Fiche-projet (Modèle ARED 2026)

Demande n° 4319 - ARED
Acronyme : SIAIAS

Identification du projet	
N°	4319
Année	2026
Statut	
Date de création	
Date de modification	
Date de soumission	
Acronyme	SIAIAS
Intitulé du projet	Security of interacting AI agent systems
Dispositif	ARED

Priorité régionale	
Priorité régionale	Cybersécurité

Unité de recherche	
Nom de l'unité	IRISA
Civilité	Monsieur

Nom du responsable	Barais
Prénom du responsable	Olivier

Stratégie régionale de recherche et d'innovation (S3)	
Domaine d'innovation stratégique principal	Economie Numérique Sécurisée et Responsable
Levier thématique associé au DIS principal	Cybersécurité
Domaine d'innovation stratégique secondaire	
Levier thématique associé au DIS secondaire	
Axe transversal : transitions climatique et écologique	
Axe transversal : cohésion et inclusion sociale	

Organisme de tutelle	
Etablissement	Université de Rennes / IRISA
Commune	Rennes
Civilité	Madame
Nom	Pinchinat
Prénom	Sophie
Adresse électronique	sophie.pinchinat@irisa.fr
Tel	+33 (0)2 99 84 72 54

Tel	
Courriel	

Données complémentaires de l'unité de recherche

Libellé	**Obligatoire (oui/non)**	**Valeur**
Statut et N° de l'unité de recherche (UMR/USR/EA)	Non	UMR 6074
Le cas échéant, nom de l'équipe de recherche	Non	LogicA
Le cas échéant, nom du responsable de l'équipe de recherche	Non	Sophie Pinchinat
Activités de recherche et points forts de l'équipe	Oui	Les points forts de l'équipe LogicA sont la mixité des compétences fondamentales à la bonne réussite de ce

		projet : fondements logiques des systèmes multi-agents, expertise de design et déploiement de solutions logiciels à base de LLMs, et applications de techniques de traitement de données avec apprentissage machine et IA génératives.
Si implication dans un/plusieurs programmes européens ou internationaux, préciser	Non	

Co-directeur de thèse	
Civilité	Dr
Nom	Maarek
Prénom	Manuel
Unité de recherche	Computer Science Department
Etablissement de tutelle	Heriot-Watt University

Description du projet

Lieu principal de déroulement du projet en Bretagne : Rennes

Lieu principal de déroulement du projet si hors Bretagne :
Édimbourg, Écosse, Royaume-Uni

Libellé	**Obligatoire (Oui/Non)**	**Valeur**
Résumé synthétique du projet (2 000 caractères maximum)	Oui	Le paradigme des systèmes multi-agents englobe tous les types de systèmes, de la distribution classique de systèmes (composés d'éléments déterministes décrits par du code pur) aux systèmes de l'intelligence artificielle agentive, y compris ceux que nous appelons ici « les systèmes d'agents intelligents interagissants », qui combinent du code pur et des composants à

		base de LLMs (modèles de langage de grande taille). Même si l'analyse de sécurité des systèmes distribués classiques est un domaine bien établi doté d'un large éventail d'outils tels que l'analyse de tainting, l'analyse de sécurité des systèmes d'agents intelligents interagissants est balbutiante, pour les raisons suivantes. D'une part, la présence d'éléments LLM rend difficile la traçabilité des données, car les données originales sont modifiées par les couches d'attention des transformateurs de manière extrêmement complexe. D'autre part, il a été observé que la composition d'éléments LLM validés séparément peut donner lieu à un comportement global indésirable. Le sujet de cette thèse est l'analyse de sécurité des systèmes d'agents intelligents interagissants. Typiquement, les attaques par prompt injection est un risque important pour les systèmes contenant un LLM ; des solutions ont été proposées pour atténuer ce danger,

		comme le fine-tuning séparant les données utilisateur et les instructions dans un prompt. Cependant, les risques se décuplent dès lors qu'un LLM interagit avec d'autres éléments du système global (par exemple, le fine-tuning d'un modèle risque de ne pas offrir une défense suffisante), impliquant la nécessité de mettre en place des protections au niveau supérieur.
Caractère structurant/stratégique du projet (descriptif/exemple : voir article 5 du Règlement ARED 2026 – 2 000 caractères maximum)	Oui	Les principaux problèmes de recherche que cette thèse vise à résoudre sont les suivants. 1) Quelle est l'efficacité des défenses existantes contre les attaques basées sur des prompts dans le cas d'un système comportant différentes formes de composition LLM (l'efficacité peut varier en fonction du type d'architecture) ? 2) Quels instrumentations individuelles de LLMs permette de corroborer les caractériques de sécurité des LLMs? 3) Comment les instrumentations des LLM peuvent-ils appuyer un raisonnement compositionnel sur la

		sécurité des systèmes d'agents intelligents interagissants ? 4) Quelles sont les instrumentsations des LLM peuvent servir à concevoir des protocoles d'interaction entre agents afin de garantir un comportement sécurisé pour les systèmes d'agents interagissants ? Il y aura quatre étapes principales dans cette thèse : 1) Conception et mise en œuvre de systèmes d'agents intelligents interagissants qui illustrent le contournement des solutions contre les attaques basées sur des prompts. 2) Proposition d'instrumentations individuels (par exemple, à l'aide de métriques dérivées des modèles) des LLMs permettant de caractériser la sécurité de l'inférence LLM. 3) Expérimentations pour sélectionner des instruments individuels qui démontrent des phénomènes d'interaction entre composantes révélant des défauts de sécurité et permettant le raisonnement formel.

