Sujet de stage de M2 IA

Développement et évaluation d'un agent conversationnel adaptatif dédié au soutien psychologique

en lien avec la société Mandarine CODI

Public visé : étudiants en Master 2 IA

Durée : 5 ou 6 mois (dates proposées : 01/02/2026 - 31/07/2026)

Encadrantes : Nadia Yacoubi Ayadi, LIRIS - UCBL (nadia.yacoubi-ayadi@univ-lyon1.fr)
Stéphanie Jean-Daubias, LIRIS-UCBL
Virginie Corvellec, Mandarine CODI

Mots-clés : Intelligence artificielle, Santé mentale, Modèles linguistiques à grande échelle

Rémunération : Environ 640 € / mois (environ 3 200 € pour 5 mois)

Contexte

Les troubles de santé mentale connaissent une progression préoccupante à l'échelle mondiale, accentuée par la pénurie de cliniciens et par les barrières socio-économiques limitant l'accès aux soins. Cette situation impose la recherche de solutions innovantes pour renforcer l'accompagnement psychologique, notamment en amont d'une prise en charge clinique. Dans ce contexte, l'intelligence artificielle (IA) et plus particulièrement les modèles de langage de grande taille (LLM) apparaissent comme des outils prometteurs.

De récentes revues de la littérature montrent que les LLM peuvent soutenir l'expression émotionnelle des individus et proposer une orientation adaptée (Guo et al., 2024). Toutefois, leur utilisation en santé mentale soulève des risques importants, en particulier liés à la fiabilité, aux biais et à la transparence.

Application concrète

Les résultats de ce travail de recherche seront appliqués à Buddaia, un projet partenarial impliquant le laboratoire LIRIS et la société MandarineCODI. Ce projet a pour objectif de développer une application mobile favorisant une meilleure santé mentale des salariés en difficulté psychologique, dans une démarche préventive qui ne s'inscrit pas dans le domaine médical. Cette application sera associée à une démarche d'accompagnement humaine existante nommée TheCompliment, apportée par la société Mandarine CODI, soutenue par le monde médical et notamment des psychologues, garants du bien-fondé de la démarche.

L'objectif est qu'un salarié en difficulté psychologique puisse utiliser l'application pour obtenir des informations sur son état. Il pourra échanger avec l'application dans un premier temps par écrit, et à terme à l'oral. L'application mobile à concevoir proposera à l'utilisateur des informations sur la problématique évoquée et des orientations pour la résoudre, y compris des contacts de praticiens ou structures vers lesquelles se tourner. À terme, selon le vocabulaire utilisé par la personne, l'application pourrait déterminer l'émotion de l'utilisateur et formuler une réponse adaptée au mal être de la personne.

État de l'art et verrous scientifiques

Les chatbots destinés à la santé mentale, tels que *Woebot* ou *Youper*, ont ouvert la voie à l'utilisation d'agents pour fournir un soutien psychologique de première ligne. Cependant, la qualité de leur accompagnement reste très variable et leur validation clinique limitée. La construction de modèles fiables dépend largement de la disponibilité de données pertinentes pour entraîner/fine-tuner les modèles. Bien que plusieurs corpus existent, ils demeurent restreints en taille et en diversité. Des initiatives récentes, comme MentalChat16K (Xu et al., 2025), combinent dialogues synthétiques et données réelles anonymisées pour couvrir différents troubles (dépression, anxiété, deuil, etc.). Ces approches permettent de pallier en partie le déficit de données, mais soulèvent encore des questions méthodologiques sur la validité des dialogues générés artificiellement.

Par ailleurs, l'évaluation de ces solutions d'IA représente un véritable défi. Steenstra & Bickmore (2025) ont proposé une taxonomie des risques propres aux agents de psychothérapie alimentés par IA, allant de la diffusion d'informations inexactes à la dépendance psychologique excessive des utilisateurs. Cette taxonomie constitue un outil clé pour anticiper les dérives et mettre en place des garde-fous.

