



HAL
open science

Évaluation des modèles génératifs (LLMs) concernant la détection des problèmes mentaux en basant sur les réseaux sociaux chinois : le cas de l'humeur dépressif

Jinyuan Xu, Tian Lan, Pierre Magistry, Mathieu Valette

► To cite this version:

Jinyuan Xu, Tian Lan, Pierre Magistry, Mathieu Valette. Évaluation des modèles génératifs (LLMs) concernant la détection des problèmes mentaux en basant sur les réseaux sociaux chinois : le cas de l'humeur dépressif. Atelier sur l'évaluation des modèles génératifs (LLM) et challenge d'extraction d'information few-shot, Institut des sciences informatiques et de leurs interactions - CNRS Sciences informatiques [INS2I-CNRS], Jul 2024, Toulouse, France. hal-04678038

HAL Id: hal-04678038

<https://hal.science/hal-04678038v1>

Submitted on 26 Aug 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Évaluation des modèles génératifs (LLMs) concernant la détection des problèmes mentaux en basant sur les réseaux sociaux chinois : le cas de l'humeur dépressif

Jinyuan XU¹ Tian LAN Pierre MAGISTRY¹ Mathieu VALETTE¹

(1) Ertim, Inalco, Paris, France

jinyuan.xu@inalco.fr, gugupigt@gmail.com, pierre.magistry@inalco.fr,
mvalette@inalco.fr

RÉSUMÉ

La technologie du traitement automatique des langues (TAL) est couramment employée pour la détection des problèmes mentaux sur les réseaux sociaux, principalement à l'aide de modèles supervisés qui ont démontré leur efficacité. Face à la montée des modèles génératifs dans divers secteurs, des chercheurs ont entamé leur utilisation pour identifier les signes de dépression, menant à des évaluations initiales. Toutefois, les études évaluatives spécifiques au contexte chinois restent peu fréquentes et restreintes. Cette recherche se concentre sur l'analyse comparative de trois modèles génératifs chinois, exploitant diverses approches de prompt. Comparativement aux classifieurs classiques, les modèles génératifs offrent l'avantage de générer des explications textuelles pour les prédictions, bien qu'ils rencontrent fréquemment des difficultés liées à la cohérence logique et aux illusions textuelles. Le but de cette étude est également de développer une stratégie efficace pour pallier ces limitations.

ABSTRACT

Evaluating Large Language Models (LLMs) for Mental Health Detection on Chinese Social Networks : A Case Study of Depressive Moods

Natural Language Processing (NLP) technology is widely employed for the detection of mental health disorders on social media platforms, predominantly using supervised models that have proven their effectiveness. With the rise of generative models across various sectors, researchers have begun applying these to identify indicators of depression, leading to preliminary assessments. However, evaluative studies within the Chinese context remain scarce and constrained. This research focuses on a comparative analysis of three Chinese generative models, utilizing a variety of prompt strategies. In comparison to classical classifiers, generative models possess the capability to produce textual explanations for predictions, although they frequently encounter challenges related to logical coherence and textual hallucinations. The objective of this study is also to devise an effective strategy to address these challenges.

MOTS-CLÉS : Traitement Automatique des Langues, Modèles Génératifs (LLMs), Problèmes Mentaux, Détection de dépression, Prompt Ingénieur, Extraction des informations.

KEYWORDS: Natural Language Processing, Large Language Models (LLMs), Mental Health, Depression Detection, Prompt Engineering, Information Retrieval.

1 Introduction

Selon les statistiques de l'Organisation mondiale de la santé (OMS)¹, près d'un milliard de personnes dans le monde souffrent de troubles mentaux, l'anxiété et la dépression étant les plus courantes. En Chine, environ 54 millions de personnes souffrent de dépression et 41 millions d'anxiété. Actuellement, les consultations psychologiques sous supervision professionnelle et les échelles et questionnaires de santé mentale sont les méthodes les plus efficaces pour détecter les maladies mentales. Toutefois, ces méthodes requièrent la participation de professionnels médicaux et les patients peuvent être plus stressés à l'idée de rechercher des consultations médicales en raison de l'impact de la stigmatisation(Corrigan & Watson, 2002).

