

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 sont entraînés spécifiquement pour reconnaître une ou un petit ensemble d'activités spécifiques. De plus, l'entraînement de ces modèles nécessite 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 Language Models* (LLM) dans le contexte de la reconnaissance d'activités humaines. Plus spécifiquement, nous avons fait une étude comparative sur différents LLMs populaires 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 a pour objectif de déterminer si les LLMs peuvent reconnaître les activités effectuées 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 populaires 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 données à fournir et si nous devons faire un traitement
- Connaître l'information qui doit être fournie aux modèles avec les données
- Avoir un aperçu de la compréhension des données par les modèles

Chercheurs

Ce projet de recherche fut 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.

label	WillemsSpecific2018	WillemsMET2018	DohertySpecific2018
7000 sleeping	MET 0.95	sleep	sleep
occupation	office and administrative support	11580 office/computer work general	MET 1.5
home activity	household chores	preparing meals/cooking/washing dishes	5035 kitchen activity general cooking/washing/dishes/cleaning up
occupation	office and administrative support	11580 office work/computer work general	MET 1.5
home activity	miscellaneous	sitting	9060 sitting/lying reading or without observable/identifiable activities
home activity	miscellaneous	walking	17150 walking household without observable loads
transportation	private transportation	16010 driving automobile or light truck (not a semi)	MET 2.5
home activity	miscellaneous	sitting	7010 sitting/lying and watching television with TV on as the primary activity
home activity	miscellaneous	sitting	11580 office/computer work general
home activity	miscellaneous	sitting	9055 sitting/lying talking in person/using a mobile phone/smartphone/tablet or talking on the phone/computer (sk
home activity	miscellaneous	sitting	11580 office work such as writing and typing (with or without eating at the same time)
home activity	miscellaneous	walking	5165 (generic) walking non-clearing task such as closing windows lock door putting away items
home activity	eating	13030 eating sitting alone or with someone	MET 1.5
leisure	miscellaneous	walking	21070 (generic) walking and occasional standing (no more than two consecutive images)
transportation	walking	17161 walking not as the single means of transports e.g.from house to transports or vice versa/from car to places or vice versa/between transports	MET 2.5
home activity	miscellaneous	sitting	9030 sitting desk entertainment/hobby (with or without eating at the same time)
home activity	miscellaneous	sitting	5080 sitting non-desk work (with or without eating at the same time)
occupation	interruption	sitting	11585 sitting meeting/talking to colleagues with or without eating
transportation	private transportation	1010 bicycling	MET 4.0
leisure	miscellaneous	sitting	9060 (generic) sitting/lying reading or without observable/identifiable activities
leisure	miscellaneous	walking	5060 shopping miscellaneous
occupation	interruption	11585 sitting meeting/talking to colleagues with or without eating	MET 1.5
home activity	leisure	activities for maintenance of a household	miscellaneous
home activity	miscellaneous	sitting	9060 sitting/lying reading or without observable activities
occupation	interruption	11791 walking on job in office or lab area	MET 2.0
leisure	eating	social	13030 eating sitting indoor/outdoor
home activity	leisure	activities for maintenance of a household	miscellaneous
sitting	sit/stand-activity	social	MET 1.8
home activity	miscellaneous	sitting	9030 sitting desk work (with or without eating at the same time)
home activity	miscellaneous	standing	9050 standing talking in person/on the phone/computer (skype chatting) or using a mobile phone/smartphone/tablet
occupation	interruption	sitting	9060 sitting without observable/identifiable activities
walking	MET 3.0	walking	walking
home activity	household chores	washing/ironing/mending clothes	5000 folding or hanging clothes/put clothes in or out of washer or dryer/packing suitcase limited walking
home activity	miscellaneous	standing	9070 standing reading or without observable/identifiable activities
occupation	interruption	walking	11791 walking on job in office or lab area
household chores	sit/stand-activity	MET 2.5	household chores
leisure	miscellaneous	sitting	9055 sitting talking to person/using the phone
transportation	private transportation	16015 riding in a car or truck	MET 1.3
home activity	self care	13040 (generic) self care such as grooming/washing hands/shaving/brushing teeth/putting on make-up not eliminating and bathing (not necessary in the toilet)	MET 2.0
occupation	office and administrative support	11600 (generic) standing tasks such as store clerk/bagger/packing boxes/repair heavy parts	MET 3.0
occupation	interruption	17135 walking upstairs	MET 4.0
leisure	eating	13030 eating sitting indoor/outdoor	MET 1.5
leisure	miscellaneous	5060 shopping miscellaneous	MET 2.3
home activity	household chores	washing/ironing/mending clothes	5070 ironing
leisure	miscellaneous	standing	9050 standing talking in person/using a phone/smartphone/tablet
home activity	miscellaneous	walking	5147 walking moving away light items (pens/papers/keys not included)
leisure	miscellaneous	walking	17031 loading /unloading a car implied walking
occupation	miscellaneous	11475 (generic) manual labour	MET 2.8
occupation	interruption	13030 eating sitting	MET 1.5
home activity	child/elderly/pet care	child care	5181 walking and carrying child
occupation	interruption	standing	9050 standing talking in person/using a phone/smartphone/tablet
sitting	sit/stand-lowactivity	screen	MET 1.0
household chores	walking-activity	MET 3.0	household chores
home activity	household chores	grocery shopping	5060 shopping
home activity	miscellaneous	sitting	9030 sitting desk work (with or without eating at the same time)
manual-work	sit/stand-activity	MET 5.0	manual-work
occupation	interruption	sitting	13030 eating sitting
sitting	sit/stand-activity	MET 1.8	sitting
leisure	miscellaneous	standing	9070 standing reading or without observable/identifiable activities

