Laboratoire Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France

Basé à Paris-Saclay, le CEA List est l'un des quatre instituts du CEA Tech, direction de la recherche technologique du CEA. Dédié aux systèmes numériques intelligents, il contribue au développement de la compétitivité des entreprises par le développement et le transfert de technologies.

L'expertise et les compétences développées par les 800 ingénieurs-chercheurs et techniciens du CEA List permettent à l'Institut d'accompagner chaque année plus de 200 entreprises françaises et étrangères sur des projets de recherche appliquée s'appuyant sur 4 programmes et 9 plateformes technologiques. 21 start-ups ont été créées depuis 2003. Labellisé Institut Carnot depuis 2006, le CEA List est aujourd'hui l'institut Carnot Technologies Numériques. Le Laboratoire d'Analyse Sémantique des Textes et des Images (LASTI) est une équipe de 25 personnes (chercheurs, ingénieurs, doctorants) menant des travaux de recherche sur les technologies de description et de compréhension du contenu multimédia (image, texte, parole) et des documents multilingues, en particulier à grande échelle. Les enjeux scientifiques sont :

-développer des algorithmes efficaces et robustes pour l'analyse et l'extraction de contenu multimédia, leur classification et analyse sémantique ;

-reconstitution ou fusion de données hétérogènes pour interpréter des scènes ou documents ;

-développer des méthodes et des outils pour la construction, la formalisation et l'organisation des ressources et connaissances

---

Based in Paris-Saclay, CEA List is one of the four institutes under CEA Tech, the technological research branch of CEA. Specializing in intelligent digital systems, it contributes to enhancing the competitiveness of businesses through technology development and transfer.

The expertise and skills cultivated by the 800 research engineers and technicians at CEA List enable the institute to support annually over 200 French and international companies in applied research projects. These projects are based on four programs and nine technological platforms. Since 2003, 21 start-ups have been created as a result of these efforts. Designated as a "Carnot Institute" since 2006, CEA List is currently recognized as the "Digital Technologies Carnot Institute".

The Laboratory of Semantic Analysis of Texts and Images (LASTI) is a team comprising around 25 individuals, including researchers, engineers, and doctoral students. They are engaged in research activities focusing on technologies for describing and understanding multimedia content (images, text, speech) and multilingual documents, especially at a large scale. The scientific challenges include:

- Developing efficient and robust algorithms for the analysis and extraction of multimedia content, their classification, and semantic analysis
- Reconstructing or fusing heterogeneous data in order to interpret scenes or documents
- Creating methods and tools for constructing, formalizing, and organizing resources and knowledge.

Laboratoire Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France

Recherche de méthodologies frugales orientées données pour l'apprentissage et l'adaptation de modèles génératifs textuels (LLM)

Contacts :

- Evan Dufraisse evan.dufraisse@cea.fr
- Julien Tourille julien.tourille@cea.fr

Sujet de stage :

Les modèles de langage génératifs ont été, jusqu'à une période récente, au centre d'une compétition axée sur le nombre de paramètres, élément considéré alors comme central pour augmenter la généralisation et la performance des modèles [1](Kaplan et al. 2020). Toutefois, la pertinence de ce facteur s'est relativisée au regard des nouvelles lois d'échelle établies par DeepMind [2] (Hoffman et al. 2022), lois à la source du modèle Chinchilla. Ces lois permettent de définir les triplets optimaux pour le pré-entraînement de modèles dans l'espace tri-dimensionnel défini par la taille du modèle, la quantité de données et les ressources computationnelles.

Cependant, ces études accordent une attention limitée à l'impact de la qualité des données de pré-entraînement. Par ailleurs, d'autres recherches indiquent qu'il est possible de réduire substantiellement la taille des données de pré-entraînement en employant des métriques permettant de sous-échantillonner des données de haute qualité, tout en préservant la performance des modèles [3](Marion et al. 2023), [4](Sorscher et al. 2022), démontrant ainsi que les lois d'échelles se vérifient à jeu de données fixés.

Ces conclusions préliminaires suggèrent que, grâce à une optimisation des données, il est envisageable de réduire la quantité de ressources computationnelles requises pour la convergence des modèles, contrainte majeure dans l'accessibilité et la diffusion des modèles génératifs textuels au sein de la société.