De plus en plus de technologies sont développées pour la détection des problèmes mentaux. Ces technologies peuvent compléter les méthodes traditionnelles de détection à une plus large échelle. Les réseaux sociaux, où les gens d'aujourd'hui documentent leur vie quotidienne et expriment leurs émotions, sont devenus une source de données en raison de la facilité d'accès aux données. Bien que de nombreux modèles génératifs basés sur des données en anglais aient été utilisés pour détecter les humeurs dépressives (Wang *et al.*, 2024) (Xu *et al.*, 2024b) (Yang *et al.*, 2023), les recherches correspondantes dans la communauté chinoise sont relativement rares et limitées (Lan *et al.*, 2024)(Lai *et al.*, 2023). En plus, des études montrent une tendance forte à l'hallucination et à l'incohérence logique lorsqu'ils traitent des émotions (Ji *et al.*, 2023)(Alhamed *et al.*, 2024).

Par conséquent, cette étude vise à évaluer les performances des modèles génératifs les plus populaires, en utilisant différentes stratégies pour la détection des humeurs dépressives non pathologiques sur les plateformes sociales chinoises. Les humeurs dépressives sont une émotion négative et accompagnent souvent une perte durable de la capacité à éprouver du plaisir ou de l'intérêt. Elles peuvent probablement être considérées comme des risques de dépression clinique précoce. Cette étude inclut également des expériences comparatives avec des modèles de classification. Elle cherche à améliorer la capacité de ces modèles à récupérer des informations sur la dépression dans de longs textes, tout en réduisant les risques d'hallucination du modèle.

2 État de l'art

2.1 Méthodes pour la détection des problèmes mentaux

Actuellement, de nombreuses recherches se penchent sur la détection automatique des problèmes mentaux, notamment la dépression. Ces travaux (Le Glaz *et al.*, 2021)(Chancellor & De Choudhury, 2020), qui cherchent à fournir des explications textuelles plus accessibles, reposent sur l'utilisation de techniques d'apprentissage automatique traditionnelles, telles que *Support Vector Machine* (SVM)(Cortes & Vapnik, 1995), *eXtreme Gradient Boosting* (XGBoost)(Chen & Guestrin, 2016), *Random Forest* (RF)(Breiman, 2001), *K-Nearest Neighbors* (KNN)(Peterson, 2009), Régression Logistique (LR) (Yu *et al.*, 2011) et etc., en combinant des caractéristiques statistiques au niveau lexical, textuel et des activités sociales des utilisateurs ou des patients. Ces caractéristiques incluent les sacs de mots, les n-grammes, le TF-IDF (Salton & Buckley, 1988), le LIWC (Tausczik & Pennebaker, 2010) et la textométrie (Valette, 2016)(Lafon, 1980).

1. OMS : <https://www.who.int/health-topics/mental-health>

Même si ces méthodes permettent de capturer individuellement les informations sémantiques, il est possible qu'elles omettent de nombreuses informations contextuelles. Pour mieux intégrer les sémantiques contextuelles, le plongement lexical (*word embedding*) (Devlin *et al.*, 2019) et divers réseaux de neurones sont utilisés dans la détection des problèmes mentaux (Li *et al.*, 2022)(Batbaatar *et al.*, 2019), notamment les LSTM (Graves, 2013), les CNN (Zhang & Wallace, 2015) et leurs variantes. Par ailleurs, depuis l'émergence des modèles *Transformer* (Vaswani *et al.*, 2017), le modèle de type BERT (Devlin *et al.*, 2019) est devenu la méthode la plus performante ; ses stratégies d'ajustement (*fine-tuning*) et son mécanisme d'attention renouvellent l'état de l'art en matière de détection des problèmes mentaux (William & Suhartono, 2021)(Ji *et al.*, 2022).

Avec l'avancée de technologies comme ChatGPT², de nouveaux modèles génératifs multilingues émergent, tels que Llama (Touvron *et al.*, 2023), GLM (Du *et al.*, 2022), Qwen (Bai *et al.*, 2023), Claude³, Gemini (Team *et al.*, 2023), etc. Ces modèles sont en cours d'élaboration. Ils sont utilisés dans divers domaines, y compris la détection des problèmes mentaux. Les chercheurs (Wang *et al.*, 2024)(Xu *et al.*, 2024b)(Yang *et al.*, 2023) ont évalué ces modèles génératifs dans le cadre de divers troubles mentaux. Les autres évaluations ont été réalisées en utilisant des données de santé mentale sur les réseaux sociaux chinois (Lan *et al.*, 2024)(Lai *et al.*, 2023).

