

Termes de référence (TDR) : Expert en optimisation de modèles et inférence sur systèmes contraints

Projet RHIS / Projet AMIS-RHS (Appui aux Ressources Humaines en Santé)

1. Informations générales

Domaine	Détails
Intitulé de la mission	Mission d'expertise technique pour la conception, le développement et le déploiement d'un assistant numérique de formation des personnels de santé fonctionnant hors connexion (intelligence artificielle embarquée, traitement automatique du langage multilingue français-wolof, recherche documentaire augmentée)
Maîtrise d'ouvrage et autorité contractante	Université Gaston Berger (UGB), en partenariat avec les équipes conjointes de l'Université Cheikh Anta Diop et de l'École supérieure polytechnique (UCAD/ESP)
Bénéficiaires finaux	Personnels de santé des quinze structures du district sanitaire de Keur Momar Sarr
Zone d'intervention	Sénégal (Saint-Louis, district sanitaire de Keur Momar Sarr) et à distance
Projet	Projet AMIS-RHS (Appui aux Ressources Humaines en Santé), mis en œuvre avec l'appui d'Expertise France sous le projet RHIS.
Durée totale	10 mois (mission perlée), de novembre 2026 à août 2027 (dates indicatives)
Nombre d'experts	Un (01)
Charge totale indicative	75 jours maximum

2. Contexte et justification du besoin

Le projet AMIS-RHS, lauréat de l'appel à projet subventionné par le projet RHIS Ressources Humaines Innovantes en Santé, financé par le groupe AFD, vise le renforcement des ressources humaines en santé au Sénégal. Dans le district sanitaire de Keur Momar Sarr, quinze structures de santé fonctionnent avec des personnels peu nombreux, éloignés des centres de formation et disposant d'un accès limité aux référentiels nationaux et à la formation continue.

Pour répondre à cette situation, l'UGB et l'UCAD/ESP conçoivent conjointement un assistant numérique de formation destiné aux personnels de santé du district. L'outil s'appuiera exclusivement sur les documents officiels d'enseignement et les référentiels nationaux validés (programmes de formation, protocoles du ministère de la Santé et de l'Action sociale, ordinogrammes de prise en charge). Il permettra aux personnels d'interroger ces référentiels en français et en wolof, y compris à l'oral, de suivre des parcours d'autoformation et de s'entraîner sur des cas d'école, sans connexion à Internet.

L'interaction s'effectue par entrée vocale ou écrite ; la restitution est écrite et visuelle à l'écran, la synthèse vocale ne relevant pas du périmètre de la mission.

La conception de cet outil se heurte à trois difficultés, qui structurent le périmètre de la mission.

La première tient aux ressources linguistiques. Le wolof est une langue peu dotée en corpus numériques et il n'existe pas, à ce jour, de corpus de parole transcrite couvrant le vocabulaire de la santé dans des conditions d'usage réelles, où le wolof et le français alternent au sein d'un même

échange. Constituer un corpus vocal transcrit et validé, numériser et structurer la base documentaire de référence : ces travaux sont des préalables du projet et sont intégrés à la mission, avec les charges correspondantes.

La fiabilité des contenus constitue une deuxième difficulté. Un outil de formation en santé ne peut restituer que des contenus vérifiables. L'assistant reposera sur un moteur pédagogique associant, d'une part, la recherche documentaire augmentée (RAG), dont chaque réponse cite sa source dans la base de référence et qui s'abstient explicitement lorsque la question sort de la couverture documentaire, et, d'autre part, la génération de contenus d'entraînement (questionnaires, cas d'école) soumis à validation académique. La fiabilité sera mesurée sur des jeux de test construits à cet effet et validés par les référents médicaux du projet.

Reste la contrainte matérielle et énergétique. L'outil doit fonctionner intégralement hors connexion, dans des structures aux ressources limitées. Plusieurs architectures d'exécution sont envisageables : exécution embarquée sur des terminaux durcis, serveur local installé dans chaque structure et accessible en réseau local, ou configuration hybride répartissant les traitements entre les deux. Plutôt que d'arrêter ce choix sur dossier, la mission prévoit une expérimentation comparative en conditions réelles dans le district ; la décision d'équipement définitive sera prise sur la base des résultats mesurés (latence, autonomie, robustesse, coût d'acquisition et charge de maintenance, énergie comprise).