Description de l'offre :

En s'intégrant dans ce contexte, l'objet du stage consiste donc à développer des méthodes de fine-tuning et de pré-entraînement frugales en ressources computationnelles pour la création ou la spécialisation de modèles de langues génératifs selon l'axe d'optimisation des données. Dans ce cadre, le candidat s'intéressera aux techniques de Data-Pruning [3](Marion et al. 2023) [5](Paul et al. 2023), de Curriculum-Learning [6](Wang et al. 2021), voire de distillation de modèles [7](Gou et al. 2021).

Plus concrètement, le stagiaire sera amené à:

- Se familiariser avec les domaines du Data-Pruning textuel, du Curriculum-Learning et de la distillation de modèles avec une étude bibliographique approfondie.
- Concevoir, en collaboration avec l'équipe encadrante, une stratégie d'entraînement ou de spécialisation pour adapter les modèles à de nouveaux domaines/tâches.

- Réaliser les expérimentations sur le cluster FactoryIA du CEA pour évaluer la stratégie envisagée.
- Améliorer de manière itérative la stratégie au regard des résultats.

Mots-clés :

- Apprentissage profond génératif, Traitement automatique des langues, Frugalité, Finetuning, Adaptation au domaine.

Profil du candidat/de la candidate

Niveau demandé :	Formation d'ingénieur et/ou M2 en informatique avec un fort intérêt pour l'apprentissage artificiel.
Durée :	6 mois
Rémunération :	entre 700 € et 1300 € suivant la formation.
Compétences requises : • environnement de travail : linux ; • notions de base en apprentissage automatique et en réseaux de neurones ; • programmation : Python + PyTorch	

Références :

[1] Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. “Scaling Laws for Neural Language Models.” arXiv, January 22, 2020. http://arxiv.org/abs/2001.08361.

[2] Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, et al. Training Compute-Optimal Large Language Models. arXiv, March 29, 2022. http://arxiv.org/abs/2203.15556

[3] Marion, Max, Ahmet, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When Less Is More: Investigating Data Pruning for Pretraining LLMs at Scale. arXiv, September 8, 2023. http://arxiv.org/abs/2309.04564

[4] Sorscher, Ben, Robert Geirhos, Surya Ganguli, Shashank Shekhar, and Ari S Morcos. Beyond Neural Scaling Laws: Beating Power Law Scaling via Data Pruning, n.d. Neurips 2022https://proceedings.neurips.cc/paper_files/paper/2022/hash/7b75da9b61eda40fa35453ee5d077df6-Abstract-Conference.html

[5] Paul, Mansheej, Surya Ganguli, and Gintare Karolina Dziugaite. Deep Learning on a Data Diet: Finding Important Examples Early in Training. arXiv, March 28, 2023. http://arxiv.org/abs/2107.07075

[6] Wang, Xin, Yudong Chen, and Wenwu Zhu. A Survey on Curriculum Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021. https://doi.org/10.1109/TPAMI.2021.3069908

[7] Gou, Jianping, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge Distillation: A Survey. International Journal of Computer Vision 129, no. 6 (June 2021) https://arxiv.org/abs/2006.05525

Laboratoire Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France

Traitement automatique des langues pour la proposition automatique d'examens à partir de comptes-rendus médicaux

Contacts :

- Romaric Besançon romaric.besancon@cea.fr
- Sondes Souihi sondes.souihi@cea.fr

Sujet de stage :

Le stage s'effectuera en co-encadrement avec l'Institut Gustave Roussy (IGR) et s'inscrit dans le cadre d'une collaboration entre le CEA LIST et l'IGR qui a pour objectif d'utiliser l'IA pour faciliter les processus administratifs des dossiers patients. Plus précisément, l'objectif du stage sera de faciliter la rédaction de demandes d'examens complémentaires après la première consultation d'un patient. Cette procédure est actuellement réalisée manuellement et est coûteuse en temps pour les médecins, alors que les informations nécessaires sont présentes dans les comptes-rendus. Le stage vise à mettre au point des méthodes pour rendre cette procédure la plus automatique possible, en utilisant les outils d'IA pour interpréter automatiquement les comptes-rendus médicaux de consultation et identifier les éléments d'intérêt qui permettent de statuer sur les examens à réaliser à la suite de la consultation.

