REF : DIT 2- 2026-IAG1
Pour postuler : faicel.chamroukhi@irt-systemx.fr – Merci de mentionner le numéro de référence

Thèse de Doctorat

Apprentissage frugal de modèles génératifs multimodaux pour la gestion de connaissances en contexte industriel à partir de données spécifiques

CONTEXTE DE LA THESE

Au sein de l'Institut de Recherche Technologique SystemX, situé au cœur du campus scientifique d'excellence mondiale de Paris-Saclay, vous prendrez une part active au développement d'un centre de recherche technologique de niveau international dans le domaine de l'ingénierie numérique des systèmes. Adossé aux meilleurs organismes de recherche français du domaine et constitué par des équipes mixtes d'industriels et d'académiques, l'institut a pour mission de générer de nouvelles connaissances et solutions technologiques en s'appuyant sur les percées de l'ingénierie numérique et de diffuser ses compétences dans tous les secteurs économiques.

Plus particulièrement au sein de l'IRT SystemX, le doctorant sera rattaché à l'axe « sciences des données, IA et Interaction ». Le sujet de thèse a été défini dans le cadre du projet IAG1 « Gestion de connaissances techniques en ingénierie des systèmes complexes » du programme « IA Générative pour l'industrie » (IAG) de l'institut. Le projet vise à développer un cadre scientifique et technique permettant d'évaluer des modèles d'IAG sur des benchmarks et des cas d'usage industriels ou industrialisables, et de les spécialiser sur des données multimodales collectées dans les corpus de connaissances techniques des ingénieries. Son objectif est d'assister la prise de décision en générant des réponses adaptées aux métiers de l'ingénierie, sous forme de rapports de différentes natures comprenant des résumés, des analyses et des recommandations de scénarios métier, alignés sur les connaissances techniques sous-jacentes aux données multimodales exploitées.

La direction de la thèse sera assurée par Faïcel Chamroukhi, professeur des universités et responsable scientifique de l'axe 1 à l'IRT SystemX. La thèse sera inscrite à l'école doctorale STIC de l'Université Paris-Saclay (#580). Le poste est basé sur le site de l'IRT SystemX à Palaiseau. Hormis les déplacements en conférences internationales, un séjour dans un laboratoire à l'étranger peut être envisagé.

La date souhaitée de démarrage de la thèse est le 01/01/2026.

La rémunération de la thèse est de 2784€ brut mensuel sur 3 ans.

SUJET DE THESE

1. Contexte et motivations

Les données produites par les systèmes d'ingénierie en milieu industriel sont multimodales et complexes : données de capteurs, rapports de diagnostic et de maintenance de pannes, dessins et schémas techniques, référentiels techniques, normes et standards,... La richesse et l'hétérogénéité de ces données ouvrent des opportunités uniques pour l'aide à la décision et à la conception en industrie. Leur exploitation optimisée est un levier majeur d'amélioration du processus industriel dans son ensemble et est un vecteur de compétitivité économique.

Toutefois, ces données posent aussi des défis scientifiques. Elles sont hétérogènes, de modalités variées, souvent incomplètes, notamment dans le cas de situations rarement observées. Leur intégration cohérente représente un enjeu clé pour les acteurs industriels.

La recherche menée dans cette thèse porte principalement sur l'exploitation simultanée et efficace de données hétérogènes, collectées et/ou simulées à partir de différentes sources en ingénierie de systèmes complexes. Trois défis scientifiques principaux structurent la thèse :

1. La multimodalité : comment représenter de façon cohérente des données hétérogènes (texte, séries temporelles, diagrammes/schémas techniques, données tabulaires), tout en préservant et en intégrant la sémantique métier.
2. La frugalité liée aux données, mais aussi aux modèles et à leur apprentissage : comment traiter et générer de telles données efficacement, en temps et en espace (manque de données/annotations, contraintes de ressources, besoin de temps réel).
3. Apprentissage multitâche : pouvoir exploiter ces représentations dans diverses tâches (prévision et détection d'anomalies, génération de rapports, augmentation de données et génération de données synthétiques, interaction et questions-réponses) et s'en servir sur de nouveaux cas d'usage.