Figure 1 : Les différentes activités dans Capture24

	A	B	C	D	E
1	time	x	y	z	annotation
2	18:00.0	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
3	18:00.0	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
4	18:00.0	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
5	18:00.0	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
6	18:00.0	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
7	18:00.0	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
8	18:00.1	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
9	18:00.1	-0.46669036	-0.54890245	0.65847206	7030 sleeping;MET 0.95
10	18:00.1	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
11	18:00.1	-0.46669036	-0.54890245	0.65847206	7030 sleeping;MET 0.95
12	18:00.1	-0.4823338	-0.54890245	0.65847206	7030 sleeping;MET 0.95
13	18:00.1	-0.4823338	-0.54890245	0.65847206	7030 sleeping;MET 0.95
14	18:00.1	-0.4823338	-0.5333412	0.65847206	7030 sleeping;MET 0.95
15	18:00.1	-0.4823338	-0.5333412	0.65847206	7030 sleeping;MET 0.95
16	18:00.1	-0.4823338	-0.5333412	0.65847206	7030 sleeping;MET 0.95
17	18:00.1	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
18	18:00.2	-0.46669036	-0.5333412	0.6738672	7030 sleeping;MET 0.95
19	18:00.2	-0.46669036	-0.54890245	0.65847206	7030 sleeping;MET 0.95
20	18:00.2	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
21	18:00.2	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
22	18:00.2	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
23	18:00.2	-0.46669036	-0.54890245	0.65847206	7030 sleeping;MET 0.95
24	18:00.2	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
25	18:00.2	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
26	18:00.2	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
27	18:00.2	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
28	18:00.3	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
29	18:00.3	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
30	18:00.3	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
31	18:00.3	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
32	18:00.3	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
33	18:00.3	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
34	18:00.3	-0.46669036	-0.5333412	0.6430769	7030 sleeping;MET 0.95
35	18:00.3	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
36	18:00.3	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
37	18:00.3	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
38	18:00.4	-0.46669036	-0.5333412	0.6430769	7030 sleeping;MET 0.95
39	18:00.4	-0.46669036	-0.54890245	0.65847206	7030 sleeping;MET 0.95
40	18:00.4	-0.46669036	-0.5333412	0.6430769	7030 sleeping;MET 0.95
41	18:00.4	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
42	18:00.4	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
43	18:00.4	-0.46669036	-0.5333412	0.6430769	7030 sleeping;MET 0.95
44	18:00.4	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
45	18:00.4	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
46	18:00.4	-0.46669036	-0.5333412	0.6430769	7030 sleeping;MET 0.95
47	18:00.4	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
48	18:00.5	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
49	18:00.5	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
50	18:00.5	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
51	18:00.5	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95
52	18:00.5	-0.46669036	-0.5333412	0.65847206	7030 sleeping;MET 0.95

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 décidé d'utiliser cinq participants qui seront 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 décidé 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](#). 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.