La mission nécessite l'intervention complémentaire de trois experts en intelligence artificielle :

- Expert 1, optimisation de modèles et inférence sur systèmes contraints : architecture d'exécution, compression des modèles, inférence locale.
- Expert 2, traitement automatique du langage multilingue et ingénierie de corpus : constitution du corpus vocal, reconnaissance vocale, traitement du wolof et du français.
- Expert 3, recherche documentaire augmentée, évaluation et développement d'applications : moteur pédagogique, jeux d'évaluation, application destinée aux praticiens, coordination des livrables.

3. Objectifs et résultats attendus

Objectif général :

Appuyer l'UGB et l'UCAD/ESP dans la conception, le développement, l'évaluation et le déploiement d'un assistant clinique hors connexion pour les personnels de santé des quinze structures du district de Keur Momar Sarr.

Objectifs spécifiques :

- Constituer un corpus de parole wolof et français transcrit et validé, représentatif des usages des personnels de santé, ainsi qu'une base documentaire de référence numérisée, structurée et validée à partir des documents officiels d'enseignement et des protocoles nationaux.
- Développer un module de reconnaissance vocale wolof-français atteignant un taux d'erreur sur les mots (WER) inférieur ou égal à 30 % sur un jeu de test indépendant, sous réserve de la décision du jalon 1 décrit en section 4, ainsi qu'un moteur pédagogique dont toute réponse est sourcée et qui s'abstient hors couverture documentaire.

- Concevoir l'architecture d'exécution hors connexion, l'arbitrer par une expérimentation comparative en conditions réelles (terminaux, serveur local, configuration hybride) et optimiser les modèles pour les configurations candidates puis pour la configuration retenue.
- Développer l'interface destinée aux praticiens, déployer l'assistant dans les quinze structures, former les utilisateurs et transférer la maintenance et la procédure de mise à jour hors connexion aux équipes de l'UGB et de l'UCAD/ESP.
- Mesurer l'effet de l'assistant sur les connaissances des personnels : mesure de référence avant la mise en service expérimentale, mesure finale en fin de mission, suivi de l'utilisation effective.

Résultats attendus :

À l'issue de la mission, les quinze structures du district disposent d'un assistant de formation opérationnel, les équipes de l'UGB et de l'UCAD/ESP sont autonomes pour la maintenance et la mise à jour, et le projet a produit trois ressources réutilisables : le corpus vocal transcrit, la base documentaire structurée et les jeux d'évaluation.

4. Description de la mission

La mission est découpée en sept étapes. Les charges par expert, exprimées en jours, sont indicatives ; elles totalisent 215 jours sur huit mois, en mission perlée. Les échéances sont exprimées en semaines à compter de T0 ; les dates correspondantes, calculées pour un début de mission au 1er décembre 2026, figurent au calendrier de la section 7.

Étape	Activités principales par expert	Répartition de la charge	Livrables et échéances
Étape 1 : Cadrage, conformité et protocole d'expérimentation	Audit des contraintes matérielles, énergétiques et de connectivité des quinze structures ; conception du protocole d'expérimentation comparative, avec rotation des configurations entre structures pilotes afin de dissocier l'effet site de l'effet architecture ; préparation d'options d'achat conditionnelles (devis et délais garantis) pour chaque architecture candidate ; choix de la localisation des moyens de calcul d'entraînement.	1 j.	Note de cadrage technique et méthodologique, audit énergétique, protocole d'expérimentation comparative, dossier CDP prêt au dépôt, remis à l'UGB (T0+6).
Étape 2 : Constitution des corpus vocal et documentaire	Appui à la spécification du matériel de collecte audio.	1 j.	Convention de transcription et guide d'annotation (T0+7). Premier lot de 10 heures validées (T0+16). Base documentaire structurée et validée (T0+18). Corpus complet de 30 heures validées et rapport de qualité incluant l'accord inter-annotateurs (T0+23).
Étape 3 : Développement du module de reconnaissance	Préparation des variantes de modèles candidates à l'inférence locale.	5 j.	Mesure intermédiaire du WER sur le premier lot de corpus (T0+18), base d'instruction du jalon 1.

