**Rapport Final**

**Apprentissage Profond**

**8INF892**

**Kévin Bouchard**

**Université du Québec à Chicoutimi**

**Département d’Informatique**

**Jean-Philippe Larouche**

**LARJ07089501**

**29/04/2024**

# Introduction

# Contexte

Ce projet s’inscrit dans le domaine de la reconnaissance d’activité humaine (RAH), la tendance des LLMs et comme première étape d’un plus grand projet de recherche. L’approche actuelle en RAH est d’utiliser des modèles d’apprentissage profond qui son entrainer spécifiquement pour reconnaitre une ou un petit ensemble d’activités spécifiques. De plus, l’entrainement de ces modèles nécessitent beaucoup de données de qualité. Cela demande un grand effort d’annotation, nettoyage, de structuration et de collecte de la part de chercheurs.

Dans le contexte du cours d’apprentissage profond (8INF892) à l’UQAC, nous devions entreprendre un projet avec l’objectif d’avancer nos connaissances du domaine. Dans cette optique, nous avons pris la décision d’entreprendre un projet sur les *Large Lanugage Models* (LLM) dans le contexte de la reconnaissance d’activités humaines. Plus spécifiquement, nous avons fait une étude comparative sur différents LLMs populaire en ce qui concerne leurs capacités à interpréter des données brutes provenant de différents capteurs.

# Objectifs

Tel que mentionné, ce projet est une première étape d’un plus grand projet de recherche qui à l’objectif de déterminer si les LLMs peuvent reconnaitre les activités performées par un humain en utilisant les données brutes provenant des capteurs situés au sein d’un habitat intelligent. Dans ce projet, nous commençons par évaluer les capacités de différents LLMs à traiter des données de capteurs simples dans le contexte de la RAH. Nous nous sommes inspirés de l’article *HARGPT : Are LLMs Zero-Shot Human Activity Recognizers?* de Ji et al. [1] qui effectue une analyse similaire avec GPT4. L’objectif de ce projet est de déterminer les capacités des différents LLMs populaire sur le marché afin d’être en mesure d’établir ceux qui seront utilisés pour les prochaines étapes du plus grand projet. Comme objectifs secondaires, nous voulons également :

* Évaluer la structure des réponses des modèles
* Déterminer la quantité de donnée à fournir et si nous devons faire un traitement
* Connaître l’information qui doit être fournit aux modèles avec les données
* Avoir un aperçu de la compréhension des données par les modèles

# Chercheurs

Ce projet de recherche fût effectué seul par Jean-Philippe Larouche, étudiant à la maîtrise de recherche en Informatique.

# Méthodologie

# Données

Les données représentent une des composantes les plus importantes de ce projet. Tel que mentionnée dans l’introduction, de bonnes données sont essentielles afin d’être en mesure de faire de la RAH. Initialement, nous avions prévus utilisés des données provenant d’un environnement « complexe », soit les données provenant du laboratoire de domotique du LIARA. Cependant, nous avons rapidement réalisé l’ampleur de la tâche de structuration de ces données afin de pouvoir les utilisés. Étant donné que notre objectif est d’évaluer les modèles de type LLM et non de faire de la structuration de donnée complexe, nous avons alors décidé de uniquement prendre l’ensemble de donnée Capture24 [2].

L’ensemble de donnée Capture24 contient des données d’un capteur de poignet porter par 151 participants pendant 24 heures. Cela nous apporte une grande variété d’activités quotidiennes performer par les humains tel que marcher avec son chien, laver la vaisselle, dormir, écouter le TV, etc. Les données sont annotées et séparées par participants de base et les données de capteur sont recueillis à 100Hz. Alors que dans le cas du projet HARGPT les données sont réduites à 10Hz, nous avons décidés de les garder à 100Hz car nous voulons analyser les capacités des modèles dans le cas où les données sont fournies directement des capteurs sans modifications.

Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

Figure 1 : Les différentes activités dans Capture24

Une image contenant texte, capture d’écran, Rectangle

Description générée automatiquement

Figure 2 : Exemple de donnée d’un participant qui dors

Nous avons commencé par déterminer les activités que nous voulions évaluer, soit : marcher (walking), conduire (driving), dormir (sleeping) et manger (eating). Nous avons choisi ces activités pour des raisons bien précises. Premièrement, marcher et dormir présente des activités « simple » avec des données qui ont tendances à être plus constantes, mêmes de participants différents. Deuxièmement, les activités manger et conduire sont des activités « complexes » avec une plus grande variance de données entre les différents participants et même pour un même participant.

Pour ce qui est des participants, nous avons décidés d’utiliser cinq participants qui serons commun à tous les tests pour toutes les activités. Cela veut dire que le modèle aura déjà « vue » le participant en question. Puis, pour chaque test nous allons utiliser cinq autre participant qui seront uniquement utilisé pour l’activité testé. Cela est fait pour introduire de la variété et de la nouveauté au modèle lors des tests.