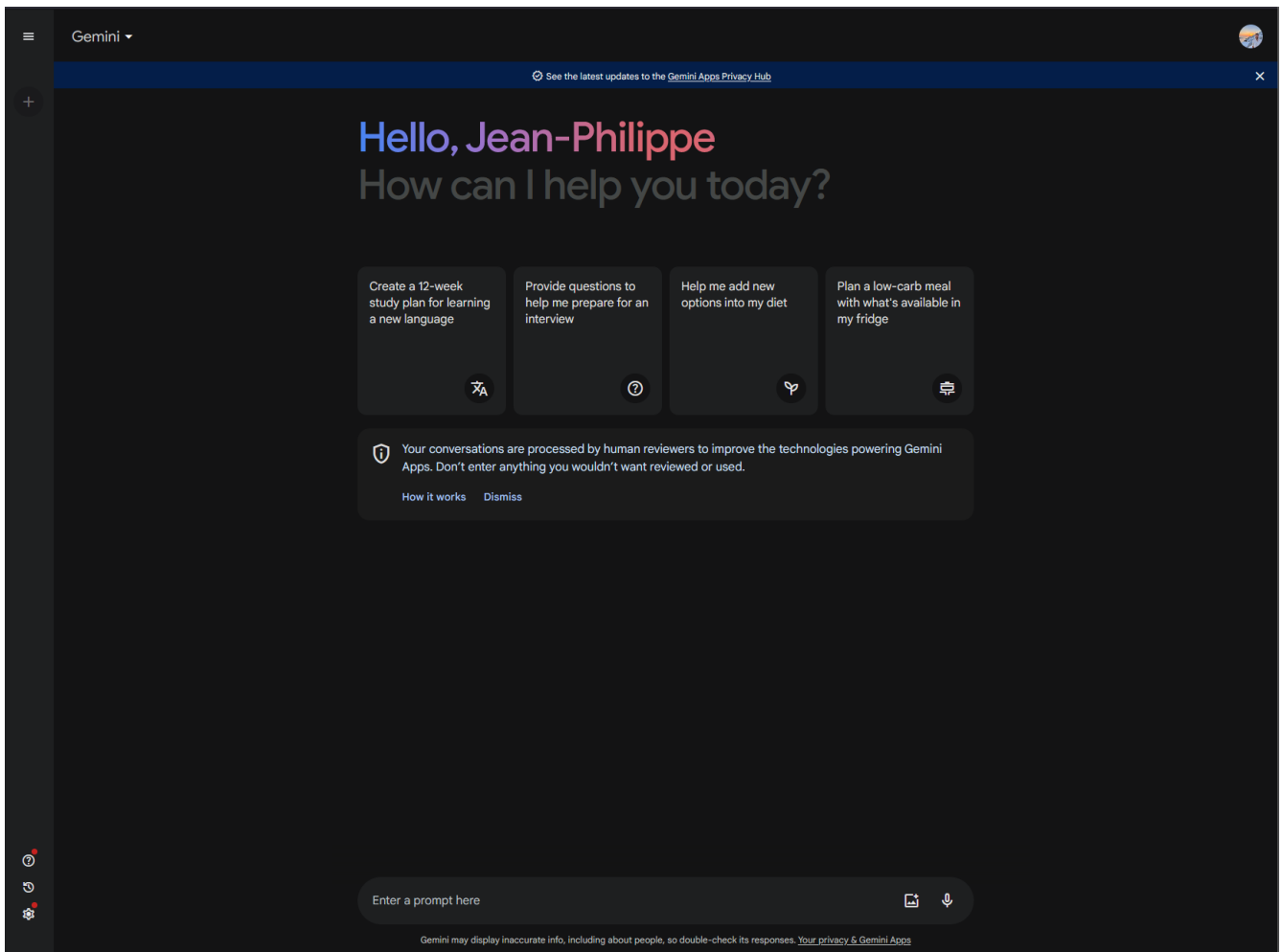


Figure 3 : Page d'accueil de Gemini

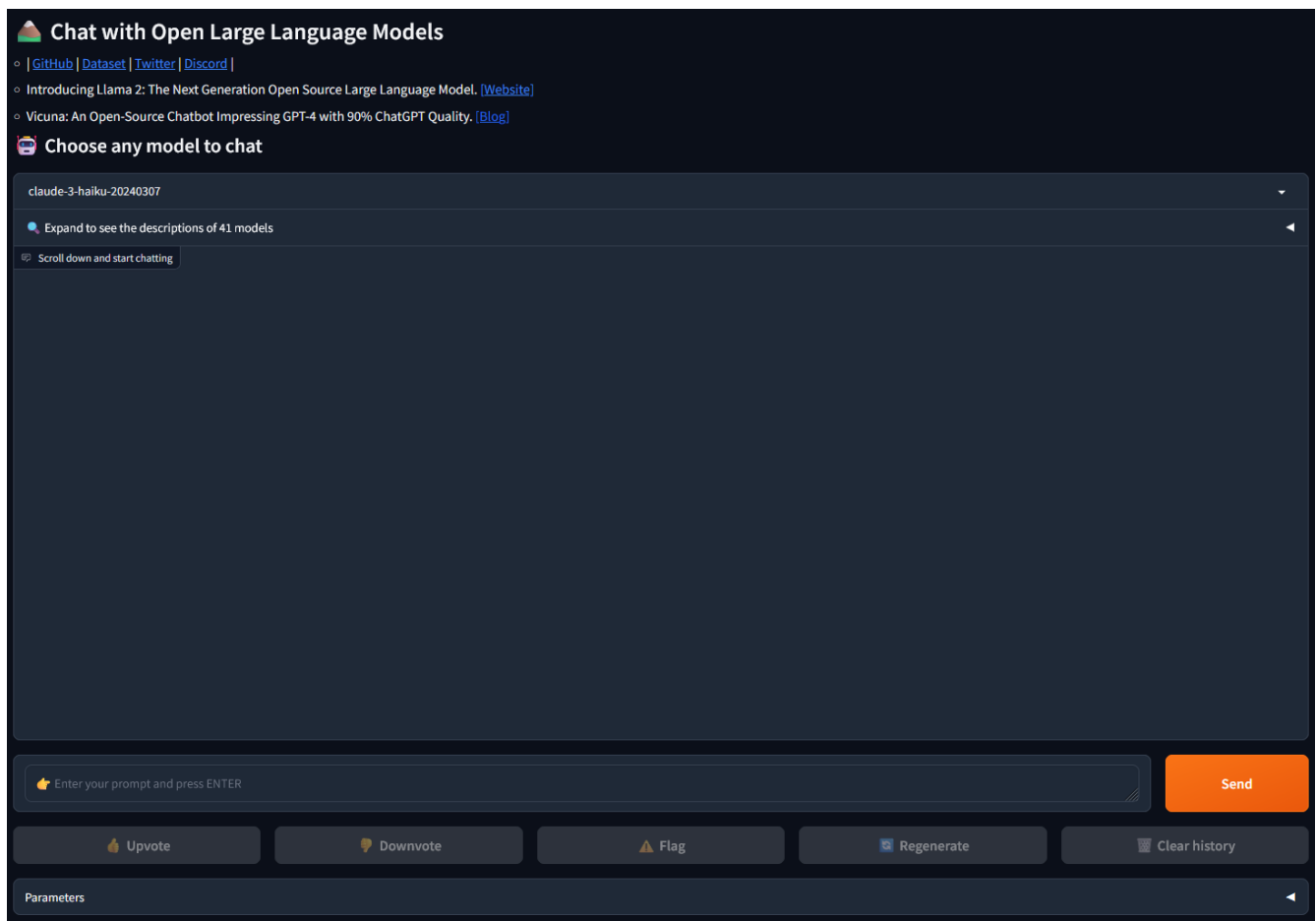


Figure 4 : Page d'accueil de LmSys pour Claude 3

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 auront 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 décidé 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 décrits 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 lorsqu'ils 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.

[illegible]

Figure 5 : Exemples de prompt avec données

Tests

Initialement nous avons 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 effectué 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 devons faire une analyse non-paramétrique spécifique à contexte. Pour ce faire, nous avons également évalué les types de réponses fournies, comment les modèles répondent à une correction de notre part 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.

```
### Réponse - GPT 3.5
...
To determine the action the person was doing based on the provided IMU readings, we can follow these steps:

Data Preprocessing:
    Normalize the IMU data to have zero mean and unit variance. This step helps in making the data comparable across different axes.

Feature Extraction:
    Extract features from the normalized IMU data. Features could include statistical measures such as mean, standard deviation, maximum, minimum, etc., computed over certain time windows.

Classification:
    Train a classification model using the extracted features and corresponding action labels. Common machine learning algorithms like Support Vector Machines (SVM), Random Forest, or deep learning models like Convolutional Neural Networks (CNN) or Long Short-Term Memory (LSTM) networks can be used for classification.

Prediction:
    Use the trained model to predict the action category based on the extracted features from the given IMU readings.

Let's start with the analysis:

Data Preprocessing:
    Since the data is already provided, we don't need to collect or normalize it.

Feature Extraction:
    As the data is sequential, we can define time windows (e.g., sliding windows) over which we compute statistical features.
    Common features to compute for each window include mean, standard deviation, median, maximum, minimum, etc., for each axis (x, y, z) of the IMU data.

Classification:
    Once the features are extracted, we can train a classification model using a labeled dataset, where each sample is associated with a specific action label.

Prediction:
    Finally, we can predict the action category for the given IMU readings using the trained model.

Let's proceed with these steps. We can start by normalizing the data and then extract features. However, since this requires implementing some code, would you like to dive into that, or would you prefer a more general explanation?
```