Description de l'offre :

D'un point de vue technique, le stage relève du traitement automatique des langues (TAL ou NLP – *Natural Language Processing*) et plus particulièrement des domaines de l'extraction d'information, pour identifier les concepts médicaux pertinents dans les comptes-rendus (comme des indications anatomiques, des pathologies, des symptômes ou des traitements), de la classification automatique, pour la prise de décision des examens à réaliser et de la génération de texte pour l'aide à la rédaction des demandes. Ainsi, les travaux à réaliser dans le cadre de ce stage aborderont les points suivants :

- Mise en place d'un environnement d'évaluation pour les modèles développés: constitution d'un benchmark de référence à partir de comptes-rendus médicaux de l'IGR ;
- Mise au point de méthodes pour la décision sur les examens complémentaires : cette tâche relève de la classification automatique et pourra s'appuyer sur des méthodes d'apprentissage à base de Deep Learning et/ou sur l'exploitation de connaissances médicales spécifiques au domaine de spécialité étudié ;
- Génération de justifications pour les examens demandés : exploration de méthodes d'extraction de passages et de méthodes s'appuyant sur les modèles d'IA générative à partir de grands modèles de langue (LLM) ;

Ce travail initial pourra être poursuivi en thèse, dans un cadre plus large.

Laboratoire Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France

Profil du candidat/de la candidate

Niveau demandé :	Formation d'ingénieur et/ou M2 en informatique avec un fort intérêt pour l'apprentissage artificiel.
Durée :	6 mois
Rémunération :	entre 700 € et 1300 € suivant la formation.
Compétences requises : • environnement de travail : linux ; • notions de base en traitement automatique des langues ; • notions de base en apprentissage automatique et en réseaux de neurones (modèles de langues pré-entraînés) ; • programmation : Python + PyTorch/TensorFlow	

Les travaux de ce stage se feront en collaboration avec l'Institut Gustave Roussy.

L'institut Gustave Roussy est un institut de soins, de recherche et d'enseignement, qui prend en charge des patients atteints de tout type de cancer, à tout âge de la vie. Son expertise des cancers rares et des tumeurs complexes est internationalement reconnue. L'Institut intègre à la fois des activités de recherche fondamentale, de recherche translationnelle et de recherche clinique, sources d'innovations thérapeutiques et d'avancées diagnostiques. Gustave Roussy axe principalement ses travaux de recherche autour de la médecine personnalisée, de l'immunothérapie et de la réparation de l'ADN, ce qui fait de lui aujourd'hui le 1er centre européen de médecine personnalisée et d'immunothérapie.

Laboratoire Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France

Large Language Models for Information Extraction

Contacts :

- Olivier Ferret olivier.ferret@cea.fr
- Julien Tourille julien.tourille@cea.fr

Sujet de stage

Large Language Models (LLM) have been widely adopted by the Natural Language Processing (NLP) community and have been applied with success to a variety of tasks (Le Scao *et al.*, 2023; Touvron *et al.*, 2023). These models have been pretrained in a self-supervised fashion on large corpora of raw text and have been tested on standardized benchmarks devised by the community. Most of the time, these benchmarks include Natural Language Understanding (NLU) tasks such as reasoning and common sense in a variety of domains (e.g. microeconomics, physics or maths) (Hendrycks et al., 2021; Srivastava et al., 2023). Other evaluate the capacities of these models to generate code or to translate a program into another language (Zheng et al., 2023). Only a few research efforts concentrate on evaluating these models on information extraction tasks. Among them, Wang et al. (2023) introduce IE INSTRUCTIONS, a benchmark composed of 32 information extraction datasets that includes Named Entity Recognition (NER), Relation Extraction (RE) and Event Extraction (EE) tasks.

Description de l'offre

In this context, we propose to further study the performance of LLMs on Information Extraction tasks. Specifically, this study will focus on their few- and zero-shot capabilities for NER in a context where the number of types of entities to identify in texts is very high, which results in a very small volume of annotated data. Among other tasks, the successful intern will have the following responsibilities:

- Perform and maintain an up-to-date literature review on the topic
- Propose and devise an evaluation framework for the application of LLMs to NER in a few- and zero-shot setting
- Evaluate state-of-the-art models in this framework by relying in our computing cluster
- Devise, implement, and evaluate new methods for zero- and few-shot NER using pretrained LLMs

This initial work could be further investigated in a following PhD starting in October 2024, whose topic will be refined with the successful candidate.

