LightOn Demonstrates the Flexibility of Its OCR Model by Adapting It to Arabic Through Targeted Training

Paris, June 12, 2026 – LightOn demonstrates the flexibility of LightOnOCR-2, its document understanding model, by adapting it to Arabic through fine-tuning. This extension is based on an internal synthetic data generation pipeline designed to cover languages that remain underrepresented in the OCR tools currently available on the market.

This demonstration is based on a dataset comprising 12,000 synthetic pages and their reference transcriptions, produced using a modified version of LightOn's synthetic document generator.

The corpus covers a wide range of document scenarios, including scanning artifacts, font variations, resolution levels, and document types. The output format remains the one used to train the *bbox* variant of LightOnOCR-2, with bounding box detection that associates text with its spatial location.

Applying OCR to Arabic presents specific challenges. The script is written from right to left, characters are connected in cursive form, and open datasets, like specialized models, remain less widely available than for Latin-based languages. For organizations processing archives, administrative, legal, or heritage documents in Arabic, these limitations can slow the automation of document workflows.

This demonstration is part of a broader effort to extend the model to a variety of domains, as reflected by its more than 3 million downloads and the fine-tunings already carried out by the community. It notably addresses needs encountered in the Middle East, where LightOn is already working with public- and private-sector organizations. This development is consistent with LightOn's positioning: providing enterprise-grade generative AI building blocks that are open, controllable, and tailored to sensitive environments.

LightOn is making the guides needed to reproduce this fine-tuning available on its Hugging Face space, with the aim of making this approach accessible to as many users as possible and adaptable to other document contexts.

LightOnOCR-2 is released as open source under the Apache 2.0 license. It plays a central role in the production document ingestion process within LightOn Console, LightOn's self-service offering.
The open model and our production engine are therefore built on the same technological foundation. The base model achieves a score of 83.2% on OlmOCR-Bench.

About LightOn

Founded in Paris in 2016, and the first European AI company listed on Euronext Growth, LightOn develops an enterprise AI platform designed to enable organizations to connect cutting-edge AI to their sensitive data. LightOn offers an integrated architecture built for large-scale production deployment, robust, efficient, and secure, allowing organizations to industrialize use cases in regulated environments. LightOn's solutions are intended in particular for the finance, industrial, healthcare, defense, and public sectors.

LightOn is listed on Euronext Growth® Paris (ISIN: FR0013230950, ticker: ALTAI-FR). The company is eligible for PEA and PEA-PME investment schemes and has been recognized as an "Innovative Company" by Bpifrance.

To learn more: www.lighton.ai

Contacts LightOn

LIGHTON invest@lighton.ai	**SEITOSEI•ACTIFIN** **Relations investisseurs** Benjamin LEHARI lighton@seitosei-actifin.com
KALAMARI **Relations médias** Maroua DERDEGA - +33 7 63 77 73 20 lighton@kalamari.agency	**SEITOSEI•ACTIFIN** **Relations presse financière** Jennifer JULLIA - +33 6 47 97 54 87 jennifer.jullia@seitosei-actifin.com