2.2 Jeux de données concernant la dépression

La spécificité du domaine rend complexe la constitution de jeux de données en santé mentale. Les jeux de données existants (Alhamed *et al.*, 2024)(Shen *et al.*, 2022)(Li *et al.*, 2022)(Bittar *et al.*, 2019)(Fernandes *et al.*, 2018), principalement issues de données médicales telles que les rapports hospitaliers ou les entretiens psychologiques, se révèlent limitées. Elles se concentrent sur les symptômes et ne permettent pas d'analyser les variations d'humeur, ni l'usage quotidien de la langue ou les changements dans les habitudes de parole des personnes à risque dépressif sur une période donnée.

Ainsi, actuellement, les chercheurs utilisent largement les données issues des réseaux sociaux pour réaliser les expériences concernant les alertes précoces et la détection de la dépression ou d'autres troubles psychologiques (Chancellor & De Choudhury, 2020)(Skaik & Inkpen, 2020). Des corpus diversifiés liés aux troubles mentaux ont été compilés dans plusieurs langues, y compris l'anglais (Yates *et al.*, 2017), le chinois (Cai *et al.*, 2023)(Wang *et al.*, 2022), l'espagnol (Leis *et al.*, 2019), le portugais (Mendes & Caseli, 2024), etc.. En matière d'annotation des données, les chercheurs s'appuient généralement sur des normes établies telles que le DSM (Kupfer *et al.*, 2008), le BDI (Beck *et al.*, 1996), et le PHQ (Mitchell *et al.*, 2016). Ils utilisent également des mesures basées sur les auto-déclarations (*self-reported*) des individus concernant la dépression clinique.

3 Constitution des données

Notre jeu de données original provient de SWDD (Cai *et al.*, 2023), qui est en accès libre et nous permet de télécharger librement sur GitHub⁴. Les auteurs originaux ont collecté ces données publiées par les utilisateurs concernés sur les réseaux sociaux via l'API de Sina Weibo. Les données sont

2. ChatGPT : <https://chatgpt.com/>

3. Claude : <https://claude.ai/chats>

4. SWDD : <https://github.com/ethan-nicholas-tsai/SWDD>

organisées par utilisateur, chaque utilisateur ayant plusieurs tweets. Ces données sont annotées par les auteurs en référence au DSM-5⁵, en classant les utilisateurs comme dépressifs (positifs) ou non dépressifs (négatifs), basées sur les symptômes de dépression manifestés dans les tweets. De plus, nous avons contacté les auteurs originaux et obtenu l'autorisation de réutiliser ces données.

La détection des troubles psychologiques ou de la dépression est une recherche pilotée par les données. Les biais dans les normes d'annotation des données peuvent potentiellement causer des erreurs importantes dans les résultats expérimentaux. Après avoir observé les données et leurs annotations, nous avons trouvé difficile de juger la persistance des symptômes de dépression exigée par le DSM-5 sur une période de plus de deux semaines dans les textes sur Internet. Pour assurer la pertinence du corpus, nous avons filtré les données pour ne conserver que les utilisateurs positifs qui se sont autodéclarés dépressifs (*self-reported*), formant ainsi notre nouveau jeu de données. De plus, afin de simuler un environnement internet réel (étant donné que, généralement, même les utilisateurs positifs ne parlent pas exclusivement de symptômes ou d'émotions dépressives, leurs sujets étant variés), et pour explorer les capacités des LLMs à rechercher de longs textes sur les émotions et les symptômes dépressifs, nous avons uniquement conservé les utilisateurs ayant plus de 80 tweets. Pour les utilisateurs de contrôle, nous les avons sélectionnés aléatoirement pour assurer un nombre comparable à celui des utilisateurs positifs. Toutes les procédures de collecte des données sont similaires à celles des travaux de *Xu et al. (2024a)*. Pour respecter les règlements sur la protection des données, nous avons éliminé toutes les informations personnelles permettant potentiellement d'identifier une personne, telles que la description du profil, l'âge, le sexe, l'adresse IP, etc. Nous avons anonymisé le nom de tous les utilisateurs en les remplaçant par des chiffres, comme utilisateur1, utilisateur2, etc.

