à la publication du meilleur modèle multilingue de représentation textuelle, brique fondamentale pour la recherche d'information (RAG), la classification et l'estimation de qualité (de résumés, traductions).

Ces types de modèles sont indispensables en traitement automatique des langues (NLP) et figurent parmi les plus téléchargés sur Hugging Face depuis de nombreuses années. Leur capacité à capturer avec précision le sens et le contexte des phrases, offrant une compréhension linguistique fine et approfondie, sont essentielles au développement d'applications d'intelligence artificielle avancées. Ce nouveau modèle EuroBERT est disponible le 10 mars 2025 sous licence Apache 2.0 sur la plateforme Hugging Face.

Le projet de recherche a été mené par Nicolas Boizard, doctorant Cifre chez Diabolocom, avec la contribution majeure d'Hippolyte Gisserot-Boukhlef, doctorant Cifre chez Artefact, et de Duarte Alves doctorant à l'Instituto Superior Técnico (IST), dans le cadre des recherches initiées par Pierre Colombo, professeur associé à CentraleSupélec, et placés sous la supervision de Céline Hudelot, directrice du MICS, et André Martins, professeur associé à IST. Les résultats sont détaillés dans un article publié sur arXiv le 10 mars 2025 : https://arxiv.org/abs/2503.05500

Un nouveau saut technologique franchi en matière d'encodage de textes

EuroBERT se distingue des autres encodeurs actuellement disponibles sur cinq points :

- Il est souverain et entièrement open source, tant pour son code source que pour ses ensembles de données.
- Il prend en charge 8 langues européennes majeures ainsi que 7 langues extra-européennes parmi les plus parlées au monde.
- Entraîné sur 5000 milliards de tokens, soit deux fois plus de données que les *encoders* classiques ou des modèles génératifs tels que Llama 2 (2000 milliards de tokens), EuroBERT offre ainsi des capacités optimales sans coût d'utilisation supplémentaire.
- La famille EuroBERT constitue la meilleure base pour des tâches de recherche d'information (RAG), de classification et d'estimation de qualité (de résumés, traductions).

- Il excelle dans des tâches jusque-là délaissées comme le traitement de données mathématiques ou de langages de programmation.
- Il se décline en 3 tailles de modèles (210M, 610M et 2,1B) offrant un équilibre optimal entre vitesse, qualité et coût, adapté aux contraintes des entreprises utilisatrices.

EuroBERT transforme ainsi les applications de traitement automatique des langues reposant sur des représentations de phrases, telles que l'analyse de texte, la recherche d'information, la classification ou l'extraction d'information.

La force et la valeur ajoutée de la recherche partenariale

Comme pour les modèles CroissantLLM et EuroLLM publiés sur Hugging Face en 2024, c'est grâce à une étroite et riche collaboration public-privé, ancrée dans l'écosystème de Paris-Saclay et élargie à l'échelle européenne, que ces avancées scientifiques ont été rendues possibles. Les équipes du MICS, de l'IST, de Diabolocom, d'Artefact et d'Unbabel ont travaillé ensemble dans le cadre des trois thèses en cours, ainsi qu'avec le supercalculateur français Adastra propulsé par la technologie AMD Instinct™ Accelerators and AMD EPYC™ processors.

Reconnu à l'échelle mondiale pour son excellence scientifique en mathématiques et informatique, le laboratoire MICS de CentraleSupélec pilote et mène de multiples travaux, programmes et projets de recherche en collaboration avec des partenaires privés et publics qui sans cesse repoussent encore plus loin les limites de l'intelligence artificielle. Diabolocom, avec son produit d'aide à la relation client, a apporté son expertise du traitement de langage, intégré dans leur produit. Artefact, leader européen du conseil en IA et data, a amené son expertise multi-secteurs et sa vision transversale sur les nombreuses applications déployées en entreprise. Enfin, Unbabel, leader tech en traduction automatique, a apporté son savoir-faire en IA multilingue.

