Press Release
10 March 2025

ARTIFICIAL INTELLIGENCE AND NLP: EUROBERT'S SUITE OF ENCODERS MODELS REACHING A NEW MILESTONE

Trained on 5 trillion tokens, this suite offers sovereign, open-source models that deliver the best text representation performance for European languages, as well as for mathematics and code-related tasks.

The collaboration between the MICS laboratory at CentraleSupélec, Diabolocom, Artefact and Unbabel, with technological support from AMD an CINES, has led to the release of the best multilingual text representation model, a fundamental building block for information retrieval, classification and quality estimation (of abstracts, translations).

These types of models are indispensable in automatic language processing (NLP) and have been among the most popular downloads on Hugging Face for many years. Their ability to accurately capture the meaning and context of sentences, offering deep and detailed linguistic understanding, is essential to the development of advanced artificial intelligence applications. This new EuroBERT model is available under Apache 2.0 on the Hugging Face platform starting March 10, 2025.

The research project was led by Nicolas Boizard, a Cifre doctoral student at Diabolocom, with major contributions from Hippolyte Gisserot-Boukhlef, a Cifre doctoral student at Artefact, and Duarte Alves at the Instituto Superior Técnico (IST) as part of the research initiated by Pierre Colombo, associate professor at CentraleSupélec, under the supervision of Céline Hudelot (director of MICS) and André Martins, associate professor at IST. The results are detailed in an article published on arXiv on 10 March 2025.

A new technological leap in text encoding

EuroBERT stands out from other available encoders in five key aspects:

- It is sovereign and fully open source, including both its source code and datasets.
- It supports eight major European languages as well as seven of the most widely spoken non-European languages.
- Trained on 5 trillion tokens—twice as many as standard encoders or generative models like Llama 2 (2 trillion tokens)—EuroBERT provides optimal capabilities without additional usage costs.
- The EuroBERT family demonstrates strong performance across various tasks, particularly in information retrieval, classification and quality assessment (of abstracts, translations).
- It excels in previously overlooked tasks, such as processing mathematical data and programming languages.
- It offers three model sizes (210M, 610M, and 2.1B), providing an optimal balance between speed, quality, and cost to meet enterprise requirements.

EuroBERT reshapes the landscape for applications leveraging Natural Language Processing (NLP) and sentence representations, including text analysis, information retrieval, text classification, and information extraction.

The strength and added value of collaborative research

As with the CroissantLLM and EuroLLM models published on Hugging Face in 2024, these scientific advancements were made possible through a close and rich public-private collaboration, rooted in the Paris-Saclay ecosystem and expanded to a European scale. The teams from MICS, IST, Diabolocom, Artefact, and Unbabel worked together on the three ongoing PhD projects, alongside the French supercomputer Adastra powered by AMD Instinct™ Accelerators and AMD EPYC™ processors.

Globally recognized for its scientific excellence in mathematics and computer science, the MICS laboratory at CentraleSupélec leads and conducts numerous research projects, programs and collaborations with both private and public partners, constantly pushing the boundaries of artificial intelligence. Diabolocom, with its customer relationship management product, contributed its expertise in language processing, integrated into their product. Artefact, a European leader in AI and data consulting, brought its multi-sector expertise and a cross-functional vision for the many applications deployed in businesses. Finally, Unbabel, a tech leader in machine translation, contributed its multilingual expertise, particularly through their datasets.

"*A month after the AI Action Summit held in Paris, we are particularly excited to announce the availability of EuroBERT. This suite of 'encoder' models for European languages is currently the most comprehensive and high-performing for handling document-level tasks. In today's AI landscape, 'encoder' models are often overlooked despite their importance in NLP applications. For example, BERT—introduced in 2017—now enjoys nearly 5 million downloads per month on Hugging Face, surpassing LLaMA and other similar models*", says Céline Hudelot, Professor at CentraleSupélec and Director of the MICS laboratory.

With the creation of Diabolocom Research in early 2025 and the collaboration of its research center on academic projects, Diabolocom strengthens its capabilities to provide concrete and efficient solutions to the market's needs for reliable, sovereign, and high-performance AI systems.
"*Multidisciplinary collaboration and contributions to open-source projects are at the heart of our strategy to stay at the forefront of innovation. EuroBERT, the latest project from our research center, addresses several limitations of existing encoders. It will contribute to enhancing the functionality of several of our solutions, such as automatic information retrieval, automatic classification, and agent-based systems*", says Frédéric Durand, CEO at Diabolocom.

Artefact, on its side, has committed to AI research through its research center, inaugurated a year ago.
"*Our goal is to develop and share models that are useful and usable for practical applications in business. All our publications and algorithms are open source. The advancements in document encoding models represented by EuroBERT open new possibilities for improving the performance and relevance of document classification and organization, intelligent information retrieval and NER (Named Entity Recognition). By focusing on its role in document encoding, it addresses a recurring and essential need for text analysis in business*", adds Emmanuel Malherbe, Research Director at Artefact.

