PROPOSITION SUJETS DE THESE
CONTRATS DOCTORAUX
2025—2028

X Contrat doctoral fléché FR Agorantic
X Contrat doctoral fléché EUR InterMEDIUS
☐Dans le cas où vous souhaitez candidater sur les deux contrats doctoraux, veuillez cochez les deux cases.

Directeur·rice de thèse :	SANJUAN Eric	**Mail :**	eric.sanjuan@univ-avignon.fr
Laboratoire :	LIA/JPEG	**Téléphone :**	0686230119
Co-directeur·rice et/ou co-encadrant·e :	Chistina Koumpli Gaël Depoorter		
Laboratoire :	JPEG		

Titre en français : Les grands modèles de langue libres : approches interdisciplinaires sur l'intelligence artificielle artificielle générative embarquée, la protection des données et les responsabilités socio-légales.

Titre en anglais : Open Large Language Models: Interdisciplinary Approaches to Embedded Generative Artificial Intelligence, Data Protection, and Socio-Legal Responsibilities.

Résumé (7 lignes maximum) : Le projet de thèse intitulé explore les défis et opportunités liés aux modèles internalisables, utilisables de manière décentralisée, dans la lignée d'outils comme Ollama. Ces modèles, souvent dérivés de LLMs centralisés émis par de grandes multinationales, posent question sur les responsabilités des émetteurs et des usagers. La thèse examine notamment les implications potentielles du RGPD dans l'éventualité où ces modèles seraient assimilés à des bases de données fiables et non à des systèmes stochastiques. Pour cela il est essentiel d'analyser l'impact du partage des paramètres sur la protection des données personnelles, les secrets industriels, les usages possibles et les responsabilités juridiques.

Mots clés : IA générative, licenses logiciels libres, Open Data, EU AI act, RGPD, systèmes stochastiques, systèmes experts, bases de données relationnelles

1- Présentation du sujet (3 pages maximum)

Ce sujet fait suite à l'interven Nicolas Berkouk, expert scientifique au sein du service IA de la CNIL, sur la régulation de l'Intelligence Artificielle lors de la Journée d'étude sur L'intelligence artificielle générative et l'enjeu des données personnelles organisée le CREIS[1] à la MSH de Paris Nord 29 novembre 2024.

Il s'agit d'explorer les défis et opportunités liés aux larges modèles de langues (LLMs) internalisables, utilisables de manière décentralisée, popularisés par des outils tels Ollama[2]. Ces modèles, souvent dérivés de LLMs centralisés développés par de très grandes multinationales, posent question sur les responsabilités partagées entre émetteurs initiaux et utilisateurs, notamment en cas d'usage problématique vis-à-vis du RGPD et du EU AI Act. On envisage en particulier le cas où ces modèles seraient assimilés à des bases de données fiables ou à des systèmes experts, plutôt qu'à des systèmes stochastiques. Le principal objectif est ainsi de clarifier pour le législateur, la différence de nature entre LLMs et bases de données, bien que les LLMs soient utilisés avec succès pour enrichir des bases de données de connaissances[3].

L'utilisation décentralisée et internalisée des LLMs répond à un nombre croissant d'usages en Europe, car elle permet des traitements localisés conformément aux exigences du RGPD. Ces recherches requièrent ainsi une analyse sociologique de ces usages pour en établir une cartographie complète des pratiques actuelles et émergentes. Il s'agit d'établir une cartographie détaillée des besoins et des risques associés à l'utilisation de ces modèles, dans des contextes variés tels que l'industrie, la recherche, ou les services publics. Ces premiers travaux contribueront directement :

- au processus législatif européen, en apportant une meilleure compréhension des implications sociétales, éthiques et pratiques des LLMs internalisés par des entreprises ou organismes européens.
- à la proposition d'un nouveau type de licences libres dites "Open Weights"[4]. Ces licences visent à permettre le partage des paramètres pré entraînés des modèles sans imposer la divulgation des données d'entraînement, tout en préservant la confidentialité et la protection des données sensibles. Elles limitent aussi la responsabilité juridique des émetteurs initiaux face aux usages selon les principes de transparence et de partage propres aux logiciels libres.

Cependant la prise en compte dans le travail législatif de ce nouveau type de licences nécessite une compréhension approfondie de la nature même des grands modèles de langage (LLMs) à l'intersection de plusieurs paradigmes technologiques. Le RGPD repose sur une distinction claire entre bases de données et algorithmes de traitement. Or les LLMs partagent des similitudes avec les logiciels, en raison de leur capacité à exécuter des tâches sur la base de programmes entraînés, mais également avec les bases de données, en ce qu'ils encapsulent de vastes quantités d'information dérivées des ensembles de données utilisés lors de leur entraînement.