Laboratoire Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France

Profil du candidat/de la candidate

Niveau demandé :	The ideal candidate should have an engineering and/or Master 2 (M2) degree in computer science with a strong interest in artificial intelligence and natural language processing. She/He must have an interest in scientific research.
Durée :	6 months
Rémunération :	700 € - 1300 €
Non-exhaustive list of required skills : • Able to work in a Linux environment • Background in natural language generation and language modeling • Familiarity with pre-trained language models and large language models • Familiarity with Python and specifically with pytorch and other AI/NLP related libraries	

References

Hendrycks et al. (2021). *"Measuring Massive Multitask Language Undestanding"*. ICLR
Le Scao et al. (2023). *"BLOOM: A 176B-Parameter Open-Access Multilingual Language Model"*. arXiv 2211.05100
Srivastava et al. (2023). *"Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*". arXiv 2206.04615
Touvron et al. (2023). *"Llama 2: Open Foundation and Fine-Tuned Chat Models".* arXiv 2307.09288
Wang et al. (2023). *"InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction"*. arXiv 2304.08085
Zheng et al. (2023). *"CodeGeeX: A Pre-Trained Model for Code Generation with Multilingual Evaluations on HumanEval-X".* arXiv 2303.17568

Laboratoire Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France

Apprentissage auto-supervisé de contenus visuels sous contraintes

Contact :

- Adrian Popescu adrian.popescu@cea.fr

Sujet de stage :

Le stage s'inscrit dans le cadre du projet de recherche collaborative PROPEOS dont l'objectif global est de proposer des systèmes de recommandation respectueux de la vie privée.Un des cas d'usage du projet est la recommandations touristiques personnalisées aux voyageurs à partir de l'analyse de leurs images. Les outils d'apprentissage profond qui peuvent être déployés sur les smartphones des utilisateurs sont une composante essentielle de ces systèmes. Ils permettent des inférences utiles pour la recommandation sans compromettre la vie privée. L'apprentissage auto-supervisé est devenu un standard de fait pour entraîner des modèles profonds généralistes. Toutefois, la plupart de ces modèles font l'hypothèse que leur déploiement se fera sans contraintes calculatoire et que les données disponibles sont très abondantes. Ces hypothèses ne sont pas vérifiées dans le cadre de PROPEOS. L'objectif du stage est de proposer des techniques d'apprentissage auto-supervisé qui soient utilisables sur smartphone, tout en conservant la qualité des gros modèles actuels.

Description de l'offre :

Techniquement, le stage relève de la vision par ordinateur, et plus précisément de la classification d'images pour le domaine du tourisme. En collaboration avec des ingénieurs chercheurs du CEA, il s'agira d'étudier la faisabilité de l'auto-supervision pour des modèles profonds frugaux en paramètres, d'entraîner de tels modèles et d'évaluer leur utilisation dans le cadre applicatif de PROPEOS. Le stage est conçu comme une initiation à la recherche et aura comme objectif la publication d'un article scientifique si les résultats obtenus sont probants.

Ce travail initial pourra être poursuivi en thèse, dans un cadre plus large.

Profil du candidat/de la candidate

Niveau demandé :	Formation d'ingénieur et/ou M2 en informatique avec un fort intérêt pour l'apprentissage artificiel.
Durée :	6 mois
Rémunération :	entre 700 € et 1300 € suivant la formation.
Compétences requises : • environnement de travail : linux • connaissance des techniques en vision par ordinateur. • connaissances en apprentissage automatique et en réseaux de neurones. • programmation : Python + PyTorch/TensorFlow	

Laboratoire Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France

Extraction d'information à partir de documents PDF complexes

Contact :

- Benjamin Labbé benjamin.labbe@cea.fr
- Anne-Laure Daquo Anne-Laure.daquo@cea.fr
- Nicolas Allezard nicolas.allezard@cea.fr

Sujet de stage :

Le format PDF (Portable Document Format) crée en 1992 et aujourd'hui largement adopté, préserve la mise en page des documents telle que définie par leur auteur. Les outils NLP nécessitent pour la plupart de convertir chaque PDF en texte brut séquentiel. Malheureusement la structure du document ou de tableaux révélée par la mise en forme est souvent détériorée.

Historiquement, les travaux en extraction d'information se sont d'abord concentrés sur les données non-structurées (texte brut), puis peu à peu le sujet s'est ouvert au traitement des données semi-structurés plus largement vers 2008, sans devenir majeur. Le traitement de ce type de données constitue un défi actuel en extraction d'information.

Il existe à l'heure actuelle nombre d'outils pour océriser les documents PDF ou image : Amazon Textract , Google API Cloud Vision, ABBYY FineReader PDF, OCRmobile (Meelo), OCRmyPDF, Tesseract 4.0 (open source), Parsr (open source), etc. Une fois OCRisé, la seconde étape d'une approche classique est d'extraire et de structurer le texte pour reconstituer les paragraphes, identifier les titres de section et organiser le contenu des tableaux. La plupart des outils d'extraction de texte sont open-source, parmi lesquels on trouve pdftotext, PdfMiner, Tika, Grobid, etc.