Ces problématiques concernent en particulier l'apprentissage de modèles génératifs frugaux à partir de modalités au-delà du seul texte, notamment des séries temporelles, dessins et images techniques, diagrammes, ainsi que des données tabulaires, avec des contraintes liées à l'intégration du contexte et des connaissances métier.

Aujourd'hui, les modèles de fondation multimodaux [20] appelés General Purpose AI (GPAI) models, entraînés sur de gros corpus hétérogènes, sont prometteurs pour apprendre des représentations intégrées de données très diverses (texte, image, séries temporelles). Leur potentiel reste largement inexploré pour les données industrielles, qui présentent des contraintes et des exigences spécifiques (volumétrie, qualité, spécialisation, ressources limitées).

Cette thèse a pour objectif de proposer des modèles génératifs probabilistes multimodaux légers, capables d'apprendre des représentations communes de données industrielles, éventuellement partagées entre modalités, et de générer des données simulées cohérentes, sous des contraintes de ressources et de temps de traitement. Des modèles parcimonieux sont nécessaires pour répondre aux contraintes de frugalité. Des données synthétiques permettront également d'étudier des situations difficilement observables en pratique (rareté, anomalies, complétion d'annotations coûteuses à obtenir), de simuler des scénarios d'aide à la décision (ex. maintenance prédictive) et de tester des modèles prédictifs en aval.

2. État de l'art et méthodologie

La conception de modèles de fondation multimodaux est un axe de recherche en plein essor. Au-delà des LLMs principalement dédiés au texte, l'état de l'art en IA générative multimodale comprend deux familles principales :

- Les modèles vision–langage (VLM) qui apprennent un espace commun texte–vision par (i) alignement contrastif entre encodeurs séparés, où image et texte sont encodés séparément puis rapprochés dans un espace commun via une perte contrastive, emblématisé par CLIP–ViT, par exemple en zéro-shot [1], (ii) une architecture de type Transformer traite texte et régions dans l'image conjointement, comme VisualBERT [8]. D'autres approches récentes associent étroitement la conception des modèles et la préparation des données. Ainsi, BLIP [6] s'auto-améliore grâce à un mécanisme de bootstrapping de légendes, générant de nouvelles paires image–texte de meilleure qualité. LLaVA [5], de son côté, relie de façon modulaire un encodeur visuel à un LLM via un projecteur, puis affine l'ensemble par instruction-tuning. CogVLM [13] pousse encore plus loin l'alignement texte–vision en intégrant directement un « expert visuel » dans les couches du LLM. Enfin, Molmo/PixMo [9] souligne que la qualité et la traçabilité des données ouvertes jouent un rôle aussi déterminant que l'architecture elle-même pour garantir une performance améliorée.
- Pour les données texte et séries temporelles, on peut citer trois approches. La première convertit la série en texte afin d'exploiter un LLM sans réentraînement, comme LLMTIME ou PromptCast, privilégiant la frugalité et le zéro/few-shot au détriment parfois de la précision [11, 4]. La deuxième aligne directement les séries avec le LLM via des modules légers, à l'image de GPT4TS/FPT [12], TIME-LLM [10], Time-LlaMA [7] ou TEMPO [3]. Enfin, la troisième approche vise le raisonnement sur des séries temporelles en contexte conversationnel de question–réponse. Time-MQA s'appuie sur un vaste corpus multitâches enrichi de contexte textuel [14], tandis que ChatTime introduit des tokens numériques spécialisés [2], et que ChatTS adopte une approche inspirée des VLM avec des données synthétiques ouvertes pour produire des réponses explicatives allant au-delà de la prévision [15].