Cette étude devra s'appuyer sur des protocoles expérimentaux clairs et reproductibles, visant à élucider plusieurs questions fondamentales. Parmi celles-ci, une priorité sera d'examiner la possibilité d'établir une preuve d'entraînement sur des données illicites ou protégées. Cette problématique implique de travailler sur des mécanismes permettant d'auditer les modèles à posteriori, tout en respectant les contraintes de confidentialité liées aux données d'entraînement. Une telle analyse devra également évaluer la faisabilité technique et les limites des approches actuelles, comme les traces de gradients, les tests d'extraction de données (model inversion), ou les empreintes statistiques laissées par les données d'entraînement.

Enfin, le développement et la validation de ces licences devront intégrer des expériences de terrain, afin de tester les scénarios d'utilisation des modèles dans des environnements réels. Ces expériences

---

[1] https://www.lecreis.org/?p=3589

[2] https://ollama.com/

[3] Li, J., Hui, B., Qu, G., Yang, J., Li, B., Li, B., ... & Li, Y. (2024). Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, *36*.

[4] Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J. S., ... & Kembhavi, A. (2024). Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

permettront de vérifier si les protocoles et les mécanismes proposés, notamment en termes de partage des paramètres et de limitation de la responsabilité des émetteurs, sont adaptés aux besoins des utilisateurs tout en répondant aux exigences des régulateurs. Ce travail pourra bénéficier du cadre du GDR MADICS et son action Simple Text[5].

Pour ne pas écarter l'éventualité où les licences de type Open Weights s'avéreraient incompatibles avec la législation européenne, le travail de thèse inclura également une exploration des LLMs entièrement entraînés sur des données ouvertes. Ces modèles, basés sur des jeux de données accessibles à tous, offrent une alternative intéressante aux modèles nécessitant le partage restreint ou confidentiel des paramètres. En effet, leur transparence et leur conformité native aux principes des données ouvertes (Open Data) en font des candidats particulièrement adaptés aux exigences européennes, notamment celles du RGPD et du EU AI Act. Dans ce cadre, la thèse s'appuiera sur des exemples de modèles construits exclusivement à partir de données publiques, tels que ceux présentés par Hugging Face avec leur initiative autour des Common Models[6]. Ces modèles reposent sur des corpus largement accessibles et documentés, comme The Pile ou Common Crawl pour garantir une traçabilité et une transparence totales dans leur processus de formation. En suivant cette approche, le travail de thèse analysera l'alternative qu'ils représentent en termes d'éthique, de conformité légale et d'indépendance technologique pour l'Europe.

2- Profil du/de la candidat(e)

Le sujet de la thèse s'aligne parfaitement avec les qualifications, compétences et expériences de Hichem Semmar. Le candidat possède une double compétence technique et éthique, acquise grâce à son European Master in Law, Data and AI et son Master en Systèmes Informatiques Intelligents. Il possède de solides compétences en apprentissage automatique et IA générative. Il a en particulier développé des projets dans des domaines variés, notamment la segmentation pulmonaire via des réseaux convolutifs, la classification des données environnementales, et la reconnaissance optique de caractères, montrant sa capacité à créer et évaluer des architectures complexes de deep learning. Ces compétences sont directement applicables à l'étude technique des LLMs et au développement de protocoles expérimentaux clairs.

Il maîtrise aussi les enjeux liés au RGPD et à la gouvernance des données de par son expérience chez PrivacyEngine. Cette expertise est cruciale pour explorer la compatibilité des licences Open Weights avec la législation européenne. Son implication dans des études sur les biais algorithmiques (article sur le découpage d'images par IA) et l'analyse de directives comme l'ePrivacy montre une capacité à aborder des problématiques complexes au croisement de la technologie et de la législation. Cette compétence sera précieuse pour contribuer au travail législatif via la cartographie sociologique des usages des LLMs. En tant que consultant associé, il a ausi dirigé une équipe pour simplifier des cadres juridiques et développer des outils basés sur des normes comme ISO 42001. Cette capacité à travailler en équipe sur des projets interconnectés entre technologie et législation est essentielle dans le cadre de cette thèse.

