Université de technologie de Compiègne – Thesis proposal

Doctoral project		
Thesis title	RAG and graph-RAG techniques assessment for health technologies regulatory compliance	
Thesis speciality	Science and technology of information and systems	
Thesis supervision	**Thesis** director**: Julie FOLLET, PhD – UTC, laboratoire BMBI, Axe BioMovE** Thesis co-director: Pr Anthony FLEURY, PhD and HDR, Institut Mines Télécom Nord Europe (IMTNE), Centre d’Enseignement, de Recherche et d’Innovation Systèmes Numériques (CERI SN), animateur de l’Axe Humain, interaction, Data Ecosystem (HIDE) Contact: julie.follet@utc.fr (+33. 3.44.23.49.86) and anthony.fleury@imt-nord-europe.fr (+33.3.27.71.23.81)	
Research laboratory	**BMBI - Biomechanics and Bioengineering** Research group: BioMovE axis website: https://bmbi.utc.fr/recherche/axes-de-recherche/biomove-2/	International Cotutelle : non
Starting time	01/10/2026	
Location	UMR CNRS 7338 BMBI, University of Technology of Compiègne, Daniel-Thomas Innovation Center, BP 60319, 57 avenue de Landshut, 60203 Compiègne, France	
Funding	PhD fellowship from the French Ministry of Higher Education, Research and Space	
Supervision conditions	The PhD student is expected to have a strong appetite for the subject and a high degree of autonomy and organization. In addition to the follow-up procedures provided for the Individual Follow-up Committee for PhD students in France, progress points for the student's work will be planned at least every two weeks with the thesis director and co-director. The PhD student will have to define the agenda, prepare in advance the deliverables to be presented, explain its difficulties and identify possible solutions that will be discussed with the director and co-director. The PhD student will present his/her thesis topic and progress during team meetings (BioMove Axis for UTC, HIDE Axis for IMTNE), BMBI laboratory days, regional (CPER TecSanté) and national (Econom'IA, Deep Learning for Science, JetSan) scientific days, as well as international conferences and congresses (e.g. International Conference on Advanced Machine Learning and Data Science). The PhD student will have to produce posters and articles for publication in peer-reviewed international scientific journals. The PhD student will benefit from the training sessions offered by the UTC doctoral school, others as part of the ‘HR Excellence in Research’ label, and training sessions on a series of work tools directly provided by the BMBI laboratory members. As the thesis topic is part of a research collaboration with the Instituto de Tecnología para la Innovación en Salud y Bienestar of the Chilean University Andrés Bello (see below), the PhD student is expected to present and conduct part of his work in Chile (from a few weeks to a few months depending on the funding opportunities obtained).	
Keywords	*Large Language Models (LLMs),* *Graph-Retrieval augmented generation (RAG) techniques,* *Text data,* *Health technologies,* *Regulatory compliance*	
Summary of thesis project	The thesis project aims to study the relevance of LLMs (Large Language Models) improved by graph-RAG (Retrieval augmented generation) techniques to assist academic health technology designers and their hospital operators, under equipped with expert legal profiles, in their regulatory compliance procedures, with among other things the search for contradictory requirements between texts. This project is the follow-up of a work initiated as part of a Master 2 internship, on a corpus of nearly 700 legal texts at various stages of development (from the bill to the decree), that includes the construction of tokenized databases and context chunks databases for the	

	assessment of instruction models improved by RAG techniques. Priority is given to open source models for which training databases are known and whose performance/cost ratio is the most favorable. The thesis project will focus on studying graph-RAG techniques to improve the results obtained, and will involve the identification of the most relevant content to be converted into graphs, the research and the implementation of optimized prompt engineering strategies exploiting the knowledge bases thus constructed (chunk bases and graph bases). A substantial part of the project will be devoted to the research and testing of the evaluation criteria of the most relevant models in view of the tasks delegated to them. Depending on the use cases selected during the implementation of the project, the exploration of fine tuning strategies may be considered. The project will be conducted in collaboration with the Instituto de Tecnología para la Innovación en Salud y Bienestar (ITISB) of the Chilean University Andrés Bello (UNAB), and will give to the PhD candidate the opportunity to conduct a part of his/her research in Chile.
Thematics	Generative artificial intelligence applied to health technologies regulatory compliance
Domain	Science for the Engineer
Objectives	The thesis project aims to : -identify the steps of the health technologies regulatory compliance process for which the use of LLMs will improve performance while integrating an appropriate risk management plan ; -design and implement improvement protocols with graph-RAG techniques of LLMs trained on legal text databases in Latin languages and open source, for, among other things, the identification of inter-text contractions.
Context	Depending on the legal status, purpose of use, means of action, connectivity and disruptive nature of health technologies, their design, evaluation, placing on the market, and operation in the European Union, involve simultaneously meeting the requirements of many regulations and directives. To this body of mandatory texts must be added methodological guides for voluntary application, which are supposed to ease regulatory compliance at the operational level (up to 70 reference documents per regulation), as well as legal particularities of each Member State. For designers (academic or SMEs) and hospital operators of these technologies, low qualified about law concerns and yet required to know, understand and respect on a daily basis all these requirements is a real challenge, especially in a context of permanent overload, prohibiting any regulatory watch, yet essential in an unstable legal environment. Offering these professionals a tool to help them comply with regulations based on an LLM seems appropriate. Software solutions are also available. Made by consulting firms, standardization bodies, or by specialised companies, how such solutions work remain opaque, as the detailed characteristics of the models, their training databases, assessment methodology, and environmental impact. The current geopolitical context calls for prioritizing research on the development of sovereign solutions, truly open source (Liesenfeld & Dingemanse, 2024), eco-responsible, that participe in the resilience of States (Glasze et al., 2023). Recently, some authors have proposed using LLMs to extract specific data from legal databases (Hassani, 2024), automate the regulatory intelligence process (Ioannidis et al., 2023), or identify guidelines for the marketing of pharmaceutical products and derive relevant information from them in relation to user requests (Kim & Min, 2024). With regard to the regulatory compliance of health technologies, it appears that one of the first interesting issues to be addressed is the automatic search for intra- and inter-text contradictions in order to prioritize the working time of lawyers on the analysis of contradictions rather than on their identification, particularly time-consuming (up to several months depending on the texts). To this end, the thesis project will continue the work currently carried out as part of a Master 2 internship, which aims to search for contradictions between legal texts produced by the Institutions of 9 South American countries, on personal data and artificial intelligence use.
Methods	The project consists in the implementation of optimizing strategies of existing LLMs structuring and function by a prompt engineering technique called « retrieval augmented generation » (RAG) (Tiezzi et al., 2025).

