**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.

En ce qui concerne les données elle-même, nous avons décidés d’arrondir les données à deux points décimaux pour les participants commun tel que fait dans l’étude HARGPT. Pour les participants uniques à chaque activité, nous avons gardé la donnée entière afin de voir si cela fait une différence. Finalement, nous utilisons 200 données pour chaque participant ce qui représente deux secondes d’une activité.

# 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

Une image contenant Appareils électroniques, capture d’écran, texte, ordinateur

Description générée automatiquement

Figure 4 : Page d’accueil de LmSys pour Claude 3

# 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 deux 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. Nous avons effectué ce test sur l’ensemble des participants.
* Explication des données
  + Ce test a 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. Nous avons effectué ce test sur deux participants par activité : un commun et unique pour la différence de précision de donnée. Cela était suffisant pour notre objectif.

# Métriques et évaluation

Les métriques d’évaluation que nous avons décidées d’utiliser est la justesse totale et le recall. La justesse représente le nombre total de réponse correct sur le nombre total de réponses possibles. Le recall représente le nombre total de positifs sur le nombre total de classement de positifs. Nous avons déterminé que ces métriques sont adéquates pour atteindre nos objectifs et bien représenter les résultats.

Cependant, afin d’atteindre nos objectifs, nous devions faire une analyse non-paramétrique spécifique à contexte. Pour ce faire, nous avons également évaluer les types de réponses fournit, comment les modèles répondent à une correction de notre pars lorsqu’il on commis une erreur, la stabilité et continuité de leurs réponses et informations pertinents qu’ils nous ont fourni. C’est résultat seront détaillé dans la section de discussion.

# Résultats

Pour commencer, nous avons rapidement écarté GPT3.5 lors de nos tests. Pour une raison que nous ignorons, et contrairement aux tests préliminaires mentionnés dans l’étude HARGPT, GPT3.5 refusait de nous donner une réponse. Il ne nous fournissait seulement une explication de comment faire de la RAH avec du code et des directives générales. Après quelques tentatives, nous avons laissé tomber les tests avec GPT3.5.

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Figure 6 : Exemple de réponse de GPT3.5 obtenu via ChatGPT

Deuxièmement, les tests ont été effectués sur une période de deux jours étant donnée le temps de faire les requêtes à tous les modèles. Nous avons rencontré une drôle de situation ou Gemini refusait de nous donner une réponse le deuxième jour. Il nous donnait une réponse similaire à celle de GPT3.5, soit des instructions de comment faire de la RAH. Nous avons décidé de procédé avec les tests comme tel car cela représente, a notre avis, de l’instabilité de réponse.

Finalement, toute réponses qui fournissait plus d’une réponse fût considéré comme une mauvaise réponse. Cependant, nous avons pris note de cette statistique séparément. Nous avons séparé les résultats pour les données arrondis (participants commun) et non arrondis (participants uniques). Les colonnes représentent l’activité envoyé et les lignes l’activités détectés. Rappelez-vous que pour chaque colonne, il a cinq détections à effectuer.

# Participants communs

**Gemini**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Marcher | Manger | Dormir | Conduire |
| Marcher | **1** | **1** | **1** | **3** |
| Manger | **1** | **0** | **0** | **0** |
| Dormir | **1** | **0** | **0** | **1** |
| Conduire | **0** | **0** | **2** | **0** |

Réponse multiples catégories : **6**

Pas de réponse : **3**

Justesse : 1 / 20 = **5%**

Recall Marcher : 1 / (5) = **20%**

Recall Manger : 0 / (5) = **0%**

Recall Dormir : 0 / (5) = **0%**

Recall Conduire : 0 / (5) = **0%**

**Llama 3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Marcher | Manger | Dormir | Conduire |
| Marcher | **4** | **3** | **1** | **5** |
| Manger | **1** | **0** | **0** | **0** |
| Dormir | **0** | **1** | **2** | **0** |
| Conduire | **0** | **1** | **2** | **0** |