Ont émergés récemment des outils génératifs tels Vision LLM, LLava, GPT-4V qui pourraient permettre une meilleure reconnaissance et sémantisation de la structure des PDF et d'éléments internes (comme les tableaux). Ces modèles pourraient in fine améliorer les performances d'extraction d'information et autres outils NLP.

Description de l'offre

L'objectif de ce stage est d'évaluer l'influence de différentes solutions d'OCR et d'extraction de texte, voire de LLM sur la performance en extraction d'informations à partir de documents PDF complexes. Le stage s'intéressera aussi à la qualité de l'extraction d'information dans du texte et des tableaux. Cette influence sera évaluée en termes d'impact quantitatif sur la capacité d'extraction d'entités (concepts) et de relations binaires ou n-aires d'un outil pensé pour prendre en entrée du texte brut.

Le stage se déroulera selon les étapes suivantes :

- Identifier et se familiariser avec les divers outils logiciels utilisés actuellement pour effectuer l'OCR et l'extraction de texte voire des outils génératifs orientés vision tels Vision LLM, LLava, GPT 4V, etc.
- Se familiariser avec les outils logiciels du laboratoire (et alternatives open-source) pour l'extraction d'informations : LIMA, SpaCy
- Évaluer les différents outils et approches grâce à
    - la constitution d'un jeu d'évaluation à partir de documents PDF internes et/ou l'identification d'un jeu de données académique,
    - la mise en place de pipelines de traitement de documents PDF pour l'extraction d'information (pour une évaluation indirecte à travers cette tâche NLP)
    - Une analyse quantitative et/ou qualitative de l'extraction d'information comparant les performances dans différentes structures des documents PDF : texte vs tableaux. Propositions d'amélioration.
- Optimisation des pipelines de traitement de documents les plus prometteurs pour améliorer les performances.
- Rédaction du rapport de stage.

Pour débuter :

- Tutoriel introductif ACL 2020 dédié à l'extraction d'information à partir de données non-, semi-structurées : https://sites.google.com/view/acl-2020-multi-modal-ie

Profil du candidat/de la candidate

Niveau demandé :	Formation d'ingénieur et/ou M2 en informatique avec un fort intérêt pour l'apprentissage artificiel.
Durée :	6 mois
Rémunération :	entre 700 € et 1300 € suivant la formation.

Compétences requises :

- Environnement de travail : Linux
- Maîtrise d'un langage de programmation : Python (ou C++)
- Maîtrise des méthodes d'évaluation des modèles de Machine Learning ou Deep Learning en NLP
- Notion de base en apprentissage automatique et en réseaux de neurones
- Notions de base en traitement automatique des langues et en vision par ordinateur.
- Expérience appréciée avec une bibliothèque de type Transfomers, Tensorflow, PyTorch, etc.

Laboratoire Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France

Detection of AI-generated text-based cybersecurity attacks

Contact :

- Sondes Souihi sondes.souihi@cea.fr

Internship subject :

The rise of artificial intelligence has revolutionized not only the way we live and work but also the tactics employed by cyber adversaries. The emergence of AI-generated cybersecurity attacks has paved the way for a new era of digital threats. AI-generated text-based cybersecurity attacks represent a new breed of cyber threats where artificial intelligence technologies are used to create and execute different malicious activities (phishing, spear phishing, fake news, disinformation, social manipulation, etc). These attacks leverage text generation models, such as GPT-3, to create convincing and contextually relevant messages, emails, or other forms of textual content. The primary goal of these attacks is to deceive individuals, systems and even nations, leading to various harmful consequences. In this context, it becomes imperative to understand the threats brought by AI-generated text-based cybersecurity attacks and develop innovative strategies to mitigate them. The aim of this internship consists in developing AI-based techniques of detecting AI-generated text-based cyber attacks in order to equip network experts with precise tools for identifying patterns of misuse and malicious behaviors generated by AI.

Internship description:

Technically, the internship involves the fields of machine learning (ML) and natural language processing (NLP), and more specifically natural language generation (NLG) and classification of texts based on their authorship (human, AI, specific generative model). In collaboration with CEA research engineers, the aim will be to train classification models capable of distinguishing texts written by humans from those generated by AI or by a particular generative model, and to evaluate them in the cybersecurity field. This internship is meant to be an introduction to research, with the goal of publishing a scientific article if the obtained results are conclusive. The implemented models may also be used to participate in a shared task like AuTexTification (https://sites.google.com/view/autextification/home) and CLIN33 (https://sites.google.com/view/shared-task-clin33/home) or in a challenge like MLMAC (https://mlmac.io/).