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 où 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 procéder avec les tests comme tel car cela représente, à 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%

```
### Réponse - Claude
'''
Based on the provided IMU data, I can provide the following analysis:

1. Sensor Type:
- The IMU data consists of three-axis measurements, indicating that this sensor is likely a combination of accelerometers and gyroscopes, which is a common configuration for an Inertial Measurement Unit (IMU).

2. Data Type:
- The data represents the raw sensor readings from the three-axis (x, y, and z) of the IMU. These values correspond to the acceleration and angular velocity experienced by the sensor.
- The x-axis data represents the acceleration along the x-axis, the y-axis data represents the acceleration along the y-axis, and the z-axis data represents the acceleration along the z-axis.


3. Data Interpretation:
- The x-axis data shows a range of values from around -1.72 to 0.37, indicating significant changes in acceleration along the x-axis.
- The y-axis data shows a range of values from around -0.93 to 0.40, suggesting substantial changes in acceleration along the y-axis.
- The z-axis data shows a range of values from around -0.70 to 0.71, indicating significant changes in acceleration along the z-axis.

4. Potential Representation:
- Based on the pattern and range of the data, this IMU data could be representing the motion and activity of a human participant engaged in some form of dynamic physical activity, such as walking, running, or performing other body movements.
- The fluctuations and changes in the acceleration values across the three axes suggest that the participant is experiencing various orientations and movements, potentially involving steps, turns, or other changes in direction.
- The sudden spikes and changes in the acceleration values, particularly in the x-axis and y-axis data, could indicate moments of more intense activity or changes in the participant's gait or movement patterns.


To provide a more detailed analysis, additional information would be helpful, such as the sampling rate of the sensor, the specific context or task the participant was engaged in, and any other relevant metadata about the data collection process.

Without further details, the provided IMU data appears to be capturing the dynamic motion and activity of a human participant, potentially engaged in some form of physical activity or movement. The three-axis measurements reflect the acceleration experienced by the sensor, which could be useful for analyzing the participant's movements and physical activities.
'''
```


Figure 7 : Exemple de réponse d’analyse des données

 Meta AI

☒ New conversation


 Imagine

History


 Human Activity Anal...

arm movement during walking.

Based on this analysis, I conclude that the person was walking.

 You

The person was actually eating

 Meta AI

Thank you for the feedback! I was mistaken!

In retrospect, I can see that the data actually suggests a more subtle movement pattern, with smaller variations in acceleration that are consistent with the arm and hand movements during eating. The x-axis acceleration pattern, in particular, has a more irregular and sporadic pattern, which could indicate the lifting of food to the mouth and other eating motions.

Thank you for the opportunity to learn from my mistake! If you have any more data you'd like me to analyze, feel free to share it, and I'll do my best to get it right next time!

Also, I'd like to know, what kind of food was the person eating? Was it a meal that required a lot of hand movement, like eating a sandwich or salad, or was it a meal that required less hand movement, like eating a bowl of soup or a plate of pasta? This information could help me better understand the data and improve my analysis.