Concernant le nouveau jeu de données constitué, nous avons utilisé 75 % des données comme ensemble d'entraînement pour entraîner les classifieurs classiques comme références et 25 % des données comme ensemble de test pour évaluer les modèles de classification et les modèles génératifs.

Dataset	Catégorie		Utilisateurs	Messages
SWDD- auto- déclaration	ensemble d'entraîne- ment	Dépressif (Po- sitif)	733	131 187
		Non Dépressif (Négatif)	733	138 743
		Total	1466	269 930
	ensemble de test	Dépressif (Po- sitif)	264	36 429
		Non Dépressif (Négatif)	249	36 890
		Total	513	73 319

TABLE 1 – Jeux de données

5. DSM5 : <https://www.psychiatry.org/psychiatrists/practice/dsm/educational-resources/dsm-5-fact-sheets>

4 Méthodologie

Dans cette partie, nous avons mené des expérimentations comparatives. Tout d'abord, nous avons réalisé des expériences de contrôle en utilisant différentes caractéristiques combinées avec des algorithmes classiques d'apprentissage automatique. Ensuite, nous avons exploré trois modèles génératifs et les différentes stratégies de prompt ainsi que l'ajustement pour chacun de ces modèles.

4.1 Modèles de classification

Nous considérons ces expériences comme un groupe de contrôle pour l'évaluation des modèles génératifs dans la tâche de détection des humeurs dépressives sur les réseaux sociaux chinois. Pour la sélection des caractéristiques, nous nous référons à l'expérimentation de [Xu et al. \(2024a\)](#). Toutes les données d'entrée et les extractions de caractéristiques sont organisées par utilisateur. Nous utilisons comme caractéristiques l'heure de publication des utilisateurs (la proportion de messages postés par un utilisateur entre 23h et 7h le lendemain), les traits sémantiques obtenus par des méthodes textométriques (la fréquence des segments répétés ([Salem, 1986](#)) extraits par la spécificité de Lafon ([Lafon, 1980](#)) liés à la dépression au sein de l'ensemble d'entraînement.), et les plongements sorties du modèle pré-entraîné RoBERTa-chinoise ([Cui et al., 2020](#)) (extraction des plongements de toutes les phrases d'un utilisateur, puis calcul de la moyenne pour obtenir une représentation vectorielle de l'utilisateur.), puis nous combinons ces trois type de caractéristiques et ajoutons des algorithmes classiques de machine learning tels que SVM, XGBoost, Régression Logistique en tant que classifieurs.

4.2 Modèles génératifs

Pour le choix du modèle génératif, nous optons pour trois modèles les plus largement utilisés et librement accessibles dans les recherches en chinoise. Et nous en utilisons des versions allégées, avec moins de paramètres, plus facile à déployer localement.

Llama3-8B-Chinese-Chat⁶ : Ce modèle a été spécifiquement ajusté pour les utilisateurs chinois et anglais grâce à ORPO ([Hong et al., 2024](#)) à partir du modèle Meta-Llama-3-8B-Instruct⁷. Par rapport au modèle original, cette version réduit considérablement les problèmes de réponses en anglais à des questions posées en chinois, ainsi que le mélange de chinois et d'anglais dans les réponses.

Chatglm3-6b⁸ : C'est un modèle basé sur l'architecture spécifique GLM ([Du et al., 2022](#)), entièrement gratuit et en source ouverte, facile à déployer, et qui présente des performances compétitives dans les évaluations de moins de 10 milliards de paramètres sous différents aspects tels que la sémantique, les mathématiques, le raisonnement, le codage, et la connaissance.

Qwen1.5-7B-Chat⁹ : C'est un modèle *Transformer* de type "*decoder-only*" avec activation SwiGLU, RoPE, et attention multi-têtes ([Bai et al., 2023](#)).

