Research internships 2025

"Responsible and Trustworthy AI" chair Polytechnique / Crédit Agricole

Laboratoire d'Informatique de l'École Polytechnique (LIX) – Orailix team
Palaiseau, France

Four research internships

This booklet presents the four research internships proposed in the context of the research chair between École Polytechnique and Crédit Agricole. The internships will take place at LIX in Palaiseau, within the Orailix team.

1. Safety of multi LLM-agents for information retrieval
2. Explainability of multi LLM-agents for information retrieval
3. Hallucination and memorization in generative document models
4. State space models for document AI

How to apply?

Please send your CV and a short paragraph to introduce yourself by email to jeremie.dentan@polytechnique.edu and mohamed.dhouib@polytechnique.edu with subject "Application: internship 2025".

Safety of multi LLM-agents for information retrieval

Internship proposal

Laboratoire d'Informatique de Polytechnique (LIX) – Orailix team – Palaiseau, France
Research chair "Responsible and Trustworthy AI" Polytechnique / Crédit Agricole SA

INTRODUCTION

Large Language Models are increasingly used as goal-driven agents interacting with external environment to solve a wide range of tasks, such as developing source code to solve a problem, analyzing a bibliography to answer a question, exploring an environment to play a game [3, 6–9].

These *scaffoldings* containing multiple LLM agents have demonstrated very strong performance and reliability, sometime being the first computer-based solution able to solve a task [5]. However, their deployment raises security challenges such underspecification of their goal, undesirable behaviors, and emergent abilities of groups of LLM agents [1].

Our team propose two internships to explore the benefits of multi LLM-agents for information retrieval. The first subject will focus on the safety of the system and its robustness to undesirable behaviors. The second one will focus on designing explainable multi LLM-agents. **This proposal concerns the first subject, focusing on safety.**

OBJECTIVE AND METHODOLOGY

First, the intern will make a bibliographic review of existing multi LLM-agents approaches applied to information retrieval. The intern will also review existing work on LLM robustness, hallucination and undesirable behaviors.

Second, the intern will develop a framework to assess information retrieval solutions on academic benchmarks such as RAGbench [2] or internal data at Crédit Agricole. This framework should measure the accuracy (mean performance), reliability (proportion above a minimal performance threshold), robustness to under-specified tasks, apparition of undesirable behaviors (hallucination, harmful content), as well as the computation time and cost.

Third, the intern will develop a multi LLM-agents approach and optimize it. The intern will compare his/her solution to traditional RAG approaches [4], and evaluate the impact of the architecture and of the choice of LLM for the system.

When implementing these experiments, the intern will have to collaborate with another intern focusing on the explainability of multi LLM-agents. A particular attention will be paid to the quality of the implementation, its adaptability, and its reproducibility.

EXPECTED ABILITIES

We are looking for a Master's level intern with a background in machine learning, applied mathematics, natural language processing, and/or computer vision. An important part of the internship will be devoted to implementing new machine learning methods. In addition to scientific skills, the intern must have an appetite for software development. The intern should be familiar with libraries such as `transformers` and `PyTorch`.

Promoting diversity in science. It has been observed that women tend to hesitate more than men to apply for jobs when they are not 100% qualified.[1] We recall that prior experience with multi-agents systems and information retrieval is not a prerequisite for the internship. The most important qualification is scientific curiosity and the ability to progress in these fields with help and guidance.

PRACTICAL DETAILS

The internship will take place within the Orailix team of the Laboratoire d'Informatique de Polytechnique (LIX). Depending on the course of the internship and the intern's interest, this work can be continued in a PhD thesis.

The internship will take place within the "Responsible and Trustworthy AI" research chair between École Polytechnique and Crédit Agricole SA.

Internship supervision.

- Sonia Vanier (Professor)
- Jesse Read (Professor)
- Jérémie Dentan (PhD Candidate)

To apply. Please send your CV and a short paragraph to introduce yourself by email to jeremie.dentan@polytechnique.edu and mohamed.dhouib@polytechnique.edu with subject "Application: internship 2025".

Joint application. We plan to recruit two interns to work collaboratively on multi LLM-agents: one with a focus on safety, and one with a focus on explainability. If you have a peer from your program with whom you enjoy collaborating, we encourage you both to apply and mention this in your applications.