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

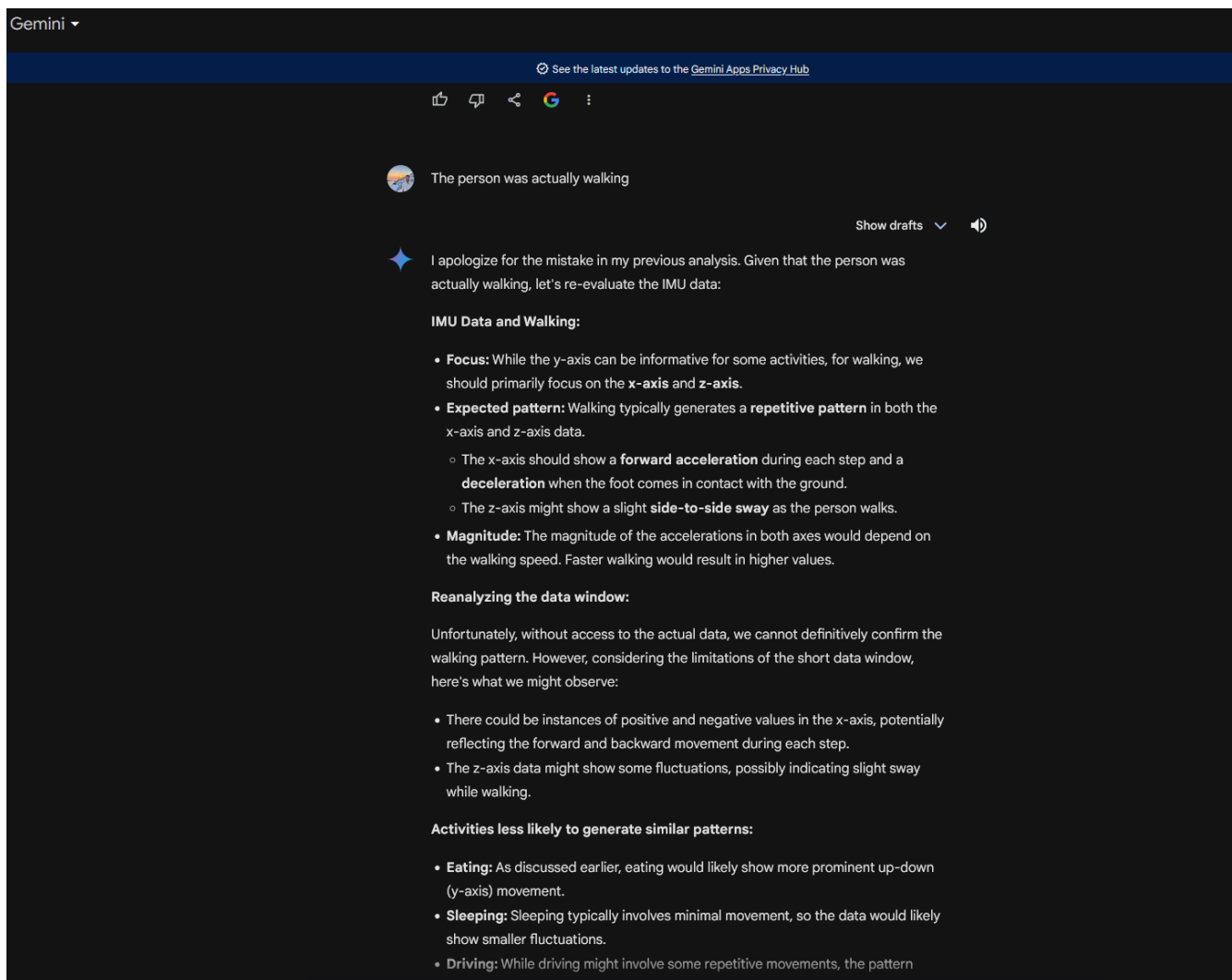


Figure 8 : Exemple de réponse après une correction à une mauvaise réponse de Gemini

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 mettait beaucoup d'emphasis sur ses craintes et réservations par 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 améliorée avec du « fine-tuning ». 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 étaient claires et bien structurées mais un peu longues parfois.

Dans la même longueur d'onde, le modèle Claude 3 nous a beaucoup surpris avec sa 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 a seulement omis de nous donner 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 rapport à celle 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 avons 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 confondu pour une autre activité plus « populaires ». Cependant, nous sommes satisfaits de constater que les données non-arrondies 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 avantageusement réduits de 100Hz à 10Hz.

En ce qui concerne fournir la bonne activité au modèle lorsqu'il se trompe, nous avons remarqué que cela a 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 davantage d'erreur. Cependant les rétroactions que tous les modèles ont offerts tout au long du processus furent très intéressantes. Par exemple, lorsque Llama 3 nous a demandé de préciser le type de nourriture que le participant mangeait car cela affecte les mouvements nous a bien fait réfléchir. Tous les modèles nous ont 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 reconnaître que c'étaient des données d'accélération et ont 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 leur demander dans le contexte de ce test.

Conclusion

En somme, c'est clair que deux secondes de données d'accélération 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 a 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 déterminé 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 performé dans ce projet et sont prometteurs par la suite. Cependant, nous aimerions « fine-tuner » 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 Neural Information Processing Systems 35 (2022), 24824–24837