This work may be followed by a PhD in a broader context.

Laboratoire Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France

Profil du candidat/de la candidate

Niveau demandé :	Engineering degree and/or Master 2 (M2) degree in computer science with a strong interest in artificial intelligence and natural language processing.
Durée :	6 mois
Rémunération :	entre 700 € et 1300 € suivant la formation.

Required skills :

- working environment : linux
- knowledge of text classification techniques
- background in natural language generation and language modeling
- familiarity with pre-trained language models and large language models
- basic knowledge of the cybersecurity field
- programming : Python + PyTorch/TensorFlow

Laboratoire Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France

Evaluation de la fiabilité et de la robustesse des grands modèles de langue sur des tâches d'annotation en TAL

Contact :

- Nasredine Semmar nasredine.semmar@cea.fr

Description du stage :

Les modèles de langue tels que BERT et GPT-3 affichent des performances impressionnantes sur une variété de tâches du Traitement Automatique des Langues (TAL) et leur fine-tuning permet de spécialiser ces modèles génériques en modèles performants et spécifiques. Cependant, le fait que le fonctionnement interne de ces modèles est difficile à saisir, constitue un frein à leur utilisation dans des applications nécessitant un niveau élevé de fiabilité et de robustesse comme c'est le cas dans le domaine de la santé.

Le stage consistera, d'une part, à constituer un framework pour l'évaluation des performances des grands modèles de langue (LLMs, pour Large Language Models) et de leur utilisation dans le domaine de la santé, et d'autre part, à évaluer ces modèles sur des tâches de structuration et d'extraction d'informations à partir de comptes-rendus médicaux. Cette évaluation sera réalisée selon deux approches différentes : une évaluation humaine à petite échelle dans laquelle les prédictions de ces modèles seront comparées à une référence créée manuellement par des praticiens de santé, et une évaluation automatique utilisant des métriques pour les tâches de structuration et d'extraction d'informations.

Le stage se déroulera selon les étapes suivantes:

• Etude bibliographique sur les méthodes et outils d'évaluation des grands modèles de langue (Chang et al., 2023 ; Zeng et al. 2023).

• Recherche de benchmarks académiques pour l'évaluation des LLMs dans le domaine de la santé (Reddy, 2023 ; Tang et al., 2023).

• Constitution d'un dataset de référence pour l'évaluation des LLMs sur des tâches de structuration et d'extraction d'informations à partir de comptes-rendus médicaux.

• Réalisation d'une évaluation comparative des performances des LLMs en zero ou few shot learning pour le NER sur des comptes rendus médicaux fictifs. Dans l'analyse comparative, des modèles équivalents à GPT 3, GPT 4, falcon, Llama, Bloom peuvent être étudiés.

• Evaluation manuelle (humaine) et automatique des LLMs sur diverses tâches de TAL. Un intérêt particulier sera accordé à l'étude du phénomène d'hallucination des LLMs (Ji et al., 2023)

• Développer une chaine de traitements pour analyser, interpréter et contrôler les prédictions des LLMs.

Mots-clés :

Traitement automatique des langues, grands modèles de langue, extraction d'informations, benchmarks d'évaluation.

Références:

• Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, Xing Xie. 2023. A Survey on Evaluation of Large Language Models. arXiv.

• Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, Danqi Chen. 2023. Evaluating Large Language Models at Evaluating Instruction Following. arXiv.

• Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, Vol. 55, No. 12.

• Sandeep Reddy. 2023. Evaluating large language models for use in healthcare: A framework for translational value assessment. Informatics in Medicine, No. 41.

• Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, Yifan Peng. 2023. Digital Medicine, 6:158.

Profil du candidat/de la candidate

Niveau demandé :	Ingénieur, Master 2
Durée :	6 mois
Rémunération :	entre 700 € et 1300 € suivant la formation.

Compétences requises :

- Environnement de travail : linux
- Maîtrise d'un langage de programmation : Python (ou C++)
- Maîtrise des méthodes d'évaluation des modèles de Machine Learning ou Deep Learning en NLP
- Expérience avec une bibliothèque de type Transfomers, Tensorflow, PyTorch, etc.
- Notion de base en apprentissage automatique et en réseaux de neurones
- Notions de base en traitement automatique des langues