# Modèles

Pour effectuer notre analyse, nous avons choisi quatre modèles de type LLM. Nous avons pris les quatre modèles les plus populaires mais qui demeurent facilement accessible (gratuite) étant donnée la nature du projet (pas de financement). De plus, ces modèles demeurent les modèles les plus utiliser parmi la majorité des gens ce qui rend l’analyse intéressant. Finalement, HARGPT effectue leurs analyses sur GPT 4 alors la différence de performance avec ce modèle est fort intéressante à analyser. Les modèles choisis sont :

* GPT 3.5
* Gemini
* Llama 3
* Claude 3

L’interaction avec les modèles se fait via leurs plateformes web respective a l’exception de Claude 3 qui se fait via la plateforme [LmSys](https://chat.lmsys.org/). Nous avons initialement envisagé utiliser LMStudio pour certains modèles afin de les rouler en local. La demande sur le système rendait le processus très lent et n’était pas optimale pour notre contexte alors nous avons utilisés que les plateformes web.

Une image contenant texte, capture d’écran, logiciel, Logiciel multimédia

Description générée automatiquement

Figure 3 : Page d’accueil de Gemini

Figure 4 : Page d’accueil de LM Studio

# Invite de texte

Étant donné que la communication avec les modèles se fait uniquement par invite de texte, la structure et formulation de ce dernier auras un grand impact sur comment le LLM performe. Nous nous sommes inspirés de ce que les chercheurs ont fait dans le projet HARGPT et avons décidés de prendre une structure similaire vu qu’ils ont eu de bonne performance avec cette structure et cela va également nous permettre de comparer nos résultats avec leurs travaux.

La première phrase de l’invite de texte est une instruction qui indique le contexte au modèle. Ensuite, nous fournissons les données de l’activité en question avec les paramètres spécifiques du test. Par exemple, un des tests est de fournir les données avec une liste des activités possibles alors ceux-ci seront décrit à cette étape avec les données. Finalement nous demandons au modèle de nous fournir une réponse. Des travaux récents [3] ont démontré que les LLMs performent mieux lorsque ont leurs demandes d’expliquer leurs réponses étape par étape. Ce type d’instruction exécute un processus « chain of thought » chez le modèle qui apporte un raisonnement supérieur. Alors, nous demandons une réponse ouverte étape par étape au modèle.

Une image contenant capture d’écran, texte

Description générée automatiquementUne image contenant capture d’écran, texte

Description générée automatiquement

Figure 5 : Exemples de prompt avec données

# Tests

Initialement nous avions prévus une grande série de tests à effectuer sur les modèles. Cependant, les temps octroyés à ce projet était limité par la nature du contexte académique étudiant (cours, session, etc.). Nous avons alors identifié trois tests distincts qui sont les plus pertinents afin d’atteindre nos objectifs. Voici les tests qui ont été effectués :

* Zero-Shot avec explication des données et liste d’activités
  + Ce test ressemble beaucoup à celui effectuer par le projet HARGPT. Nous fournissons une description des données avec une liste d’activités possibles au modèle. Cela nous permet de voir la performance du modèle dans un scénario réel et nous permet également de facilement comparer avec les résultats de HARGPT.
* Zero-Shot avec retrait de données
  + Ce test est identique au précédent à l’exception que nous avons retirés des données de manière aléatoire dans les trois axes. L’objectif est d’évaluer la résilience du modèle lorsque des données serait manquante ce qui est très possible dans un contexte de reconnaissance en temps réel.
* Explication des données
  + Ce test à pour but d’évaluer la capacité du modèle à interpréter et comprendre les données. On lui fournit un jeu de données sans explications et on lui demande ce qu’il pense que les données sont et ce qu’elles représentent.

# Métriques

Les métriques d’évaluation que nous avons utilisées sont parmi les plus utilisés dans le domaine. Nous avons choisi c’est métrique pour faciliter la compréhension des résultats et de fournir un comparatif facile entre notre travail et d’autres projets de recherches. Les métriques sélectionnées sont :

* Precision
  + Cette métrique représente le nombre d’activités correctement identifier sur le nombre total de requête identifier comme positif.
* Recall
  + Cette métrique représente le nombre de vrais positifs sur le nombre total de vrais positifs pour toutes les activités.
* F1-score
  + Cette métrique est basée sur les deux métriques précédentes afin de lui donner un *score*.

# Résultats

# Conclusion

# Références

[1] Ji S., Zheng X., Wu C. 2024. HARGPT: Are LLMs Zero-Shot Human Activity Recognizers. arXiv:2403.02727 [cs.CL]. <https://doi.org/10.48550/arXiv.2403.02727>

[2] Chan Chang, S., Walmsley, R., Gershuny, J., Harms, T., Thomas, E., Milton, K., Kelly, P., Foster, C., Wong, A., Gray, N., Haque, S., Hollowell, S., & Doherty, A. (2021). Capture-24: Activity tracker dataset for human activity recognition. University of Oxford.

[3] Wei J., Wang X., Schuurmans D., Bosma M., Xia F., Chi E., Le Q. V., Zhou D., et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in NeuralInformation Processing Systems 35 (2022), 24824–24837