LightOn démontre la flexibilité de son modèle d'OCR en l'adaptant à la langue arabe via entraînement ciblé

Paris, le 12 juin 2026 – LightOn démontre la flexibilité de LightOnOCR-2, son modèle de compréhension documentaire, en l'adaptant à l'arabe par fine-tuning. Cette extension repose sur un pipeline interne de génération de données synthétiques, conçu pour couvrir des langues encore peu représentées dans les outils OCR du marché.

Cette démonstration repose sur un jeu de données composé de 12 000 pages synthétiques et de leurs transcriptions de référence, produit à l'aide d'une version modifiée du générateur de documents synthétiques de LightOn.

Le corpus couvre une diversité de situations documentaires : artefacts de numérisation, variations de polices, niveaux de résolution et types de documents. Le format de sortie reste celui utilisé pour l'entraînement de la variante bbox de LightOnOCR-2, avec détection des boîtes englobantes, qui permettent d'associer au texte sa localisation spatiale.

L'OCR appliqué à l'arabe présente des défis spécifiques. L'écriture s'effectue de droite à gauche, les caractères se lient en cursive, et les jeux de données ouverts comme les modèles spécialisés demeurent plus rares que pour les langues latines. Pour les organisations qui traitent des archives, des documents administratifs, juridiques ou patrimoniaux en arabe, ces limites peuvent ralentir l'automatisation des chaînes documentaires.

Cette démonstration s'inscrit dans un mouvement plus large d'extensions du modèle à des domaines variés, comme en témoignent ses plus de 3 millions de téléchargements et les fine-tunings déjà réalisés par la communauté. Elle répond notamment aux besoins rencontrés au Moyen-Orient, où LightOn est déjà présent auprès d'acteurs publics et privés. Cette évolution s'inscrit dans la continuité du positionnement de LightOn : proposer des briques d'IA générative d'entreprise, ouvertes, maîtrisables et adaptées aux environnements sensibles.

LightOn met à disposition les guides nécessaires à la reproduction de ce fine-tuning sur son espace Hugging Face, afin de rendre cette approche accessible au plus grand nombre et adaptable à d'autres contextes documentaires.

LightOnOCR-2 est diffusé en open source sous licence Apache 2.0. Il est central au processus d'ingestion de documents en production dans LightOn Console, l'offre self-service de LightOn. Le modèle ouvert et le moteur de production reposent ainsi sur une même base technologique. Le modèle de base atteint un score de 83,2% sur OlmOCR-Bench.

À propos de LightOn

Fondée en 2016 à Paris et première société européenne de l'IA cotée sur Euronext Growth, LightOn développe une plateforme d'IA d'entreprise conçue pour permettre aux organisations de connecter une IA de pointe sur leurs données sensibles. LightOn propose une architecture intégrée, pensée pour le passage en production à grande échelle, robuste, frugale et sécurisée, permettant d'industrialiser des cas d'usage dans des environnements régulés. Les solutions LightOn s'adressent notamment aux secteurs de la finance, de l'industrie, de la santé, de la défense et du secteur public.

LightOn est cotée sur Euronext Growth® Paris (ISIN : FR0013230950, mnémonique : ALTAI-FR). La société est éligible au PEA et au PEA PME, et est qualifiée « Entreprise innovante » par Bpifrance. Pour en savoir plus : www.lighton.ai/fr

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