Press Release

ARTIFICIAL INTELLIGENCE AND NLP: THE EUROBERT ENCODER SUITE REACHES A NEW MILESTONE

Paris, March 10, 2025 - Trained on 5,000 billion tokens, this suite offers sovereign, open-source models delivering the best text representation performance for European languages, as well as for tasks related to mathematics and coding.

The collaboration between CentraleSupélec's MICS laboratory, Diabolocom, Artefact, and Unbabel, supported by the technological expertise of AMD and CINES, has resulted in the release of the most advanced multilingual text representation model. This model serves as a fundamental building block for information retrieval (RAG), classification, and quality estimation (summarization, translation).

These types of models are essential in natural language processing (NLP) and have been among the most downloaded on Hugging Face for many years. Their ability to accurately capture the meaning and context of sentences, offering a refined and in-depth linguistic understanding, is crucial for the development of advanced artificial intelligence applications. The new EuroBERT model is available as of March 10, 2025, under the Apache 2.0 license on the Hugging Face platform.

The research project was led by Nicolas Boizard, a Cifre PhD candidate at Diabolocom, with major contributions from Hippolyte Gisserot-Boukhlef, a Cifre PhD candidate at Artefact, and Duarte Alves, a PhD candidate at Instituto Superior Técnico (IST). It builds upon research initiated by Pierre Colombo, Associate Professor at CentraleSupélec, and was conducted under the supervision of Céline Hudelot, Director of MICS, and André Martins, Associate Professor at IST. The results are detailed in a paper published on arXiv on March 10, 2025: https://arxiv.org/abs/2503.05500

A new technological leap in text encoding

EuroBERT stands out from currently available encoders in five key ways:

- It is sovereign and fully open-source, including both its source code and datasets.
- It supports 8 major European languages as well as 7 of the most widely spoken non-European languages.
- Trained on 5 trillion tokens, twice the amount of data used for standard encoders or generative models such as Llama 2 (2 trillion tokens), EuroBERT offers optimal capabilities without additional usage costs.
- The EuroBERT family provides the best foundation for information retrieval (RAG), classification, and quality estimation (summarization, translation).
- It excels in previously underexplored areas such as mathematical data processing and programming languages.

It is available in three model sizes (210M, 610M, and 2.1B), offering an optimal balance between speed, quality, and cost, tailored to the needs of enterprise users.

EuroBERT is thus transforming natural language processing applications based on sentence representations, such as text analysis, information retrieval, 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 dynamic public-private collaboration rooted in the Paris-Saclay ecosystem and extended across Europe. The teams from MICS, IST, Diabolocom, Artefact, and Unbabel worked together within the framework of three ongoing PhD projects, supported by the French supercomputer Adastra, powered by AMD Instinct™ Accelerators and AMD EPYC™ processors.

Recognized worldwide for its excellence in mathematics and computer science, CentraleSupélec's MICS laboratory leads multiple research programs and projects in partnership with private and public organizations, continually pushing the boundaries of artificial intelligence. Diabolocom, through its customer relationship support product, has contributed its expertise in language processing, which has been integrated into its product. Artefact, a European leader in AI and data consulting, provided its cross-sector expertise and strategic vision on numerous enterprise applications. Finally, Unbabel, a tech leader in machine translation, contributed its expertise in multilingual AI.

"One month after the AI Action Summit in Paris, we are particularly excited to announce the release of EuroBERT. This family of encoder models for European languages is the most comprehensive and high-performing solution for 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, still sees nearly five million downloads per month on Hugging Face, surpassing LLaMA and other similar models," emphasizes **Céline Hudelot, Professor at CentraleSupélec and Director of the MICS laboratory.**

With the establishment of its research center, Diabolocom Research, in early 2025, Diabolocom is equipping itself with new resources to provide concrete and efficient solutions to market demands 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, our latest research initiative, addresses several limitations of existing encoders. This model will enhance the functionality of multiple solutions, including automatic information retrieval, automated classification, and agent-based systems," explains **Frédéric Durand, President and Founder of Diabolocom.**

Artefact, for its part, has been actively engaged in AI research through its research center, inaugurated a year ago. *"Our goal is to develop and distribute useful, practical models for concrete business applications. As a result, all our publications and algorithms are open-source. The advancements in document encoding represented by EuroBERT open up new possibilities for improving the efficiency and relevance of document classification, intelligent information retrieval, and named entity recognition (NER). By focusing on analyzing existing documents rather than generating new ones, EuroBERT addresses a critical and recurring need for business text analysis,"* adds **Emmanuel Malherbe, Director of the Artefact Research Center.**

