RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

ONERA
THE FRENCH AEROSPACE LAB

www.onera.fr

PROPOSITION DE STAGE EN COURS D'ETUDES

Référence : **DTIS-2026-31**
(à rappeler dans toute correspondance)

Lieu : Palaiseau

Département/Dir./Serv. : DTIS

Tél. : 01 80 38 65 69 / 01 80 38 65 91

Responsable(s) du stage : S. Herbin, A Chan-Hon-Tong

Email. : stephane.herbin@onera.fr, adrien.chan-hon-tong@onera.fr

DESCRIPTION DU STAGE

Thématique(s) : IAD, RA

Type de stage : ☒ Fin d'études bac+5 ☒ Master 2 ☐ Bac+2 à bac+4 ☐ Autres

Title : Reliability of AI foundation models for robotics

Foundation models (FMs) are large deep neural networks (DNNs) that have been trained on extensive datasets. Examples include popular Large Language Models, such as ChatGPT; Vision Language Models, such as CLIP; and Vision Language Action Models, such as RT-2 or PI-Zero. FMs hold immense promise due to their versatility; they serve as knowledge sources and mappings between diverse data representation spaces.

However, as is common with DNNs, FMs can exhibit unexpected behaviors such as hallucinations, overconfidence, sensitivity to adversarial attacks, and domain shift. These behaviors pose critical safety concerns in physical systems such as robotics, as they can lead to potential physical harm or material damage. This is especially problematic because interacting with humans requires trust, which is based on reliability. Therefore, mastering the reliability of FMs for robotics is a key objective.

Since critical for the integration of AI models in robotics, few recent studies have addressed reliability issues in robotics for embodied LLM [Liu, 2024] [QI, 2024] [Wu, 2024] [Zhang, 2024] [Zhang, 2024], Vision-language-Action models [Wang, 2024] [Mao, 2024] human-robot language based interaction [Azeem, 2024], navigation [Mullen, 2024] or planning [Zhu, 2024].

After a bibliography phase, the work will mainly consist of adapting a selected article on the state of the art, coding the method, evaluating it, and propose improvements.

- Liu, A., Zhou, Y., Liu, X., Zhang, T., Liang, S., Wang, J., Pu, Y., Li, T., Zhang, J., Zhou, W., Guo, Q., Tao, D., 2024. Compromising Embodied Agents with Contextual Backdoor Attacks. https://doi.org/10.48550/arXiv.2408.02882
- Qi, Y., Kyebambo, G., Xie, S., Shen, W., Wang, S., Xie, B., He, B., Wang, Z., Jiang, S., 2024. Safety Control of Service Robots with LLMs and Embodied Knowledge Graphs. https://doi.org/10.48550/arXiv.2405.17846
- Wu, X., Chakraborty, S., Xian, R., Liang, J., Guan, T., Liu, F., Sadler, B.M., Manocha, D., Bedi, A.S., 2024. Highlighting the Safety Concerns of Deploying LLMs/VLMs in Robotics. https://doi.org/10.48550/arXiv.2402.10340
- Zhang, W., Kong, X., Braunl, T., Hong, J.B., 2024a. SafeEmbodAI: a Safety Framework for Mobile Robots in Embodied AI Systems. https://doi.org/10.48550/arXiv.2409.01630
- Zhang, W., Kong, X., Dewitt, C., Braunl, T., Hong, J.B., 2024b. A Study on Prompt Injection Attack Against LLM-Integrated Mobile Robotic Systems, in: 2024 IEEE 35th International Symposium on Software Reliability Engineering Workshops (ISSREW).), pp. 361–368. https://doi.org/10.1109/ISSREW63542.2024.00103
- Wang, T., Liu, D., Liang, J.C., Yang, W., Wang, Q., Han, C., Luo, J., Tang, R., 2024. Exploring the Adversarial Vulnerabilities of Vision-Language-Action Models in Robotics. https://doi.org/10.48550/arXiv.2411.13587
- Azeem, R., Hundt, A., Mansouri, M., Brandão, M., 2024. LLM-Driven Robots Risk Enacting Discrimination, Violence, and Unlawful Actions. https://doi.org/10.48550/arXiv.2406.08824
- Mullen, Jr, J.F., Goyal, P., Piramuthu, R., Johnston, M., Manocha, D., Ghanadan, R., 2024. "Don't forget to put the milk back!" Dataset for Enabling Embodied Agents to Detect Anomalous Situations. IEEE Robot. Autom. Lett. 9, 9087–9094. https://doi.org/10.1109/LRA.2024.3430129
- Zhu, Z., Wu, Bingzhe, Zhang, Z., Han, L., Liu, Q., Wu, Baoyuan, 2024. EARBench: Towards Evaluating Physical Risk Awareness for Task Planning of Foundation Model-based Embodied AI Agents. https://doi.org/10.48550/arXiv.2408.04449

- Mao, Z., Dai, S., Geng, Y., Ruchkin, I., 2024. Zero-shot Safety Prediction for Autonomous Robots with Foundation World Models. Presented at the ICRA 2024 Workshop Back to the Future: Robot Learning Going Probabilistic. https://arxiv.org/abs/2404.00462

Est-il possible d'envisager un travail en binôme ? Non

Méthodes à mettre en oeuvre :

☐ Recherche théorique	☒ Travail de synthèse
☒ Recherche appliquée	☐ Travail de documentation
☐ Recherche expérimentale	☐ Participation à une réalisation

Possibilité de prolongation en thèse : Oui

Durée du stage : Minimum : 4 mois Maximum : 6 mois

Période souhaitée : Mars-Août 2026

PROFIL DU STAGIAIRE

Connaissances en Intelligence artificielle et Robotique. Bonne pratique de la programmation dans les environnements de développement IA (par ex. pytorch).	Ecoles ou établissements souhaités : M2 ou Ecole Ingénieur, spécialité en Intelligence Artificielle, Vision par ordinateur ou robotique.