Enfin son engagement envers l'open source et la collaboration scientifique : Sa contribution à des projets open source comme l'amélioration de la bibliothèque PyDS illustre une compréhension des enjeux liés aux licences libres, qui sera un aspect clé de cette thèse.

3- Opportunités de mobilité à l'international du/de la doctorant(e) dans le cadre de sa thèse

Actuellement inscrit dans un European Master in Law, Data and AI (Erasmus Mundus), un programme qui rassemble des universités de plusieurs pays européens. Cette formation inclut déjà une forte dimension de mobilité internationale, ce qui montre qu'il est familier avec les attentes et les logistiques liées aux collaborations transfrontalières. Le candidat parle couramment le français (C2), l'anglais (C2, certifié IELTS), et l'arabe (langue maternelle). Ces compétences linguistiques facilitent les échanges et collaborations avec des équipes de recherche dans divers pays.

---

[5] https://www.madics.fr/actions/simpletext/

[6] https://huggingface.co/blog/Pclanglais/common-models

Le programme Erasmus Mundus auquel il participe lui donne accès à un réseau académique et professionnel international étendu, facilitant les stages, collaborations et séjours de recherche à l'étranger.

Sa participation à des projets open source et à des publications sur des plateformes globales (comme ResearchGate et GitHub) démontre déjà son intégration dans des communautés scientifiques internationales.

4- Références bibliographiques

Publications des encadrants pertinentes vis-à-vis de ce sujet.

Eric SanJuan (Informatique)

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. Document numérique, 17(1), 61-84.

Ermakova, L., SanJuan, E., Huet, S., Augereau, O., Azarbonyad, H., & Kamps, J. (2023, March). CLEF 2023 SimpleText Track: What Happens if General Users Search Scientific Texts?. In European Conference on Information Retrieval (pp. 536-545). Cham: Springer Nature Switzerland.

Vermeirsche, J., SanJuan, E., Jiménez, T., & Lagier, C. (2024, June). Analyse thématique comparative des discours politiques et de leur diffusion dans le Wikipédia francophone. In JADT (pp. 913-922).

Chistina Compli (Droi)

C. Koumpli, Les données personnelles sensibles : contribution à l'évolution du droit fondamental à la protection des données à caractère personnel. Étude comparée : Union Européenne, Allemagne, France, Grèce, Royaume-Uni, Préface : Pr. Otto Pfersmann, Avant-Propos : Pr. Judith Rochfeld, Collection thèses, Prix René Cassin 2020, Prix de l'IRJS 2020, Editions Pedone Paris, 2024

V. Barbé, S. Mauclaire C. Koumpli, Intelligence artificielle & Droits fondamentaux, Editions L'Epitoge, 2022.

C. Koumpli, « Why is European protection of personal data not a legal limit to the development of AI? A critical analysis of the relationship between the GDPR and the proposal of a European regulation on AI », Revista de Direitos Humanos e Desenvolvimento Social, Volume 5, ISSN 2675-9160 (en cours de diffusion - épreuves terminées)

C. Koumpli, « Les transferts de données UE-Etats-Unis. A la recherche du niveau élevé de protection européenne des données personnelles », Revue de droit public, Décembre 2024 ((en cours de diffusion - épreuves terminées)

C. Koumpli, « L'intelligence artificielle, au service de la surveillance ? Réflexions à propos de la position de la CNIL sur les conditions de déploiement des caméras dites "intelligentes" », in Lanna M. (dir), Villes numériques et sécurité, Mare & Martin, 2024 (en cours de diffusion - épreuves terminées)

Gael Depoorter (Sociologie)

« La communauté du Libre comme dynamique de reclassement du technicien informatique », Congrès de l'AFS, RT 18 « Relations professionnelles », session 5 « Digitalisation du travail et des relations de travail » organisé par H. Champin et C. Vincent, Aix-en-Provence, 29 août 2019.

« Bitcoin. Un modèle subversif de souveraineté fragilisé par le défi climatique ? », Colloque international "cryptomonnaies" organisé par Guillaume Champy, Avignon, mars 2023.

« What european regulation is doing to cryptocurrencies », Summer School EMILDAI, Pisa, juin 2024.

« Le "bal des hackers". Sociologie d'un exotisme postmoderne », Séminaire L'entreprenariat au prisme dessciences sociales, organisé par Manon Piazza, EHESS, Paris 1, 13 mai 2024.

☐ **J'ai informé le Directeur de mon unité du dépôt de cette proposition de sujet de thèse**