Les architectures MoE—Mélanges d'experts (Mixture of Experts – MoE) (Shazeer et al., 2017) ont montré d'excellentes performances pour concilier capacité et frugalité ; par exemple, des variantes récentes de LLM à experts (ex. Mixtral 8×7B [18], DeepSeek-MoE [19]) réduisent le coût d'inférence en n'activant qu'un sous-ensemble d'experts par jeton.

Des approches récentes de MoE parcimonieux montrent que l'activation sélective des experts, via des mécanismes de compétition (*CompeteSMoE* [16]) ou de routage dynamique optimisé (*HyperRouter* [17]), permet d'accroître la capacité des modèles tout en limitant les coûts. Elles offrent ainsi un moyen efficace et frugal pour spécialiser et exploiter des modèles de fondation multimodaux en contexte industriel.

Dans le cadre de cette thèse, une première direction consistera à explorer l'approche des Mélanges d'experts pour traiter l'hétérogénéité des données. Chaque modalité pourra être prise en charge par un expert spécialisé (texte, séries temporelles, données tabulaires, imagerie/diagrammes). Un mécanisme de *gating* permet de combiner dynamiquement leurs contributions selon la tâche (prédiction, génération). Cette approche permet de favoriser la frugalité computationnelle, la modularité et l'interprétabilité. Le cadre des MoE s'articule naturellement avec les modèles de fondation multimodaux/GPAI, en permettant de combiner plusieurs modèles spécialisés de manière efficace.

Un enjeu réside dans la gestion du phénomène d'effondrement des représentations (*representation collapse*), particulièrement critique en situation multimodale : il correspond au cas où tous les experts (ou composantes du modèle)

apprennent des représentations trop similaires, ce qui réduit la diversité, la spécialisation et, in fine, la capacité effective du modèle. En contexte multimodal, cette problématique est amplifiée : chaque modalité doit être représentée de manière à la fois distincte et correctement intégrée.

Concernant l'adaptation de ces modèles, l'approche privilégiée sera celle de la combinaison d'experts spécialisés (MoE), orchestrés dynamiquement. Chaque modalité ou sous-tâche pourra ainsi être prise en charge par un expert adapté, activé de manière sélective, ce qui garantit modularité, frugalité computationnelle et meilleure interprétabilité. En complément, plusieurs scénarios seront étudiés : l'évaluation zero-shot / few-shot de modèles pré-entraînés existants sur des tâches spécialisées ; l'adaptation efficace en paramètres (PEFT) — comme LoRA, *prefix-tuning* [21]— qui constitue une alternative économe en données et en calcul ; et, de façon comparative, le fine-tuning complet, envisageable si les ressources et la qualité des données le permettent.

Cette démarche permettra de tirer parti de la puissance des modèles de fondation tout en garantissant leur adaptation aux spécificités des données industrielles (volume limité, multimodalité, rareté, hétérogénéité), sous des garanties contre l'effondrement de représentations, afin de préserver et d'exploiter pleinement la contribution de chaque modalité.

Cas d'usage : Les travaux de thèse seront illustrés et validés sur un ou plusieurs cas d'application issus du monde industriel dans le cadre du projet IAG1:

- génération conditionnelle de séries temporelles synthétiques pour l'augmentation de données (ex. anomalies rares, enrichissement et annotation de bases de données métier),
- génération de rapports de diagnostic à partir de détections sur séries temporelles et texte métier,
- génération de scénarios de maintenance exploitant des retours experts,
- assistance au résumé et à la vérification de dessins techniques

3. Références

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

[2] Chengsen Wang, Qingquan Song, Qi Zhang, et al. (2024). ChatTime: A Unified Multimodal Time Series Foundation Model Bridging Numerical and Textual Data. arXiv:2412.11376.

[3] Defu Cao, Fang Jin, Jie Feng, et al. (2023). TEMPO: Prompt-based Generative Pre-trained Transformer for Time Series Forecasting. arXiv:2310.04948.