L'un des défis majeurs réside dans la frugalité énergétique. Les LLM actuels sont notoirement coûteux en ressources computationnelles, ce qui compromet leur déploiement en contexte clinique ou grand public. Des approches comme le distillation learning, la quantization ou encore l'utilisation de LLM spécialisés de taille intermédiaire (Mistral, LLaMA 2/3, Falcon) constituent des pistes prometteuses pour réduire l'empreinte carbone sans compromettre la qualité de l'accompagnement (Xu et al., 2025 ; Guo et al., 2024).

Objectifs du stage

Ce stage vise à développer un agent conversationnel, capable d'accompagner les individus dans un cadre pré-clinique, tout en intégrant des connaissances scientifiques et cliniques validées.

Les objectifs de ce stage peuvent être résumés comme suit :

1. Implémenter un agent conversationnel adaptatif et frugal. Nous visons une approche permettant des interactions personnalisées et engageantes, tout en préservant la validité clinique des résultats.

2. S'assurer qu'un compromis est possible entre performance conversationnelle, sobriété computationnelle et respect de la confidentialité.

Méthodologie envisagée

La méthodologie suivra une approche en quatre volets complémentaires :

1. revue de littérature (solutions existantes),
2. constitution et enrichissement des données pour l'entraînement et l'évaluation de la solution proposée,
3. conception et adaptation du modèle,
4. et évaluation psychologique et informatique.

Pour assurer un déploiement frugal, nous retiendrons un modèle open-source de taille intermédiaire (par exemple Mistral 7B ou LLaMA 3 8B). Ce modèle sera affiné par un fine-tuning supervisé sur les corpus enrichis de dialogues et de questionnaires, qui permettront de relier les réponses des utilisateurs à des savoirs cliniques validés.

L'optimisation du système passera également par des techniques de quantization et de distillation, afin de réduire la taille mémoire et de rendre possible un déploiement hybride : les échanges courants pourront être traités localement (edge), tandis que les analyses plus complexes, notamment l'interprétation fine des questionnaires, seront déportées sur un serveur sécurisé (cloud).

Évaluation

L'évaluation du prototype reposera sur deux volets complémentaires :

1. une **évaluation humaine**, conduite avec la participation d'experts (notamment des psychologues), afin d'analyser la qualité perçue des interactions ;
2. une **évaluation quantitative** fondée sur diverses métriques mesurant la **cohérence** et la **fluidité des dialogues**, la **frugalité computationnelle** (consommation CPU/GPU, usage mémoire, empreinte carbone) ainsi que la **protection de la confidentialité**.

L'ensemble du processus d'évaluation respectera strictement les **standards réglementaires en vigueur** (RGPD, HIPAA).

Contributions attendues

Les résultats escomptés incluent :

- une architecture d'agent conversationnel, indépendante des GAFAM,
- des méthodes pour enrichir un LLM avec des données scientifiques et cliniques, incluant questionnaires et dialogues multi-turn,
- une analyse des compromis entre performance, frugalité et confidentialité,
- une première évaluation de l'impact d'un tel agent comme compagnon numérique dans le domaine du bien-être et de la santé mentale.

Références

Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). *Large Language Model for Mental Health: A Systematic Review*. arXiv preprint arXiv:2403.15401.

Sobowale, K., & Humphrey, D. K. (2025). *Evaluating the Quality of Psychotherapy Conversational Agents: Framework Development and Cross-Sectional Study*. JMIR Formative Research, 9, e65605.

Xu, J., Wei, T., Hou, B., Orzechowski, P., Yang, S., Jin, R., Paulbeck, R., Wagenaar, J., Demiris, G., & Shen, L. (2025). *MentalChat16K: A Benchmark Dataset for Conversational Mental Health Assistance*. arXiv preprint arXiv:2503.13509.

Steenstra, I., & Bickmore, T. W. (2025). *A Risk Taxonomy for Evaluating AI-Powered Psychotherapy Agents*. arXiv preprint arXiv:2505.15108.