« *Un mois après le sommet pour l'action sur l'IA qui s'est tenu à Paris, nous sommes particulièrement enthousiastes d'annoncer la disponibilité d'EuroBERT. Cette famille de modèles "encoder" pour les langues européennes est à ce jour la plus complète et la plus performante pour gérer les tâches au niveau des documents. Dans le paysage actuel de l'IA, les modèles "encoder" seuls sont souvent négligés malgré leur importance dans les applications de NLP. Par exemple, BERT - introduit en 2017 - bénéficie aujourd'hui de près de 5 millions de téléchargements par mois sur Hugging Face, dépassant LLaMA et d'autres modèles similaires* », souligne Céline Hudelot, Professeur à CentraleSupélec et directrice du laboratoire MICS.

Avec la création début 2025 de son centre de recherche Diabolocom Research, l'entreprise Diabolocom se dote de nouveaux atouts pour apporter des réponses concrètes et efficientes aux besoins du marché en systèmes IA fiables, souverains et performants.
"*La collaboration multidisciplinaire et la contribution à des projets open source sont au cœur de notre stratégie pour demeurer à la pointe de l'innovation. EuroBERT, qui est le dernier projet mené au sein de notre centre de recherche, répond à plusieurs limitations des encodeurs existants. Celui-ci contribuera à l'enrichissement fonctionnel de plusieurs de ses solutions comme la recherche d'information automatique, la classification automatique, les systèmes agentiques, etc.*", précise Frédéric Durand, Président et fondateur de Diabolocom.

Artefact s'est de son côté engagée dans la recherche en IA à travers son centre de recherche inauguré il y a un an.
"L'objectif étant de développer et de diffuser des modèles utiles et utilisables pour des applications concrètes en entreprise, l'ensemble de nos publications et algorithmes sont open source. L'amélioration aux modèles d'encodage de documents que représente EuroBERT ouvre des perspectives pour rendre plus performantes et pertinentes la classification et l'organisation des

documents, la recherche intelligente d'informations ou encore la NER (Identification automatique d'entités nommées). L'angle pris d'analyser un document existant et non de le générer, couvre un besoin récurrent et essentiel pour l'analyse de textes en entreprise", ajoute Emmanuel Malherbe, Directeur de la Recherche d'Artefact.

Quant à Unbabel, première plateforme d'opérations linguistiques alimentée par IA : *"EuroBERT représente une avancée majeure dans l'IA multilingue. Les modèles encodeurs sont depuis longtemps un atout caché du NLP, offrant une compréhension linguistique approfondie essentielle aux applications d'IA performantes. Contrairement aux approches purement génératives, les encodeurs excellent dans la capture du sens et du contexte, des éléments clés pour des systèmes multilingues précis et évolutifs. Chez Unbabel, nous avons une expertise solide non seulement dans le développement de solutions LLM génératives, comme nos modèles de pointe Tower, mais aussi dans la création de solutions à base d'encodeurs de référence, telles que Comet et CometKiwi. Le lancement d'EuroBERT intervient à un moment clé, comblant le manque de modèles encodeurs multilingues entraînés avec les avancées clés des modèles génératifs. Cette avancée représente une étape supplémentaire dans la construction d'une infrastructure essentielle au renforcement de notre souveraineté en IA en Europe et sommes fiers d'y contribuer à travers des des projets comme EuroBERT et EuroLLM, qui renforcent les capacités européennes et sécurisent notre avenir numérique commun",* ajoute Nuno Miguel Guerreiro, chercheur chez Unbabel.

Ce projet fut également rendu possible grâce aux GPU de pointe AMD Instinct™ MI300A Accelerators, intégrés à Adastra, le supercalculateur hyper-efficace français.
Pour Julien Ruiz, directeur France d'AMD, "*le développement d'EuroBERT marque une étape importante dans nos efforts pour améliorer les capacités de traitement du langage naturel pour les langues européennes conduites en France. En utilisant les GPU MI300 d'AMD et leur architecture de mémoire unifiée, il a été possible d'atteindre des performances et une efficacité sans précédent. Ce projet illustre l'engagement d'AMD en faveur de l'innovation et de l'excellence dans le domaine de l'intelligence artificielle*."

Ont également participé au développement d'EuroBERT des équipes de l'Université Grenoble Alpes, CNRS, LISN, d'Illuin Technology, de l'IRT Saint-Exupéry et du CINES.