	The technical limitations of RAG-enhanced LLMs will be addressed by developing coupling strategies with « explainable » algorithms to increase the reliability of model output data, and user confidence of such a tool, while reducing the carbon and water footprints of its operation. Encouraging resuslts have been obtained with RAG-enhanced LLMs calling legal texts structured in graphs (GraphRAG techniques). This method makes it possible to navigate efficiently within the same text (references to the various articles, annexes, etc.), by following the nodes and edges of the graph created from the text (Galli et al., 2026; Garza et al., 2024), but also from one text to another (reference from one regulation to another reference). Following the same approach, it is possible to: - extract specific requirements for each ‘role’ defined in a text (e.g. according to the European Medical Devices Regulation, a manufacturer is not subject to the same requirements as an authorized representative or importer), - to carry out a comparative analysis of different texts (from a very conceptual law to an operational guide, or of different territorial application in the United States of America vs. the European Union) (Barry et al., 2025), - even when the terminologies used to designate the same object or the same concept vary from one text to another. Given the quantity of texts to be analysed according to the use cases identified, generating a graph per text, and graphs of graphs representing the relations between the texts, seems inefficient. The challenge of the project is therefore to define the optimal strategy for coupling RAG-enhanced LLMs to knowledge bases structured in graphs. Depending on the performance obtained with the RAG and graph-RAG techniques, the thesis project does not exclude LLMs fine tuning techniques.
Expected results	- Benchmark of relevant ‘RAG-compatible’ pre-trained instruction models to the objectives of the doctoral project ; - Benchmark of relevant for LLM-as-a-Judge/Juries the automatic assessment of contents generated by the identified models ; - Benchmark of relevant LLMs for automatic knowledge graph generation ; - Identification of relevant assessment criteria of the tasks devolved to the various LLMs thus identified; - Assessment datasets building ; - Knowledge graphs generation ; - Contribution to the development of a 6-TRL prototype software tool whose purpose is to enable researchers to integrate regulatory requirements from the early stages of the innovative devices development process, in order to ease its transfer to industry.
Bibliographical references	Barry, M., Caillaut, G., Halftermeyer, P., Qader, R., Mouayad, M., Cariolaro, D., Deit, F. L., & Gesnouin, J. (2025, janvier). GraphRAG : Leveraging Graph-Based Efficiency to Minimize Hallucinations in LLM-Driven RAG for Finance Data. *31st International conference on Computational Linguistics Workshop Knowledge Graph & GenAI*. https://hal.science/hal-04907346 Galli, F., Dal Pont, T. R., Sartor, G., & Contissa, G. (2026). Approaching the AI Act... with AI : LLMs and knowledge graphs to extract and analyse obligations. *Computer Law & Security Review*, *60*, 106230. https://doi.org/10.1016/j.clsr.2025.106230 Garza, L., Elluri, L., Kotal, A., Piplai, A., Gupta, D., & Joshi, A. (2024). *PrivComp-KG : Leveraging Knowledge Graph and Large Language Models for Privacy Policy Compliance Verification* (arXiv:2404.19744). arXiv. https://doi.org/10.48550/arXiv.2404.19744 Glasze, G., Cattaruzza, A., Douzet, F., Dammann, F., Bertran, M.-G., Bômont, C., Braun, M., Danet, D., Desforges, A., Géry, A., Grumbach, S., Hummel, P., Limonier, K., Münßinger, M., Nicolai, F., Pétiniaud, L., Winkler, J., & Zanin, C. (2023). Contested Spatialities of Digital Sovereignty. *Geopolitics*, *28*(2), 919-958. https://doi.org/10.1080/14650045.2022.2050070 Hassani, S. (2024). *Enhancing Legal Compliance and Regulation Analysis with Large Language Models* (arXiv:2404.17522). arXiv. https://doi.org/10.48550/arXiv.2404.17522 Ioannidis, J., Harper, J., Quah, M. S., & Hunter, D. (2023). *Gracenote.ai : Legal Generative AI for Regulatory Compliance* (SSRN Scholarly Paper No. 4494272). Social Science Research Network. https://doi.org/10.2139/ssrn.4494272 Kim, J., & Min, M. (2024). *From RAG to QA-RAG : Integrating Generative AI for Pharmaceutical Regulatory Compliance Process* (arXiv:2402.01717). arXiv. https://doi.org/10.48550/arXiv.2402.01717