6. Llama3-8B-Chinese : <https://huggingface.co/shenzhi-wang/Llama3-8B-Chinese-Chat>

7. Meta-Llama-3-8B : <https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

8. Chatglm3-6b : <https://huggingface.co/THUDM/chatglm3-6b>

9. Qwen1.5-7B-Chat : <https://huggingface.co/Qwen/Qwen1.5-7B-Chat>

4.3 Prompts et méthodes d'analyse explicatives

Dans le domaine des applications des modèles génératifs, l'ingénierie des prompts est cruciale. Nous élaborons des instructions spécifiques selon la tâche pour guider les LLMs, leur permettant de détecter et de classer les états dépressifs, ainsi que d'extraire et de rechercher des informations liées à la dépression. Nous avons systématiquement exploré cinq stratégies de prompt pour la détection des humeurs dépressives, l'analyse de ces états et la recherche d'informations associées. Ci-dessous, les types de prompts sont indiqués. Pour des prompts détaillés, veuillez consulter l'[annexe](#).

1) Prompt Basique (PB) : Formulez directement des instructions simples qui requièrent une réponse par "oui" ou "non".

2) Prompt Ciblé (PC) : Demander au modèle d'évaluer les expressions émotionnelles, attitudes futures, engagements sociaux, et symptômes physiques de l'utilisateur pour identifier les signes de dépression. La réponse doit être "oui" ou "non" et le modèle doit justifier cette réponse.

3) Prompt Enrichi de Connaissances (Supervision faible) (PEC) : Dans les instructions, intégrer la définition et l'explication de la dépression selon l'OMS, en ajoutant des expressions sémantiques et syntaxiques caractéristiques de textes potentiellement dépressifs, issues de la classification DSM-5 et d'autres recherches textuelles sur la dépression. Illustrer avec des exemples précis au niveau lexico-sémantique. Finalement, demander au modèle de répondre par "oui" ou "non" et de fournir une justification.

4) Prompt de Recherche (Retrieval) (PR) : Après avoir numéroté chaque texte des utilisateurs, les instructions demandent au modèle de lire les textes un par un. Si l'utilisateur évoque explicitement son diagnostic de dépression dans le texte, le modèle doit automatiquement le classer comme dépressif (réponse "oui") et extraire tous les textes pertinents pour étayer cette conclusion dans son analyse. Dans les autres cas, on demande au modèle de fournir une réponse "oui" ou "non" tout en récupérant les textes pertinents qui peuvent corroborer sa réponse.

5) Prompt Ciblé de l'Inverse (PCI) : Demander au modèle de se concentrer sur l'évaluation des expressions émotionnelles, des perspectives futures, de l'activité sociale et des symptômes physiques de l'utilisateur pour identifier les signes de dépression. Analyser d'abord le texte, puis, sur la base de cette analyse, déterminer si la réponse est "oui" ou "non".

Nous avons mis en œuvre deux stratégies d'ajustement sur certains modèles, en combinaison avec l'utilisation de certains prompts.

- La première stratégie utilise le dataset D4 ([Yao et al., 2022](#)) pour l'ajustement, qui se compose de transcriptions de dialogues de consultations psychologiques en chinois entre 1339 psychologues et leurs patients, structurés sous forme de questions-réponses entre le médecin et le patient, avec un diagnostic final donné par le médecin. À partir de ces données, nous avons conservé uniquement les paroles des patients, en remplaçant les mots-clés à haute spécificité tels que "médecin", "patient", "dépression modérée", et "dépression sévère". Après l'ajustement de *LoRA* ([Hu et al., 2021](#)), nous associons ce modèle au prompt ciblé.
- La deuxième stratégie consiste à ajuster les modèles en utilisant des réponses générées par GPT-4o¹⁰. Nous choisissons aléatoirement les textes de 50 utilisateurs négatifs et 50 utilisateurs positifs dans l'ensemble d'entraînement. Ensuite, nous utilisons un prompt de recherche (*retrieval*) et ajoutons la référence du DSM-5 pour permettre à GPT-4o de répondre

10. GPT-4o : <https://openai.com/index/hello-gpt-4o/>

par "oui" ou "non" et fournir les analyses ainsi que les textes numérotés (pour justifier les analyses). Nous sélectionnons manuellement 90 réponses correctes et de qualité, dont 45 négatives et 45 positives. Les réponses sélectionnées sont ensuite mélangées pour effectuer une époque d'ajustement supervisée. Pour le modèle après cet ajustement, nous employons le prompt de recherche (*retrieval*) sur l'ensemble de test.