REFERENCES

[1] Usman Anwar and al. 2024. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. http://arxiv.org/abs/2404.09932
[2] Robert Friel and al. 2024. RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems. http://arxiv.org/abs/2407.11005
[3] Wenlong Huang and al. 2023. Inner Monologue: Embodied Reasoning through Planning with Language Models. In *6th Conference on Robot Learning*. https://proceedings.mlr.press/v205/huang23c.html
[4] Patrick Lewis and al. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. http://arxiv.org/abs/2005.11401
[5] Meta Fundamental AI Research Diplomacy Team (FAIR)† and al. 2022. Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning. *Science* (Dec. 2022). https://doi.org/10.1126/science.ade9097
[6] Joon Sung Park and al. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *ACM UIST*. https://doi.org/10.1145/3586183.3606763
[7] Noah Shinn and al. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. http://arxiv.org/abs/2303.11366
[8] Guanzhi Wang and al. 2023. Voyager: An Open-Ended Embodied Agent with Large Language Models. http://arxiv.org/abs/2305.16291
[9] Shunyu Yao and al. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. http://arxiv.org/abs/2210.03629

[1] https://hbr.org/2014/08/why-women-dont-apply-for-jobs-unless-theyre-100-qualified

Explainability of multi LLM-agents for information retrieval

Internship proposal

Laboratoire d'Informatique de Polytechnique (LIX) – Orailix team – Palaiseau, France
Research chair "Responsible and Trustworthy AI" Polytechnique / Crédit Agricole SA

INTRODUCTION

Large Language Models are increasingly used as goal-driven agents interacting with external environment to solve a wide range of tasks, such as developing source code to solve a problem, analyzing a bibliography to answer a question, exploring an environment to play a game [3, 5–8].

These *scaffoldings* containing multiple LLM agents have demonstrated very strong performance and reliability, sometime being the first computer-based solution able to solve a task [4]. However, their deployment raises security challenges such underspecification of their goal, undesirable behaviors, and emergent abilities of groups of LLM agents [1].

Our team propose two internships to explore the benefits of multi LLM-agents for information retrieval. The first subject will focus on the safety of the system and its robustness to undesirable behaviors. The second one will focus on designing explainable multi LLM-agents. **This proposal concerns the second subject, focusing on explainability.**

OBJECTIVE AND METHODOLOGY

First, the intern will make a bibliographic review of existing multi LLM-agents approaches applied to information retrieval. The intern will also review existing work on LLM explainability techniques and how it is evaluated in the literature.

Second, the intern will develop a framework to assess information retrieval solutions on academic benchmarks such as RAGbench [2] or internal data at Crédit Agricole. This framework should measure the accuracy (mean performance), computation time and cost, and, most importantly, the quality of the explanation in the output of the system and its faithfulness.

Third, the intern will develop a multi LLM-agents approach and optimize it. The intern will compare his/her solution to traditional explainability approaches, and evaluate the impact of the architecture and of the choice of LLM for the system.

When implementing these experiments, the intern will have to collaborate with another intern focusing on the safety of multi LLM-agents. A particular attention will be paid to the quality of the implementation, its adaptability, and its reproducibility.

EXPECTED ABILITIES

We are looking for a Master's level intern with a background in machine learning, applied mathematics, natural language processing, and/or computer vision. An important part of the internship will be devoted to implementing new machine learning methods. In addition to scientific skills, the intern must have an appetite for software development. The intern should be familiar with libraries such as `transformers` and `PyTorch`.

Promoting diversity in science. It has been observed that women tend to hesitate more than men to apply for jobs when they are not 100% qualified.[1] We recall that prior experience with multi-agents systems and information retrieval is not a prerequisite for the internship. The most important qualification is scientific curiosity and the ability to progress in these fields with help and guidance.

PRACTICAL DETAILS

The internship will take place within the Orailix team of the Laboratoire d'Informatique de Polytechnique (LIX). Depending on the course of the internship and the intern's interest, this work can be continued in a PhD thesis.

The internship will take place within the "Responsible and Trustworthy AI" research chair between École Polytechnique and Crédit Agricole SA.

Internship supervision.

- Sonia Vanier (Professor)
- Jesse Read (Professor)
- Jérémie Dentan (PhD Candidate)

To apply. Please send your CV and a short paragraph to introduce yourself by email to jeremie.dentan@polytechnique.edu and mohamed.dhouib@polytechnique.edu with subject "Application: internship 2025".

Joint application. We plan to recruit two interns to work collaboratively on multi LLM-agents: one with a focus on safety, and one with a focus on explainability. If you have a peer from your program with whom you enjoy collaborating, we encourage you both to apply and mention this in your applications.