vocale et du moteur pédagogique			Module de reconnaissance vocale et moteur pédagogique documentés (T0+26).
Étape 4 : Évaluation et sécurité des contenus	Revue de la testabilité des cibles sur les configurations matérielles candidates.	2 j.	Jeux d'évaluation et de sécurité validés, typologie des erreurs critiques comprise (T0+21). Rapport d'évaluation complet (T0+27).
Étape 5 : Optimisation des modèles et inférence locale	Quantification (4 et 8 bits) des modèles pour les configurations candidates ; distillation et élagage si le compromis entre qualité et latence l'exige ; intégration sur les moteurs d'inférence locaux des deux configurations candidates prioritaires (terminaux, serveur local) ; bancs d'essai de latence, de mémoire et de consommation énergétique sur matériel réel.	25 j.	Kits d'inférence candidats pour l'expérimentation comparative (T0+24). Rapport de bancs d'essai consolidé (T0+27).
Étape 6 : Interface praticiens et intégration	Intégration des kits d'inférence dans l'application ; signature cryptographique et vérification d'intégrité des paquets de mise à jour hors connexion.	7 j.	Version expérimentale de l'application, intégrant les versions de travail des modules (T0+24). Version finale intégrant les retours de l'expérimentation (T0+30).
Étape 7 : Expérimentation comparative, déploiement, mesure d'effet et transfert	Conduite de l'expérimentation comparative des architectures sur trois structures pilotes, de T0+24 à T0+28, avec mesures de latence, d'autonomie, de robustesse, de coût d'acquisition et de charge de maintenance, énergie comprise, et rotation des configurations entre structures ; déploiement sur les quinze structures ; test de la procédure de mise à jour hors connexion ; transfert technique aux équipes de l'UGB et de l'UCAD/ESP.	24 j.	Rapport d'expérimentation comparative et décision d'architecture (T0+28). Assistant déployé dans les quinze structures et personnels formés (T0+33). Procès-verbal de recette, rapport de mesure d'effet, guide de transfert et documentation (T0+34).
Étape à confirmer	Visite sur site	5j	

Deux jalons de décision sont inscrits au calendrier et arbitrés en comité de suivi.

Le premier jalon intervient à T0+19 semaines. Il s'appuie sur la mesure intermédiaire du WER réalisée sur le premier lot de corpus (T0+18), appréciée au regard des heures effectivement disponibles à cette date, et sur la courbe de progression projetée. Si ces éléments ne permettent pas de projeter l'atteinte du seuil de 30 %, le comité décide la bascule vers une interaction vocale à vocabulaire contraint, complétée par l'interaction écrite ; la charge correspondante est redéployée au sein des étapes 3 et 6, sans remise en cause du reste de la mission.

Le second, fixé à T0+28 semaines, conditionne le déploiement généralisé : celui-ci n'est autorisé qu'en l'absence d'erreur critique sur le jeu de sécurité, au sens de la typologie arrêtée à l'étape 4. À défaut,

l'assistant est déployé en mode consultation documentaire seule, jusqu'à correction et nouvelle validation.

5. Coordination et modalités d'exécution

Rattachement institutionnel : l'équipe d'experts est rattachée directement à Expertise France pour l'Université Gaston Berger (UGB, Saint-Louis).

Pilotage : Expertise France et l'expert recherche documentaire augmentée, évaluation et développement d'applications assurent la coordination d'ensemble des livrables et la liaison avec les enseignants-chercheurs de l'UGB et de l'UCAD/ESP. Le comité de suivi, présidé par l'UGB, réunit mensuellement l'UGB, l'UCAD/ESP, EF et les experts.

Dispositif académique : l'UGB mobilise sur ses ressources propres les profils académiques nécessaires à la mission, qui ne relèvent pas du présent recrutement : un médecin référent chargé de la validation de la base documentaire et des jeux d'évaluation, un second lecteur médical désigné parmi les enseignants associés pour la double lecture des jeux d'évaluation, un expert linguistique du wolof (convention de transcription, arbitrages de normalisation, contrôle qualité), une équipe d'annotateurs pour la transcription du corpus vocal (annotation, dimensionnement vérifié au cadrage sur la base de la convention de transcription), un opérateur de collecte chargé de la conduite des sessions d'enregistrement, ainsi que les enseignants associés à la validation des contenus d'entraînement.

