Communiqué de presse
12 décembre 2024

# EuroLLM-9B marque un nouveau jalon dans la souveraineté européenne des modèles de langage

***L'équipe du projet de recherche européen EuroLLM auquel participe le laboratoire MICS de CentraleSupélec (Université Paris-Saclay) aux côtés d'autres partenaires - Unbabel, l'Instituto Superior Técnico, l'Instituto de Telecomunicações, l'Université d'Édimbourg, Aveni, Equall, l'Université d'Amsterdam, Naver Labs et Sorbonne Université -, vient de publier un article de recherche détaillant les avancées obtenues avec EuroLLM-9B.***

L'Europe franchit aujourd'hui une étape majeure dans le développement de modèles de langage multilingues avec la publication d'**EuroLLM-9B**, le modèle de référence le plus avancé de sa catégorie conçu sur le continent.

"*Issu du projet de recherche collaboratif EuroLLM, ce modèle, entièrement open-source et « open-weight », vient répondre aux besoins des langues européennes et rivalise avec les approches internationales les plus performantes*", précise le Dr Pierre Colombo du laboratoire MICS de CentraleSupélec - Université Paris-Saclay. EuroLLM-9B est disponible en accès libre sur Hugging Face (pré-entraîné : utter-project/EuroLLM-9B ; modèle post-entraîné : utter-project/EuroLLM-9B-Instruct).

**Des performances de pointe, multilingues et adaptées aux langues européennes**
Le modèle EuroLLM-9B prend en charge les 24 langues officielles de l'Union européenne ainsi qu'un ensemble de 11 autres langues stratégiques et commercialement importantes, parmi lesquelles l'arabe, le catalan, le chinois, le galicien, l'hindi, le japonais, le coréen, le norvégien, le russe, le turc et l'ukrainien.

Cette couverture linguistique étendue répond aux besoins de diversification linguistique de l'Union européenne et vise à réduire la dépendance à des modèles principalement centrés sur l'anglais. Les résultats sur les benchmarks multilingues montrent qu'il surpasse largement les autres modèles européens de taille équivalente et se montre compétitif face à des modèles non européens reconnus tels que Gemma-2-9B.

**Une architecture robuste et un entraînement à grande échelle**
EuroLLM-9B est doté d'un tokenizer optimisé pour les langues européens et a été pré-entraîné sur près de 4.000 milliards de tokens, puis affiné progressivement en trois phases (pré-entraînement initial, phase d'annealing, puis phase d'annealing vers zéro). Cette approche progressive a permis d'améliorer sans cesse la qualité des données et de l'entraînement, pour aboutir à un modèle robuste et flexible.

Le projet a bénéficié de l'infrastructure de calcul européenne EuroHPC, et tout particulièrement de la puissance du supercalculateur MareNostrum5. L'entraînement a mobilisé environ 400 GPU Nvidia H100, grâce à un accès extrême-échelle obtenu dans le cadre d'EuroHPC, soutenu par la Commission européenne.

**Post-entraînement et adaptation aux usages réels**
Après son pré-entraînement, EuroLLM-9B a fait l'objet d'un post-entraînement (instruction tuning) visant à le spécialiser dans le suivi d'instructions complexes, le dialogue multi-tour et l'adaptation à divers cas d'usage. Cette étape a été réalisée en recourant exclusivement à des jeux de données publics, garantissant à la fois transparence et reproductibilité. Les résultats sont particulièrement remarquables en traduction multilingue, où EuroLLM-9B surpasse des modèles de référence tels que Gemma-2-9B–IT ou Aya-expanse-8B.

**Ouverture, transparence et conformité avec les réglementations**

Dans la continuité des principes du projet EuroLLM, tous les éléments d'EuroLLM-9B sont mis à disposition de la communauté en open-source, des poids du modèle jusqu'aux jeux de données utilisés. Cette transparence s'inscrit dans un effort plus large de conformité avec le futur AI Act européen, afin de garantir une IA respectueuse des valeurs et des réglementations de l'Union.

**Perspectives**

EuroLLM-9B ne constitue qu'une première étape dans la mise au point de modèles européens multilingues performants, souverains et adaptés aux besoins du continent. Pour aller plus loin encore, l'équipe de recherche vient de se voir attribuer une bourse spécifique de 5 M€ de la part du consortium de supercalculateurs EuroHPC pour créer un modèle multimodal européen d'intelligence artificielle.

"*La synergie internationale qui s'est mise en œuvre dans le cadre du projet EuroLLM nourrit déjà la conception de modèles plus puissants encore. L'équipe de recherche publiera prochainement un nouveau rapport technique qui détaillera les choix de données, les configurations de modélisation, ainsi que les orientations futures pour le développement d'une nouvelle génération de modèles linguistiques européens*", se réjouit Pierre Colombo.

Les membres du laboratoire MICS impliqués dans le projet EUROLLM font partie de l'équipe Gemila. Sous l'impulsion du Dr Pierre Colombo, expert international en LLMs, l'équipe participe très activement à l'avancée des connaissances sur les modèles de langage et les modèles de fondation dans le respect des principes de l'AI Act et des principes de souveraineté, chers à CentraleSupelec.

A propos du laboratoire MICS

Créé au début des années 2000, MICS (anciennement MAS) est le laboratoire de recherche en mathématiques et d'informatique de CentraleSupélec au sein de l'Université Paris-Saclay. Ses travaux de recherche concernent l'analyse et la modélisation de systèmes et de données complexes, qu'ils soient issus de l'industrie, des sciences de la vie ou des marchés financiers, des technologies de l'information ou des réseaux. L'intelligence artificielle est un axe de recherche transverse du laboratoire avec des travaux reconnus en IA pour la santé, en apprentissage pour les données non-structurées (images, textes, documents), en IA de confiance et de la décision.

Le laboratoire est dirigé par Céline Hudelot, Professeur au département d'Informatique de l'école, responsable de la Dominante Informatique et Numérique et co-responsable de la mention IA et du MSc AI for Society. Elle anime l'axe transverse en IA du laboratoire.

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)).

**Contacts presse :**

Claire Flin : clairefline@gmail.com – 06 95 41 95 90 // Marion Molina : marionmolinapro@gmail.com - 06 29 11 52 08