[4] Hao Xue, Fuli Feng, et al. (2023). PromptCast: A New Prompt-based Learning Paradigm for Time Series Forecasting. arXiv:2210.08964.

[5] Haotian Liu, Chunyuan Li, et al. (2023). Visual Instruction Tuning. arXiv:2304.08485.

[6] Junnan Li, Dongxu Li, et al. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086.

[7] Juyuan Zhang, Weizhong Zhang, et al. (2025). Adapting Large Language Models for Time Series Modeling via a Novel Parameter-efficient Adaptation Method. arXiv:2502.13725.

[8] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, Kai-Wei Chang. (2019). VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv:1908.03557.

[9] Matt Deitke, Chunyuan Chen, et al. (2024). Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. arXiv:2409.17146.

[10] Ming Jin, Shun Wang, Yifan Feng, et al. (2023). Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. arXiv:2310.01728.

[11] Nate Gruver, Matthew Finzi, et al. (2023). Large Language Models Are Zero-Shot Time Series Forecasters. arXiv:2310.07820.

[12] Tian Zhou, Pengzhan Niu, et al. (2023). One Fits All: Power General Time Series Analysis by Pretrained LM. arXiv:2302.11939.

[13] Weihan Wang, Qiang Liu, et al. (2023). CogVLM: Visual Expert for Pretrained Language Models. arXiv:2311.03079.

[14] Yaxuan Kong, Yiming Yang, et al. (2025). Time-MQA: Time Series Multi-Task Question Answering with Context Enhancement. arXiv:2503.01875.

[15] Zhe Xie, Zihan Li, et al. (2024). ChatTS: Aligning Time Series with LLMs via Synthetic Data for Enhanced Understanding and Reasoning. arXiv:2412.03104.

[16] Yihan Dong, Yuqi Lin, Yihe Deng, et al. (2024). CompeteSMoE: Training Large Language Models with Competition Routing. arXiv:2402.02526.

[17] Giang Do, Khiem Le, Quang Pham, TrungTin Nguyen, Thanh-Nam Doan, Binh T. Nguyen, Chenghao Liu, Savitha Ramasamy, Xiaoli Li, Steven C.H. Hoi (2023). *HyperRouter: Towards Efficient Training and Inference of Sparse Mixture of Experts via HyperNetwork.* arXiv:2312.07035.
[18] Albert Q. Jiang, et al. (2024). Mixtral of Experts: Mixtral 8×7B, a Sparse Mixture of Experts Model. arXiv:2401.04088.
[19] Damai Dai, Chengqi Deng, Chenggang Zhao, et al. (2024). DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL.
[20] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. (2021). On the Opportunities and Risks of Foundation Models. arXiv:2108.07258.
[21] Li, X., Liang, P. (2021). *Prefix-Tuning: Optimizing Continuous Prompts for Generation.* arXiv:2101.00190.

PROFIL RECHERCHE

Le(la) candidat(te) doit justifier d'un Master Recherche (ou formation équivalente avec un intérêt avéré pour la recherche) dans le domaine des sciences des données et de l'Intelligence Artificielle.

Compétences attendues:

- Master Recherche ou équivalent en sciences des données et Intelligence Artificielle.
- Intérêt marqué pour la recherche et goût pour les applications.
- Solides compétences en inférence statistique et en optimisation.
- Maîtrise de l'apprentissage profond.
- Programmation en Python, avec expérience PyTorch/TensorFlow.
- Des compétences sur les modèles d'IA générative serait un plus.

CONTACT ET COMMENT POSTULER

Pour postuler, merci d'envoyer les éléments suivants au format PDF à : faicel.chamroukhi@irt-systemx.fr

1. CV détaillé
2. Lettre de motivation
3. Relevés de notes des deux dernières années d'étude de Master ou de cycle ingénieur
4. Au moins une lettre de recommandation