4.4 Résultats des expérimentations

Tous les objets de détection sont basés sur les utilisateurs, c'est-à-dire que les modèles reçoivent les informations textuelles de tous les tweets d'un utilisateur. En plus de réaliser des expériences comparatives sur trois classifieurs classiques, nous avons également appliqué les stratégies de prompt et d'ajustement mentionnées dans la section 4.3 pour effectuer 20 expériences sur 3 modèles génératifs (Chaque expérience a été réalisée 3 fois, et le meilleur résultat de ces tentatives a été sélectionné comme résultat final). Concernant les prédictions générées, nous avons utilisé des règles basées sur des mots-clés et certaines conditions restrictives pour étiqueter chaque utilisateur comme « positif », « négatif » ou « non reconnu ». En raison du problème d'hallucination du modèle, environ 5 à 10 % des prédictions générées « non reconnu » ne pouvaient pas être identifiées et ont été réétiquetées manuellement. Nous avons utilisé la précision, le rappel, le F1 score et l'exactitude pour évaluer les performances. Tous les résultats expérimentaux sont présentés dans le Tableau 2.

Model	True Negative	False Positive	False Negative	True Positive	Precision	Recall	F1 Score	Accuracy
Caractéristiques Combinées (CC) + SVM	229	20	19	245	92.45%	92.8%	92.63%	92.4%
CC + LR	227	22	18	246	91.79%	93.18%	92.48%	92.2%
CC + XGboost	230	19	24	240	91.62%	90.91%	91.78%	91.62%
1) GLM + PB	58	191	10	254	57.08%	96.21%	71.65%	60.82%
2) GLM + PC	4	245	1	263	51.77%	99.62%	68.13%	52.05%
3) GLM + PEC	44	205	8	256	55.53%	96.97%	70.62%	58.48%
4) GLM + PC + FT_D4	233	16	106	158	90.8%	59.8%	72.15%	76.22%
5) GLM + PR	19	230	9	255	52.58%	96.59%	68.09%	53.41%
6) GLM + PCI	197	52	86	178	77.39%	67.42%	72.06%	73.1%
1) Llama3 + PB	181	68	41	223	76.63%	84.47%	80.36%	78.75%
2) Llama3 + PC	178	71	37	227	76.17%	85.98%	80.78%	78.95%
3) Llama3 + PEC	235	14	83	181	92.82%	68.56%	78.87%	81.09%
4) Llama3 + PC + FT_D4	248	1	104	160	99.38%	60.61%	75.29%	79.53%
5) Llama3 + PR	232	17	73	191	91.82%	72.35%	80.93%	82.46%
6) Llama3 + PCI	187	62	43	221	78.09%	83.71%	80.8%	79.53%
7) Llama3 + PR + FT_GPT4o_DSM5	222	27	43	221	89.11%	83.71%	86.33%	86.35%
1) Qwen + PB	237	12	78	186	93.94%	70.45%	80.52%	82.46%
2) Qwen + PC	211	38	23	241	86.38%	91.29%	88.77%	88.1%
3) Qwen + PEC	207	42	34	230	84.56%	87.12%	85.82%	85.19%
4) Qwen + PC + FT_D4	243	6	97	167	96.53%	63.26%	76.43%	79.92%
5) Qwen + PR	217	32	32	232	87.88%	87.8%	87.8%	87.5%
6) Qwen + PCI	187	62	38	226	78.47%	85.61%	81.88%	80.51%
7) Qwen + PR + FT_GPT4o_DSM5	205	44	12	252	85.13%	95.45%	90%	89.08%

TABLE 2 – Résultats des expérimentations

5 Analyse des résultats

5.1 Détection de dépression

Configuration et Évaluation Progressive des Modèles à Base de Prompts : Dans le cadre de notre étude sur les LLMs, nous avons adopté une approche méthodique et hiérarchisée pour la configuration et l'évaluation des prompts. Initialement, 3 modèles, à savoir GLM, Llama, et Qwen, ont été soumis à des tests utilisant un prompt basique. Il en ressort que GLM affiche des performances inférieures, alors que Llama et Qwen maintiennent un niveau plus élevé de performance.