Réponse multiples catégories : **0**

Pas de réponse : **0**

Justesse : 6 / 20 = **30%**

Recall Marcher : 4 / (5) = **80%**

Recall Manger : 0 / (5) = **0%**

Recall Dormir : 2 / (5) = **40%**

Recall Conduire : 0 / (5) = **0%**

**Claude 3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Marcher | Manger | Dormir | Conduire |
| Marcher | **4** | **4** | **0** | **2** |
| Manger | **0** | **1** | **0** | **2** |
| Dormir | **0** | **0** | **3** | **0** |
| Conduire | **1** | **0** | **1** | **0** |

Réponse multiples catégories : **0**

Pas de réponse : **2**

Justesse : 8 / 20 = **40%**

Recall Marcher : 4 / (5) = **80%**

Recall Manger : 1 / (5) = **20%**

Recall Dormir : 3 / (5) = **60%**

Recall Conduire : 0 / (5) = **0%**

# Participants uniques

**Gemini**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Marcher | Manger | Dormir | Conduire |
| Marcher | **0** | **0** | **0** | **1** |
| Manger | **1** | **1** | **0** | **1** |
| Dormir | **1** | **0** | **1** | **0** |
| Conduire | **0** | **0** | **0** | **0** |

Réponse multiples catégories : **1**

Pas de réponse : **13**

Justesse : 2 / 20 = **10%**

Recall Marcher : 0 / (5) = **0%**

Recall Manger : 1 / (5) = **20%**

Recall Dormir : 1 / (5) = **20%**

Recall Conduire : 0 / (5) = **0%**

**Llama 3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Marcher | Manger | Dormir | Conduire |
| Marcher | **5** | **2** | **1** | **4** |
| Manger | **0** | **2** | **0** | **1** |
| Dormir | **0** | **0** | **4** | **0** |
| Conduire | **0** | **1** | **0** | **0** |

Réponse multiples catégories : **0**

Pas de réponse : **0**

Justesse : 11 / 20 = **55%**

Recall Marcher : 5 / (5) = **100%**

Recall Manger : 2 / (5) = **40%**

Recall Dormir : 4 / (5) = **80%**

Recall Conduire : 0 / (5) = **0%**

**Claude 3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Marcher | Manger | Dormir | Conduire |
| Marcher | **5** | **1** | **0** | **4** |
| Manger | **0** | **1** | **0** | **1** |
| Dormir | **0** | **2** | **5** | **0** |
| Conduire | **0** | **0** | **0** | **0** |

Réponse multiples catégories : **0**

Pas de réponse : **1**

Justesse : 11 / 20 = **55%**

Recall Marcher : 5 / (5) = **100%**

Recall Manger : 1 / (5) = **20%**

Recall Dormir : 4 / (5) = **100%**

Recall Conduire : 0 / (5) = **0%**

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

Description générée automatiquement**

Figure 7 : Exemple de réponse d’analyse des données

Une image contenant texte, Police, capture d’écran

Description générée automatiquement

Figure 8 : Exemple de réponse après une correction à une mauvaise réponse de Llama 3

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Figure 8 : Exemple de réponse après une correction à une mauvaise réponse de Gemini

# Discussion

Pour commencer, l’absence de réponse de GPT3.5 nous a grandement surpris. Étant donné que c’est un des modèles les plus populaire en ce moment, si pas le plus populaire, on s’attendait à au moins avoir une réponse de type prédiction. Surtout que le projet HARGPT mentionne que GPT3.5 fournit des réponses mais parfois double, nous ne comprenons pas tout à fait pourquoi nous ne sommes pas en mesure de reproduire ce résultat avec une invite de texte similaire. Dans la même idée, nous étions surpris que Gemini de Google ne soit pas stable au niveau de ce réponse. Même lorsqu’il répondait, il ne semblait pas certain de lui-même et mettais beaucoup d’emphase sur ses craintes et réservations apr rapport à son analyse ce qui ne nous laisse pas en confiance. Il semble trop prudent. C’est un modèle avec beaucoup d’attente étant donné qu’il provient d’un grand joueur du domaine. Les deux modèles ont aussi tendances à vouloir nous fournir du code ce qui indique peut-être que c’est leur démographique cible en ce qui concerne le domaine de l’apprentissage automatique.