		4) Conception de remédiation (par exemple, des protocoles d'interaction) contre les attaques basées sur des prompts dans les systèmes d'agents intelligents interagissants à l'aide des instruments sélectionnés précédemment.
Porteur du projet : date d'obtention de l'HDR ou dérogation	Oui	Professeure Sophie Pinchinat (HDR 2003)
Collaborations scientifiques (nature/pays) et partenariat socio-économique envisagé (3 000 caractères maximum)	Oui	Ce sujet de recherche est l'aboutissement du projet MLSEAN, de collaboration financé entre l'IRISA et Heriot-Watt University (Royaume-Uni) impliquant Sophie Pinchinat, Manuel Maarek, Michael Lones et Sasa Radomirovic. La thèse se déroulera au sein de l'équipe LogicA qui offre un environment de recherche pertinent doté de compétences théoriques en méthodes formelles pour les systèmes multi-agents, d'expérience dans la conception et le déploiement des systèmes à base de LLMs, ainsi que d'expertises appliquées dans l'IA et les modèle d'IA génératives (GenAI). Par ailleurs, la thèse repose sur la collaboration en cours

		entre l'université de Rennes et l'Université Heriot-Watt qui offre des compétences en synergie avec l'équipe LogiCA : apprentissage automatique et GenAI avec Michael Lones, vérification de protocoles de sécurité avec Sasa Radomirovic, méthodes formelles appliquées pour la sécurité des systèmes logiciels avec Manuel Maarek (co-encadrant). Afin d'atteindre ces quatre étapes, nous proposons une méthodologie basée sur les multiple compétences de l'environnement de recherche offert par l'équipe LogiCA et de la collaboration (cotutelle) avec l'Université Heriot-Watt (les deux équipes mainetnant un veille technologique sur le domaine des GenAI en pleine évolution) : 1) les implémentations des systèmes d'agents intelligents interagissants peuvent être hébergées dans un environnement déjà existant où les membres de l'équipe ont démontré leur compétence pour déployer des systèmes d'IA agentive; 2)

		l'instrumentation de LLM est assistée par la connaissance de l'équipe sur l'entraînement des LLM et les outils existants utilisant des métriques dérivées des modèles; 3) le raisonnement compositionnel est appuyé par l'expertise d'équipe en vérification formelle ainsi qu'en protocoles de sécurité.
Si Domaine d'innovation stratégique (DIS) identifié, préciser l'adéquation du projet avec le DIS (2 000 caractères maximum)	Oui	L'entrée accélérée des LLMs dans tous les domaines d'activité où l'informatique fournit des aides pose de véritables questions sur les enjeux de cybersécurité. Ce projet de thèse cible ces enjeux actuels mais aussi les enjeux à venir en considérant le contexte multi-LLM.
Si priorité régionale cochée, préciser en quoi le projet s'y rattache effectivement (200 caractères maximum)	Non	Economie Numérique Sécurisée et Responsable, levier Cybersécurité
Si priorité régionale « Politique publique Région » cochée, préciser (voir liste au point 4, Règlement ARED 2026)	Non	
Cotutelle financière internationale (oui/non)	Non	Oui
Nom(s) du(des) cofinanceur(s) envisagé(s) ou confirmé(s)	Non	Bourse de thèse Heriot-Watt University (départmentale or DLA EPSRC)
Si cofinancement industriel, justifier le choix du dispositif ARED au lieu du dispositif CIFRE	Non	

Etat du cofinancement (acquis/demandé)	Non	Demandé
Si demandé, date prévisionnelle de réponse du cofinanceur	Non	Mars 2026, mais conditionnée au cofinancement ARED
Si cofinancement refusé, autre(s) source(s) de cofinancement identifiée(s)	Non	Non
Si reversement prévu, préciser (partenaire, montant)	Non	

Candidat (futur·e doctorant·e)	
Genre	H
Nationalité	France

Collaboration internationale (dont cotutelle financière internationale)	
Pays	Écosse, Royaume-Uni
Etablissement partenaire à l'étranger	Heriot-Watt Université

Pièces jointes à déposer pour le dépôt du projet	
Type de pièce	**Commentaire**
Dépôt – CV du ·de la porteur·se	Obligatoire
Dépôt – Avis directeur.rice unité de recherche	Obligatoire
Dépôt – Justificatif cofinancement	Facultatif
Dépôt – Lettre partenaire étranger	Facultatif
Dépôt – Pièce complémentaire	Facultatif