Press Release
Paris, May 7, 2025

LightOn sets new standards for complex information retrieval (RAG) with GTE-ModernColBERT.

LightOn is proud to announce the release of **GTE-ModernColBERT**, our new state-of-the-art, open-source, multi-vector retrieval model. By leveraging ModernBERT architecture and our innovative PyLate library, we've created a solution that sets a new milestone in the field and addresses the complex challenges of modern enterprise information retrieval. This new model outperforms models of the ecosystem (Alibaba, Snowflake, Cohere, BAAI, JinaAI..) in the industry-standard LongEmbed benchmark.

Breaking New Ground in Retrieval Technology

Traditional single-vector embedding models have become standard in the industry, but as enterprise needs evolve toward handling longer contexts and specialized domains, their limitations become increasingly apparent. GTE-ModernColBERT-base represents a significant leap forward with its state-of-the-art multi-vector (late interaction) architecture, offering:

Outstanding generalization capability for long documents

GTE-ModernColBERT sets a new benchmark (SOTA — State of the Art) for generalization with long contexts. It outperforms the best existing models by a 10 point margin (LongEmbed benchmark) on documents up to 32,000 tokens, equivalent to texts spanning dozens of pages, even though it was initially trained only on 300 token excerpts from the MS MARCO dataset. These early results indicate that GTE-ModernColBERT could further extend its capabilities, delivering excellent performance even beyond this already impressive context window.

- **Extended context handling** for documents up to 32,000 tokens
- **Superior generalization** for domain-specific, confidential, or specialized content
- **Breakthrough performance** as the first model to surpass ColBERT-small on the BEIR benchmark
- **Remarkable efficiency** through ModernBERT's architectural advancements

LightOn's Technical Innovation

LightOn created GTE-ModernColBERT as an unique solution by identifying and building upon key elements:

1. **Modern encoder:** LightOn built ModernBERT to enable the creation of powerful and up to date retrieval models. GTE-ModernColBERT is a direct follow-up of this first release to extend on the very promising multi-vector approach.

2. **PyLate Library**: We developed a framework to enable streamlined implementation to experiment and train multi-vector retrieval models. Only 80 lines of code are needed to reproduce the training process.

3. **Knowledge Distillation**: By training on MS MARCO via knowledge distillation, we've created a lightweight yet powerful model that doesn't compromise on performance.

4. **Compatibility Focus**: Most major vector databases including QDrant, LanceDB,Weaviate and Vespa now support multi-vectors indexation, making enterprise adoption frictionless.

Transforming Enterprise RAG Implementations

GTE-ModernColBERT fundamentally transforms how organizations can implement Retrieval-Augmented Generation (RAG) by:

- Enhancing search quality within proprietary knowledge bases
- Maintaining high performance even with highly specialized content
- Supporting enterprise-scale document processing
- Enabling more accurate retrieval for AI-generated responses

Real-World Impact

For knowledge management teams and AI solution developers, GTE-ModernColBERT offers the ideal foundation for next-generation information systems. Its ability to process large volumes of text while maintaining contextual understanding makes it particularly valuable for:

- Legal document analysis
- Scientific research repositories
- Technical documentation search
- Customer support knowledge bases
- Internal enterprise knowledge management

Open Source Commitment

After the release of ModernBERT and ModernBERT-embed, by releasing GTE-ModernColBERT as an Apache 2.0 licensed open-source solution, LightOn continues its commitment to advancing the field of AI while enabling organizations of all sizes to benefit from cutting-edge retrieval technology and empower research through open sourcing PyLate as well.

For organizations seeking to stay ahead in Knowledge Management and RAG, GTE-ModernColBERT is now available. Try it out and (re)discover the hidden value within your documents!

🎯 Try it today on Hugging Face
📚 **Get started: PyLate Documentation**

About LightOn

Founded in 2016 in Paris and the first European generative AI company listed on Euronext Growth, LightOn is a pioneering player in the field of sovereign GenAI. Its Paradigm platform enables organizations to deploy large-scale AI while ensuring the confidentiality of their data. LightOn's technology ensures essential strategic independence by offering tailored solutions. This technological mastery is accompanied by the ability to process large volumes of data for industrial uses, with applications in various sectors such as finance, industry, health, defense, and public services.

LightOn is listed on Euronext Growth® Paris (ISIN: FR0013230950, ticker: ALTAI-FR). The company qualifies for PEA and PEA PME investment plans and is recognized as an "Innovative Company" by Bpifrance. To learn more : https://www.lighton.ai

Contacts

LIGHTON invest@lighton.ai	**SEITOSEI●ACTIFIN** **Investor Relations** Benjamin LEHARI lighton@seitosei-actifin.com
KALAMARI **Media Relations** Camille Bernisson - +33 7 64 44 14 49 Maroua Derdega - +33 7 63 77 73 20 lighton@kalamari.agency	**SEITOSEI●ACTIFIN** **Financial Media Relations** Jennifer JULLIA - +33 6 47 97 54 87 jennifer.jullia@seitosei-actifin.com