REFERENCES

[1] Usman Anwar and al. 2024. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. http://arxiv.org/abs/2404.09932
[2] Robert Friel and al. 2024. RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems. http://arxiv.org/abs/2407.11005
[3] Wenlong Huang and al. 2023. Inner Monologue: Embodied Reasoning through Planning with Language Models. In *6th Conference on Robot Learning*. https://proceedings.mlr.press/v205/huang23c.html
[4] Meta Fundamental AI Research Diplomacy Team (FAIR)† and al. 2022. Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning. *Science* (Dec. 2022). https://doi.org/10.1126/science.ade9097
[5] Joon Sung Park and al. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *ACM UIST*. https://doi.org/10.1145/3586183.3606763
[6] Noah Shinn and al. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. http://arxiv.org/abs/2303.11366
[7] Guanzhi Wang and al. 2023. Voyager: An Open-Ended Embodied Agent with Large Language Models. http://arxiv.org/abs/2305.16291
[8] Shunyu Yao and al. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. http://arxiv.org/abs/2210.03629

[1] https://hbr.org/2014/08/why-women-dont-apply-for-jobs-unless-theyre-100-qualified

Hallucination and memorization in generative document models

Internship proposal

Laboratoire d'Informatique de Polytechnique (LIX) – Orailix team – Palaiseau, France
Research chair "Responsible and Trustworthy AI" Polytechnique / Crédit Agricole SA

INTRODUCTION

Generative document models are transformer-based models designed to understand visually-rich documents and generate textual content to solve specific tasks such as visual question answering or information extraction [2–4]. These models are increasingly employed by companies to supplant humans in processing sensitive documents, such as invoices, tax notices, or even ID cards.

However, it has been demonstrated that these models can hallucinate answers that do not exist in the input document. Worse still, these hallucinations sometime contain sensitive personal data from their training set [6]. Indeed, it has been observed that document understanding models memorize part of their training data, which can be exposed at inference time [1].

The goal of this internship is to explore hallucination and memorization in generative document models, as well as the relationship between these two phenomena and the factors influencing them.

OBJECTIVE AND METHODOLOGY

We aim to accurately detect hallucinated outputs and assess the reasons why the model resorted to hallucination instead of abstaining from responding. Moreover, we intend to explore the criteria for whether or not an hallucination contain personal data memorized from the training set. The purpose of this work is to develop mitigation techniques to monitor models at inference time and prevent the harmful consequences of these phenomena.

The experiments will be conducted on pre-trained OCR-free generative document models such as Donut [3], Pix2Struct [4] or DocParser [2]. The intern will use open-source data such as the popular Document Visual Question Answering (DocVQA) dataset [5].

The intern will develop a methodology to measure and detect hallucinations and private information leakage from the training set. Then, the intern will evaluate the factors that influence these phenomena using a systematic exploration of the input space as well as the similarity between the inputs and documents from the training set of these models.

When implementing these experiments, particular attention will be paid to the reproducibility of the results and their applicability to real-life scenarios, with the aim of promoting open and reproducible research and software.

EXPECTED ABILITIES

We are looking for a Master's level intern with a background in machine learning, applied mathematics, natural language processing, and/or computer vision. An important part of the internship will be devoted to implementing new machine learning methods. In addition to scientific skills, the intern must have an appetite for software development. The intern should be familiar with libraries such as `transformers` and `PyTorch`.

Promoting diversity in science. It has been observed that women tend to hesitate more than men to apply for jobs when they are not 100% qualified.[1] We recall that prior experience with document models, hallucination or memorization is not a prerequisite for the internship. The most important qualification is scientific curiosity and the ability to progress in these fields with help and guidance.

PRACTICAL DETAILS

The internship will take place within the Orailix team of the Laboratoire d'Informatique de Polytechnique (LIX). Depending on the course of the internship and the intern's interest, this work can be continued in a PhD thesis.

The internship will take place within the "Responsible and Trustworthy AI" research chair between École Polytechnique and Crédit Agricole SA.

Internship supervision.

- Sonia Vanier (Professor)
- Jérémie Dentan (PhD Candidate)
- Mohamed Dhouib (PhD Candidate)

To apply. Please send your CV and a short paragraph to introduce yourself by email to jeremie.dentan@polytechnique.edu and mohamed.dhouib@polytechnique.edu with subject "Application: internship 2025".