	Liesenfeld, A., & Dingemanse, M. (2024). Rethinking open source generative AI : Open-washing and the EU AI Act. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, 1774-1787. https://doi.org/10.1145/3630106.3659005 Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance), 117 OJ L (2017). http://data.europa.eu/eli/reg/2017/745/oj Tiezzi, M., Casoni, M., Betti, A., Guidi, T., Gori, M., & Melacci, S. (2025). Back to recurrent processing at the crossroad of transformers and state-space models. *Nature Machine Intelligence*, *7*(5), 678-688. https://doi.org/10.1038/s42256-025-01034-6
Scientific material conditions scientifiques matérielles and financial resources for the research project	- Open-plan office equipped with a computer and a fixed monitor, keyboard, and mouse ; - Access to computing resources from the UTC's Inter-Laboratory Platform for Multidisciplinary Computation and Modeling (PILCAM2) ; - Access to the computing server of the Movement Technology and E-Health Meta-Platform (TecMovE) at the BMBI laboratory (funded by CPER TecSanté; acquisition and deployment planned for 2027) ; - Access to IMT's HPC servers and the dedicated persons working on LLMs within the IT department of IMT Nord Europe.
International cooperation	Research partnership with Prof. Carla TARAMASCO, director of the Instituto de Tecnología para la Innovación en Salud y Bienestar (ITISB) at Andrés Bello University (UNAB) in Chile, and Dr. Javier PEREIRA, a member of the Institute.
Planned collaborations	Télécom Sud Paris/Institut Polytechnique de Paris, S@MOVAR Lab for Distributed Services, Architectures, Modeling, Validation, and Network Administration.
Objectives of valorization for the research work	- Creation of a data management plan, to be updated throughout the project's lifespan, and to be promoted through the submission of a "data paper" for publication in a dedicated journal ("data journal"); - Provision of "tokenized" textual datasets and contextual data created from the legal texts analyzed (open science policy of the MESRE, UTC, and the BMBI laboratory); - Production of posters and scientific articles to be submitted to international conferences and congresses, as well as to peer-reviewed international scientific journals.
Confidentiality	**Confidential thesis: no**

Funding for the doctoral project	
Type of funding for the doctoral project	Higher education
Period	Start: 01/10/2026 End: 30/09/2029
Funding source	Funding of Ministry of Higher Education, Research, and Space
Employer	University of Technology of Compiegne
Status of the funding	acquired
Additional information about the funding	Abroad missions may be eligible for joint France-Chile funding, for which an application is currently underway (Incentive Program of the UTC Research Directorate – International Collaboration Initiative – 2026 Call for Proposals currently open). There are also plans to submit a joint application this year with IMTNE and ITISB at UNAB for the upcoming calls for projects "Evaluation and Orientation of Scientific Cooperation (ECOS) Southern Chile" and "Information and Communication Sciences and Technologies – South America (STIC AmSud)," which fund research missions in these geographic areas. Under the collaboration agreement with IMTNE for thesis supervision, IMTNE will contribute to expenses related to participation in scientific conferences and congresses, as well as those related to the thesis defense committee as a fixed sum per year.

Application	
Profile and skills required	Holds an engineering degree in computer engineering or a master degree in computer science or applied mathematics with a specialization in natural language processing.

	Solid theoretical background in machine learning algorithms, particularly large language models (LLMs), and knowledge graph structuring. Proven experience in producing tokenized and contextual text datasets from large text corpora (several hundred documents of several dozen pages each, ideally of a legal nature), benchmarking and evaluating LLMs, as well as improving them using RAG techniques (and ideally graph-RAG). Demonstrated proficiency in coding in Python and JavaScript, PyTorch libraries and frameworks, as well as GitLab/GitHub tools. Project management experience. Soft skills: strong proactivity and self-organization, scientific curiosity and dynamism, high level of rigor, strong commitment, adaptability to unforeseen circumstances, teamwork, and communication skills. The ability to understand and speak Spanish is a plus.
French level required	Proved C2 minimum (European reference CEFRL)
English level required	C2 appreciated (European reference CEFRL)

Contact first the thesis director before applying on the ADUM platform

Information about application for doctoral education (PhD) at UTC
on the Doctoral School website and on the ADUM platform