Lieu de travail : travail à distance, combiné à des regroupements techniques à l'UGB (Saint-Louis) et à des sessions de collecte, d'expérimentation et de déploiement dans le district de Keur Momar Sarr. La langue de travail de la mission et de rédaction de l'ensemble des livrables est le français.

Équipements : l'acquisition des équipements est prise en charge par le projet et ne relève pas de la mission des experts. Elle couvre les terminaux ou serveurs locaux, le matériel de collecte audio, les dispositifs d'alimentation électrique des structures (kits solaires, batteries, onduleurs) et les moyens de calcul nécessaires à l'adaptation des modèles (heures de calcul ou station de travail dédiée à l'UGB). La configuration initialement envisagée, quinze terminaux durcis, pourra être ajustée à l'issue de l'expérimentation comparative, dans la limite du budget d'équipement du projet.

6. Profils et compétences requis

Master 2, diplôme d'ingénieur ou doctorat en informatique, traitement automatique du langage, science des données ou domaine proche, ou équivalent de niveau bac+5 ; pour les seuls candidats sans diplôme de ce niveau, une expérience professionnelle probante d'au moins huit ans peut s'y substituer. Français courant indispensable. La sélection accorde une place déterminante aux réalisations vérifiables (systèmes en production, dépôts de code, publications, démonstrations).

Profil : Expert en optimisation de modèles et inférence sur systèmes contraints

- **Expérience :** au moins cinq ans en data science ou en ingénierie de systèmes d'intelligence artificielle, dont deux ans au minimum sur le déploiement de modèles en environnement contraint ou hors connexion.
- **Compétences clés :** compression de modèles (quantification 4 et 8 bits, distillation de connaissances, élagage) ; moteurs d'inférence locaux (llama.cpp, ONNX Runtime, ExecuTorch ou

équivalents) ; profilage de la latence, de la mémoire et de la consommation énergétique ; conception d'architectures à modes dégradés pour le fonctionnement hors connexion.

- Exigence : au moins une réalisation vérifiable de mise en production d'un modèle de langage ou de reconnaissance vocale sur matériel contraint.
- Atouts : déploiements en zone à faible connectivité ou à alimentation électrique instable ; expérience du secteur de la santé.

7. Calendrier indicatif

Jalon ou livrable	Échéance	Date indicative
Démarrage de la mission	T0	1er décembre 2026
Note de cadrage, audit énergétique, protocole d'expérimentation, dossier CDP remis à l'UGB pour dépôt	T0+6	12 janvier 2027
Convention de transcription et guide d'annotation	T0+7	19 janvier 2027
Autorisation ou récépissé CDP (jalón externe, conditionne le début de la collecte)	attendu T0+12	23 février 2027
Premier lot de corpus : 10 heures validées	T0+16	23 mars 2027
Mesure intermédiaire du WER sur le premier lot ; base documentaire structurée et validée	T0+18	6 avril 2027
Jalón 1 : revue des résultats intermédiaires du module de reconnaissance vocale	T0+19	13 avril 2027
Jeux d'évaluation et de sécurité validés, typologie des erreurs critiques comprise	T0+21	27 avril 2027
Corpus complet : 30 heures validées et rapport de qualité	T0+23	11 mai 2027
Version expérimentale de l'application et kits d'inférence candidats ; début de l'expérimentation comparative	T0+24	18 mai 2027
Module de reconnaissance vocale et moteur pédagogique documentés	T0+26	1er juin 2027
Rapport de bancs d'essai consolidé ; rapport d'évaluation complet	T0+27	8 juin 2027
Jalón 2 : rapport d'expérimentation comparative, décision d'architecture, autorisation de déploiement	T0+28	15 juin 2027
Version finale de l'application	T0+30	29 juin 2027

Déploiement dans les quinze structures et formation des personnels	T0+33	20 juillet 2027
Recette, rapport de mesure d'effet, guide de transfert	T0+34	27 juillet 2027
Clôture administrative de la mission		août 2027

En cas de décalage de la date effective de démarrage, le calendrier de la mission est replanifié par le comité de suivi, l'ensemble des livrables devant être remis au plus tard fin novembre 2027, la fin de la période de financement, en décembre 2027, constituant une butée non négociable.