REFERENCES

[1] Jérémie Dentan, Arnaud Paran, and Aymen Shabou. 2024. Reconstructing training data from document understanding models. In *USENIX Security*. USENIX Association, Philadelphia, PA, 6813–6830. https://www.usenix.org/conference/usenixsecurity24/presentation/dentan

[2] Mohamed Dhouib, Ghassen Bettaieb, and Aymen Shabou. 2023. DocParser: End-to-end OCR-Free Information Extraction from Visually Rich Documents. In *Document Analysis and Recognition - ICDAR 2023*. Vol. 14191. Springer Nature Switzerland, Cham, 155–172. https://doi.org/10.1007/978-3-031-41734-4_10

[3] Geewook Kim, Teakgyu Hong, Moonbin Yim, and al. 2022. OCR-Free Document Understanding Transformer. In *ECCV*. https://arxiv.org/abs/2111.15664

[4] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, and al. 2023. Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 18893–18912. https://proceedings.mlr.press/v202/lee23g.html

[5] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A Dataset for VQA on Document Images. In *IEEE/CVF WACV*. Online. http://arxiv.org/abs/2007.00398

[6] Francesco Pinto, Nathalie Rauschmayr, Florian Tramèr, and al. 2024. Extracting Training Data from Document-Based VQA Models. In *ICML*. arXiv. http://arxiv.org/abs/2407.08707

[1] https://hbr.org/2014/08/why-women-dont-apply-for-jobs-unless-theyre-100-qualified

State space models for document AI

Research Proposal

Laboratoire d'Informatique de Polytechnique (LIX) – Orailix team – Palaiseau, France
Research Chair "Responsible and Trustworthy AI" Polytechnique / Crédit Agricole SA

INTRODUCTION

Recent advancements in sequence modeling, particularly architectures like Mamba [1] and Jamba [3], have enabled more efficient handling of long-sequence data, especially in natural language processing. Originally explored with a focus on text, these architectures offer efficient solutions by addressing traditional limitations in attention-based architectures. Mamba, for instance, achieves linear-time complexity through a selective state space model (SSM), which allows it to manage long-range dependencies without the heavy memory demands typical of Transformers. This efficiency makes Mamba interesting for applications in Document AI where high-resolution images are used.

This objective of this internship is to investigate the potential of Mamba and similar architectures in Document AI tasks, exploring their ability to handle multi-modal data.

OBJECTIVE AND METHODOLOGY

The objective of this project is to assess the applicability of architectures like Mamba [1, 3] for Document AI applications, especially those involving long documents or multi-page formats. Our methodology will include:

- **Literature Review:** Conduct an in-depth review of existing work on state space models, focusing on how these architectures have been applied to sequence modeling tasks and multi-modal data [5], particularly in contrast to traditional attention-based models. This review will highlight the strengths of Mamba and similar architectures in handling long-sequence and multi-modal data efficiently, setting the foundation for their potential in Document AI.
- **Concept Development:** Develop an approach to apply state space models to multi-modal Document AI tasks. This includes conceptualizing how these architectures can be adapted to process a combination of text and high-resolution text-rich images.
- **Implementation and Benchmarking:** Implement the proposed architecture and rigorously benchmark it on established Document AI datasets such as SROIE [2] and DocVQA [4]. This evaluation will focus on computational efficiency, and accuracy, comparing the state space models' effectiveness to traditional architectures.

EXPECTED ABILITIES

We are looking for a Master's level intern with a background in machine learning, applied mathematics, natural language processing, and/or computer vision. An important part of the internship will be devoted to implementing new machine learning methods. In addition to scientific skills, the intern must have an appetite for software development. Proficiency in `PyTorch` is essential, and familiarity with `transformers` would be beneficial.

Promoting diversity in science. It has been observed that women tend to hesitate more than men to apply for jobs when they are not 100% qualified.[1] We recall that prior experience with document models, or Mamba/Jamba is not a prerequisite for the internship. The most important qualification is scientific curiosity and the ability to progress in these fields with help and guidance.

PRACTICAL DETAILS

The internship will take place within the Orailix team of the Laboratoire d'Informatique de Polytechnique (LIX). Depending on the course of the internship and the intern's interest, this work can be continued in a PhD thesis.

The internship will take place within the "Responsible and Trustworthy AI" research chair between École Polytechnique and Crédit Agricole SA.

Internship supervision.

- Sonia Vanier (Professor)
- Mohamed Dhouib (PhD Candidate)

To apply. Please send your CV and a short paragraph to introduce yourself by email to jeremie.dentan@polytechnique.edu and mohamed.dhouib@polytechnique.edu with subject "Application: internship 2025".

REFERENCES

[1] Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).
[2] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1516–1520.
[3] Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. 2024. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887* (2024).
[4] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2200–2209.
[5] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417* (2024).

[1] https://hbr.org/2014/08/why-women-dont-apply-for-jobs-unless-theyre-100-qualified