As for Unbabel, the first AI-powered language operations platform : "*EuroBERT represents a major breakthrough in multilingual AI. Encoder models have long been a hidden asset of NLP, providing the deep linguistic understanding essential to high-performance AI applications. Unlike purely generative approaches, encoders excel at capturing meaning and context, key elements for accurate and*

scalable multilingual systems. At Unbabel, we have solid expertise not only in the development of generative LLM solutions, such as our state-of-the-art Tower models, but also in the creation of reference encoder-based solutions, such as Comet and CometKiwi. The launch of EuroBERT comes at a key moment, filling the gap in multilingual encoder models trained with the key advances of generative models. This breakthrough represents a further step in building an infrastructure essential to strengthening our AI sovereignty in Europe, and we are proud to contribute to it through projects like EuroBERT and EuroLLM, which strengthen European capabilities and secure our common digital future", adds Nuno Miguel Guerreiro, Researcher at Unbabel.

This project was also made possible thanks to cutting-edge AMD Instinct™ MI300A Accelerators**,** integrated into Adastra, the highly efficient French supercomputer.
"*We are thrilled to announce the launch of EuroBERT as this project showcases our commitment to advancing language models and leveraging the full potential of cutting-edge technology from AMD. The collaboration between our teams has been instrumental in achieving these remarkable technical milestones*" said Stephanie Dismore, SVP EMEA Region, AMD.

The development of EuroBERT also involved teams from the University of Grenoble Alpes, INRIA Rennes, Illuin Technology, IRT Saint-Exupéry, and CINES.

About CentraleSupélec - www.centralesupelec.fr
CentraleSupélec is a public institution with a scientific, cultural, and professional focus, founded in January 2015 through the merger of École Centrale Paris and Supélec. Today, CentraleSupélec has 4 campuses in France (Paris-Saclay, Metz, Rennes, and Reims). It has over 5,400 students, including 3,800 engineering students, and is home to 18 research laboratories or teams. With a strong international presence (25% of its students and nearly a quarter of its faculty are international), the school has established over 170 partnerships with the world's top institutions. A leader in higher education and research, CentraleSupélec is a reference hub in engineering sciences and systems. It co-founded the University of Paris-Saclay in 2020 and chairs the Group of Écoles Centrale (CentraleSupélec, Centrale Lyon, Centrale Lille, Centrale Nantes, and Centrale Méditerranée), which operates international campuses in Beijing (China), Hyderabad (India), and Casablanca (Morocco).

About MICS Laboratory – Web Site
Created in the early 2000s, MICS brings together research in Mathematics and Computer Science at CentraleSupélec. At the heart of digital technologies, its themes focus on the modeling, simulation, analysis, and optimization of complex systems, whether they originate from industry, life sciences, markets, or information and networks. The MICS laboratory is organized into 6 research teams with shared scientific goals, along with a cross-cutting focus on Artificial Intelligence.

Press Contacts:
Claire Flin: clairefline@gmail.com – +33 6 95 41 95 90
Marion Molina: marionmolinapro@gmail.com – +33 6 29 11 52 08

About Diabolocom - www.diabolocom.com
For over 20 years, Diabolocom has transformed customer interactions with its cloud-based CCaaS solution, powered by proprietary generative AI. We equip customer service and sales teams with intelligent automation, seamless accessibility, and reliable analytics to enhance efficiency and engagement. Our AI, designed for customer relations, offers real-time transcription, satisfaction analysis, and action recommendations—while minimizing repetitive tasks. The result: hyper-personalized interactions, increased customer loyalty, and optimized sales.

With native integrations into CRMs like Salesforce, Microsoft, and ServiceNow, Diabolocom provides full visibility into every interaction, empowering global leaders such as Carrefour, Air Liquide, Meilleurtaux, and Leboncoin to transform customer relations in over 60 countries.

In 2025, we launched Diabolocom Research to tackle the challenges of building responsible, reliable, ethical, and high-performance AI systems for contact centers. Our research spans speech technology, natural language processing, conversational AI, and hardware-algorithm optimization—pushing the boundaries of innovation in customer experience.

Press Contact:
Nada Nachit : nada.nachit@diabolocom.com

About d'Artefact - www.artefact.com

Artefact is a French consulting and engineering firm specializing in data and AI, and a leader in Europe. Based in Paris, it is currently present in 23 countries across all continents and employs 1,600 people.

Its mission is to help companies harness the full potential of AI and data by developing customized solutions tailored to their business challenges. As a pioneer in this field, it combines technological expertise with operational excellence, collaborating with the largest players in the market. Its clients span key sectors of the economy – industry, retail, luxury, consumer goods, healthcare, finance – and include major international corporations.

Beyond consulting, Artefact is actively committed to ethical and accessible AI. It launched the "School of Data" to promote career transitions into tech professions and established AI research centers in Paris and Shanghai.

Press Contact:
Astrid Calippe : astrid.calippe@artefact.com