As for Unbabel, the first AI-powered Language Operations platform: *"EuroBERT represents a major breakthrough in multilingual AI. Encoder models have long been an unsung hero of NLP, providing the deep linguistic understanding necessary for high-performing 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 strong expertise not only in developing generative LLM solutions, such as our cutting-edge Tower models, but also in creating reference encoder-based solutions like Comet and CometKiwi. The launch of EuroBERT comes at a pivotal moment,*

addressing the lack of multilingual encoders trained with the latest generative model advancements. This marks another step toward building the essential infrastructure for strengthening Europe's AI sovereignty, and we are proud to contribute through projects like EuroBERT and EuroLLM, which enhance European capabilities and secure our shared digital future," adds **Nuno Miguel Guerreiro, researcher at Unbabel.**

This project was also made possible thanks to AMD Instinct™ MI300A Accelerators, integrated into Adastra, the highly efficient French supercomputer.

For Julien Ruiz, Director of AMD France, "*The development of EuroBERT marks a significant milestone in our efforts to enhance natural language processing capabilities for European languages, driven by research in France. Leveraging AMD's MI300 GPUs and their unified memory architecture, we achieved unprecedented performance and efficiency. This project underscores AMD's commitment to innovation and excellence in artificial intelligence."*

The development of EuroBERT also involved contributions from teams at Université Grenoble Alpes, CNRS, LISN, Illuin Technology, IRT Saint-Exupéry, and CINES.

About CentraleSupélec - www.centralesupelec.fr

CentraleSupélec is a public institution dedicated to scientific, cultural, and professional education, founded in January 2015 through the merger of École Centrale Paris and Supélec. Today, CentraleSupélec operates across four campuses in France (Paris-Saclay, Metz, Rennes, and Reims) and has more than 5,400 students, including 3,800 engineering students. The institution hosts 18 research laboratories or teams. With a strong international presence—25% of its students and nearly a quarter of its faculty coming from abroad—CentraleSupélec has established over 170 partnerships with top institutions worldwide. A leading school in higher education and research, it serves as a benchmark in the fields of engineering and systems sciences. In 2020, CentraleSupélec co-founded Université Paris-Saclay and currently leads the Groupe des Écoles Centrale (CentraleSupélec, Centrale Lyon, Centrale Lille, Centrale Nantes, and Centrale Méditerranée), which oversees international campuses in Beijing (China), Hyderabad (India), and Casablanca (Morocco).

About the MICS Laboratory
Founded in the early 2000s, the MICS laboratory brings together research in Mathematics and Computer Science at CentraleSupélec. Positioned at the core of digital technologies, its research focuses on modeling, simulation, analysis, and optimization of complex systems, spanning industrial applications, life sciences, financial markets, and information networks. The MICS laboratory is structured into six research teams, pursuing shared scientific objectives, along with a cross-disciplinary 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 been revolutionizing customer interactions with its cloud-based CCaaS (Contact Center as a Service) solution, enhanced by proprietary generative AI. Intelligent automation, improved reachability, and reliable analytics provide customer service and sales teams with the tools they need to succeed. Designed specifically for customer relations, Diabolocom's AI features real-time transcription, satisfaction analysis, and action recommendations while minimizing repetitive tasks. The result: hyper-personalized interactions, stronger customer loyalty, and optimized sales. By offering full visibility into every customer interaction, Diabolocom's solution helps leading companies such as Carrefour, Air Liquide, Meilleurtaux, and Leboncoin transform their customer relationships in over 60 countries.

With a presence in Europe, North America, Brazil, and the Middle East, Diabolocom supports organizations in enhancing customer relations on a global scale.

In 2025, Diabolocom launched its research center, **Diabolocom Research**, dedicated to tackling challenges in designing responsible, reliable, ethical, and high-performance contact center systems. The lab focuses on developing cutting-edge technologies in speech processing, natural language processing, conversational AI, and hardware-algorithm optimization.

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

About Artefact:

Artefact is a French consulting and engineering firm specializing in data and AI, and a European leader in the field. Headquartered in Paris, we are now present in 23 countries across all continents, with a team of 1,500 employees.

Our mission is to help businesses unlock the full potential of AI and data by developing tailored solutions that address their specific industry challenges. As pioneers in this field, we combine technological expertise with operational excellence, collaborating with major market players. From strategy to operations, we offer an end-to-end approach and solutions: data strategy, data quality and governance, data platforms, AI Factory, data-driven customer experience and marketing ROI

Our clients span all key economic sectors - industry, retail, luxury, consumer goods, healthcare, finance, and more - including large international corporations.

Beyond consulting, we are actively committed to promoting ethical and accessible AI. We launched the **"School of Data"** to facilitate career transitions into tech roles and established **Artefact AI Research Centers** in Paris and Shanghai.

Find out more at **www.artefact.com**

Press contact:
Astrid Calippe : **astrid.calippe@artefact.com**