8. Suivi-évaluation et indicateurs de performance

Livrable	Indicateurs	Sources de vérification
Corpus vocal transcrit	Au moins 30 heures validées en double passe, livrées par lots (10 heures, puis 30 heures cumulées) ; taux d'accord inter-annotateurs mesuré sur le sous-ensemble de recouvrement et documenté ; jeu de test indépendant d'au moins 2 heures de parole spontanée avec alternance codique, enregistré dans les conditions acoustiques des structures, locuteurs distincts du corpus d'entraînement.	Rapport de qualité du corpus, dépôt versionné.
Base documentaire structurée	Couverture des documents de référence identifiés au cadrage ; logique conditionnelle des ordinogrammes préservée ; validation formelle par les référents médicaux.	Registre de validation, base documentaire livrée.
Module de reconnaissance vocale	WER inférieur ou égal à 30 % sur le jeu de test indépendant, dont la composition (répartition wolof-français, part d'alternance codique) est documentée ; taux d'erreur sur les termes techniques du domaine mesuré séparément, avec un seuil fixé dans le protocole d'évaluation. En cas d'activation du repli décidé au jalon 1 : taux de reconnaissance supérieur ou égal à 90 % sur le vocabulaire contraint validé par le comité de suivi.	Rapport d'évaluation, banc d'évaluation automatisé.
Moteur pédagogique	100 % des réponses restituées avec citation de leur source ; taux de réponses correctes et fidèles aux sources supérieur ou égal à 90 % sur le jeu de référence, apprécié en double lecture ; taux d'abstention correcte supérieur ou égal à 90 % sur les questions hors couverture et taux de faux refus inférieur ou égal à 10 % sur les questions couvertes ; aucune erreur critique sur le jeu de sécurité, condition du déploiement généralisé (jalon 2), le mode consultation documentaire seule s'appliquant à défaut ; 100 % des contenus d'entraînement publiés validés par les enseignants associés.	Rapport d'évaluation, jeux de test versionnés.
Kits d'inférence	Latence mesurée de la fin de l'énoncé de l'utilisateur au premier élément de réponse affiché,	Rapport de bancs d'essai (T0+27), rapport

	reconnaissance vocale comprise : les seuils contractuels par configuration sont ceux consignés au protocole d'évaluation validé en comité de suivi, les valeurs de 5 secondes pour la première réponse et de 30 secondes pour la réponse complète constituant les cibles de conception pour la configuration serveur local, ajustables selon le matériel effectivement retenu ; autonomie mesurée sur un profil d'usage de référence (10 requêtes vocales par heure pendant 8 heures). Mesures réalisées sur chaque configuration candidate, puis confirmées sur la configuration retenue.	d'expérimentation comparative (T0+28).
Déploiement et formation	Assistant opérationnel dans les quinze structures ; sessions de formation organisées dans les quinze structures et proposées à l'ensemble des personnels, avec un objectif de 45 personnels formés, la mobilisation des personnels relevant du district sanitaire et de l'UGB ; procédure de mise à jour hors connexion, avec vérification d'intégrité, testée au moins une fois en conditions réelles.	Procès-verbal de recette, feuilles de présence, rapport de test de mise à jour.
Effet sur la formation	Mesure de référence des connaissances sur les structures pilotes avant la mise en service expérimentale, puis mesure finale en fin de mission ; progression des scores documentée ; taux d'utilisation hebdomadaire mesuré par la journalisation locale.	Rapport de mesure d'effet, journaux d'utilisation anonymisés.

9. Modalités de candidature

Le dossier de candidature comprend, pour chaque expert : un curriculum vitae détaillé, une lettre de motivation et la description d'au moins trois réalisations vérifiables en lien avec la mission (systèmes en production, publications, démonstrations).

Les candidatures peuvent être individuelles ou portées par une structure regroupant plusieurs experts. Une candidature portée par une structure comprend une note méthodologique commune de dix pages au maximum.