A propos de CentraleSupélec - www.centralesupelec.fr
CentraleSupélec est un établissement public à caractère scientifique, culturel et professionnel, né en janvier 2015 du rapprochement de l'Ecole Centrale Paris et de Supélec. Aujourd'hui, CentraleSupélec se compose de 4 campus en France (Paris-Saclay, Metz, Rennes et Reims). Elle compte plus de 5 400 étudiants, dont 3 800 élèves ingénieurs, et regroupe 18 laboratoires ou équipes de recherche. Fortement internationalisée (25 % de ses étudiants et près d'un quart de son corps enseignant internationaux), l'école a noué plus de 170 partenariats avec les meilleures institutions mondiales. Ecole leader dans l'enseignement supérieur et la recherche, CentraleSupélec constitue un pôle de référence dans le domaine des sciences de l'ingénierie et des systèmes. Elle a cofondé l'Université Paris-Saclay en 2020 et préside le Groupe des Écoles Centrale (CentraleSupélec, Centrale Lyon, Centrale Lille, Centrale Nantes et Centrale Méditerranée) qui opère les implantations internationales (Pékin (Chine), Hyderabad (Inde), Casablanca (Maroc)).

A propos du Laboratoire MICS - Site web ici
Créé au début des années 2000, MICS rassemble la recherche en Mathématiques et Informatique de CentraleSupélec. Au cœur des technologies numériques, ses thématiques concernent la modélisation, la simulation, l'analyse et l'optimisation de systèmes complexes, qu'ils proviennent du monde industriel, du vivant, des marchés ou de l'information et des réseaux. Le laboratoire MICS est structuré en 6 équipes de recherche portant des objectifs scientifiques communs et en un axe transverse en Intelligence Artificielle.

Contacts presse :
Claire Flin : clairefline@gmail.com – 06 95 41 95 90

Marion Molina : marionmolinapro@gmail.com - 06 29 11 52 08

A propos de Diabolocom - www.diabolocom.com

Depuis plus de 20 ans, Diabolocom révolutionne les interactions client avec sa solution cloud CCaaS (Contact Center as a Service) enrichie d'une IA générative propriétaire. Automatisation intelligente, meilleure joignabilité et analyses fiables : des outils sont mis à disposition des équipes de service client et commerciales pour réussir. L'IA, conçue pour la relation client, intègre des fonctionnalités telles que la transcription en temps réel, l'analyse de satisfaction et des recommandations d'actions, tout en réduisant les tâches répétitives. Résultat : des interactions hyper-personnalisées, une fidélisation accrue et des ventes optimisées. En permettant une visibilité complète sur chaque interaction, la solution aide des entreprises leaders comme Carrefour, Air Liquide, Meilleurtaux et Leboncoin à transformer leur relation client dans plus de 60 pays.

Présente en Europe, en Amérique du Nord, au Brésil et au Moyen-Orient, Diabolocom accompagne les organisations dans l'amélioration de la relation client à l'échelle internationale.
En 2025, Diabolocom a lancé son laboratoire de recherche, Diabolocom Research, avec pour mission de relever les défis liés à la conception de systèmes responsables, fiables, éthiques et performants pour les centres de contact. Pour cela, nous développons des technologies de pointe en technologie de la parole, traitement du langage naturel, IA conversationnelle et optimisation hardware-algorithme.

Contacts presse :
Nada Nachit : nada.nachit@diabolocom.com

A propos d'Artefact - www.artefact.com

Artefact est une société de conseil et d'ingénierie française spécialisée en data et IA, et leader en Europe. Basée à Paris, elle est aujourd'hui présente dans 23 pays sur tous les continents et compte 1 600 collaborateurs.
Sa mission est d'aider les entreprises à exploiter tout le potentiel de l'IA et de la data en développant des solutions sur-mesure adaptées à leurs enjeux métiers. En tant que pionnier dans ce domaine, elle allie expertise technologique et excellence opérationnelle, en collaborant avec les plus grands acteurs du marché. Ses clients couvrent tous les secteurs clés de l'économie – industrie, retail, luxe, grande consommation, santé, finance – et incluent des grands groupes internationaux.
Au-delà du conseil, elle s'engage activement pour une IA éthique et accessible. Elle a lancé la "School of Data" pour favoriser la reconversion vers les métiers de la tech, créé un centre de recherche en IA à Paris et Shanghai.

Contacts presse :
Astrid Calippe : astrid.calippe@artefact.com