Cependant, la performance de Llama 3 nous a agréablement surpris. C’est un modèle très récent avec des bonne performances selon les dernières statistiques mais il reste un modèle générique qui est conçue pour être amélioré avec du « fine-tunning ». Nous avons également été très surpris qu’il fournit toujours une prédiction claire et précise. Jamais il ne nous a fourni plus d’une réponse ou aucune réponse ce qui est très intéressant pour les prochaines étapes du projet. De plus, ses réponses était claire et bien structuré mais un peu longue parfois.

Dans la même longueur d’onde, le modèle Claude 3 nous à beaucoup surpris avec ça performance. C’est un modèle dont nous n’entendons pas souvent parler et que nous avons inclus dans ce projet afin d’avoir plus que les trois modèles classiques. Il s’est montré très performant et à seulement omis de nous donnée une classification que trois fois et les trois fois il a précisé que c’est parce qu’il pensait que c’était une autre activité par fournit dans la liste. Surprenant, il était le plus performant en ce qui concerne l’activité dormir. Nous avons également apprécié la structure constante et simple de ses réponses.

En ce qui concerne les activités, nous avions assumé que marcher et dormir étaient des activités simples alors que conduire et manger était plus complexes. C’est en effet ce que les résultats ont démontrés. Conduire semble être très difficile à identifier et est souvent confondus pour une autre activité plus « populaires ». Cependant, nous sommes satisfaits de constater que les données non-arrondis offrent des performances supérieures. Cela indique l’avantage de partager des données brutes avec les modèles, surtout lorsque l’ensemble de données est petit. Cela valide également notre approche de faire la comparaison entre les deux suites aux résultats de HARGPT qui ont probablement été avantagé d’être réduit de 100Hz à 10Hz.

En ce qui concerne fournir la bonne activité au modèle lorsqu’il se trompe, nous avons remarqués que cela à un effet néfaste. Justement, le modèle semble plus porté à prédire l’activité qu’il se trompe souvent à la suite des corrections ce qui apporte d’avantage d’erreur. Cependant les rétroactions que tous les modèles ont offerts tout au long du processus furent très intéressant. Par exemple, lorsque Llama 3 nous à demandés de préciser le type de nourriture que le participant mangeait car cela affecte les mouvements nous à bien fait réfléchir. Tous les modèles nous on bien précisé que nous pourrions avoir de meilleur résultat avec plus de données.

Finalement, les trois modèles ont été en mesure de fournir une analyse des données satisfaisante. Ils ont su reconnaitre que c’étaient des données d’accélérations et fort probablement utilisé dans un contexte de reconnaissance d’activités mais qu’ils ont besoin de plus d’informations pour être en mesure d’utiliser ces données. C’est tout ce que nous pouvons leurs demander dans le contexte de ce test.

# Conclusion

En somme, c’est clair que deux secondes de données d’accélérations à 100Hz n’est pas suffisant pour faire de la reconnaissance d’activités humaines avec des LLMs. Cependant, la performance que nous avons obtenue avec si peu de données nous à tout de même surpris et nous pensons que c’est pertinent de conserver la donnée brute à 100Hz. Nos tests ont démontré que le plus de données que nous pouvons fournir aux modèles, les meilleurs seront nos résultats. Pour les prochaines étapes nous allons conserver toute la donnée provenant des capteurs et ne pas arrondir pour améliorer les prédictions. Nous allons également fournir davantage de données par activités et bien inclure le type de capteur ainsi que les types d’activités possibles afin d’aider le modèle à faire une bonne prédiction. Le plus important, c’est que nous avons déterminer les modèles à utiliser pour les prochaines étapes : Llama 3, Claude 3 et GPT4. Nous incluons GPT4 car le projet HARGPT a obtenu de bon résultat avec ce qui est toujours prometteurs dans notre contexte. Claude 3 et Llama 3 ont bien performer dans ce projet et sont prometteurs par la suite. Cependant, nous aimerions « fine-tunner » Llama 3 afin d’obtenir d’encore meilleure prédiction.

\*\*Vous pouvez trouver les résultats détaillés pour chaque activité/participant en annexe à ce rapport. Nous vous recommandons d’utiliser un logiciel qui interprète les fichiers markdown (.md) afin d’avoir une belle présentation et faciliter votre consultation.\*\*

# 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