Phase Deux : Amélioration par Prompt Ciblé Durant la deuxième phase d'évaluation, l'utilisation d'un prompt ciblé centré sur les indicateurs de dépression a prouvé être une stratégie efficace pour améliorer la capacité des modèles à identifier les émotions dans les analyses générées. Toutefois, bien que Qwen et Llama semblent mieux équilibrés, ils n'ont pas montré d'amélioration significative des résultats. Et GLM continue d'enregistrer des scores métriques inférieurs. Cette condition peut être attribuée aux différences structurelles de GLM, qui utilise une méthode d'inférence autoregressive pour le remplissage de blancs (Du et al., 2022).

Troisième Phase : Introduction du Prompt Enrichi de Connaissances Dans la troisième étape, nous avons introduit le Prompt Enrichi de Connaissances. En fournissant des définitions, des catégories sémantiques de la dépression, et des exemples au niveau lexical, nous avons observé une diminution du rappel pour les trois modèles, avec Llama enregistrant la plus grande baisse. Ce résultat suggère que l'intégration de connaissances spécifiques rend les modèles plus prudents dans leurs jugements lors des prédictions positives.

Quatrième Phase : Prompt Ciblé et Adjustment (*Fine-tuning*) avec le Jeu de Données D4 La quatrième partie de notre étude a réutilisé le prompt ciblé avec les modèles ajustés par le jeu de données D4, qui comprend des interactions de psychologues sur la dépression. Cette méthode a entraîné une augmentation de la précision pour les trois modèles. Cependant, le rappel a baissé, ce qui pourrait être attribué à la nature prudente des modèles face aux données médicales et au potentiel risque de surajustement.

Cinquième Phase : Emploi du Prompt de Recherche Dans la cinquième partie, nous avons mis en place un prompt de recherche spécialement conçu pour la récupération d'informations liées à la dépression. Ce prompt a aidé à fournir un soutien textuel pendant les prédictions des modèles. Les résultats montrent que, contrairement à GLM, Llama et Qwen ont bien performé avec cette stratégie sur la détection et la recherche des textes.

Sixième Phase : Adaptation Spécifique pour le GLM La sixième phase a été spécifiquement conçue pour s'adapter à la structure unique du GLM. Étant donné que ce modèle diffère des modèles de décodeurs traditionnels, en pré-entraînement, il prédit le contenu de plusieurs spans extraits aléatoirement d'une phrase. Nous avons demandé au modèle de suivre les instructions du prompt pour une analyse minutieuse avant de générer des réponses. Cette approche a mené à une amélioration notable des performances de GLM.

Septième Phase : Adjustment (*Fine-tuning*) Spécifique pour Llama et Qwen Au cours de cette phase de tests, seuls les modèles Llama et Qwen ont été évalués, excluant le GLM en raison de ses performances atypiques précédemment observées. Une approche similaire à la distillation des grands modèles vers les petits a été appliquée (Yu et al., 2023). Les modèles sont d'abord ajustés sur des

réponses générées par GPT-4, en utilisant des prompts de recherche auxquels sont ajoutés des critères du DSM-5. Cette méthode renforce le comportement de récupération de texte numéroté, qui est moins observé dans le petit modèle. Après cet ajustement, les modèles ont été évalués à l'aide de prompts de recherche. Cette méthode a permis d'améliorer significativement les scores F1 et la précision de Llama et Qwen, qui ont ainsi obtenu les meilleurs résultats parmi tous les tests effectués sur les modèles génératifs.

Conclusion : Les modèles supervisés basés sur des caractéristiques sont généralement plus performants que les modèles génératifs dans la détection de dépression. Toutefois, grâce à nos stratégies de prompt et d'ajustement orientées vers la recherche d'informations, il a été possible de rapprocher les performances de certains modèles génératifs de celles des modèles de classification. Les modèles que nous avons utilisés possèdent moins de 8 milliards de paramètres ; par conséquent, l'utilisation de modèles plus grands, avec davantage de paramètres, pourrait potentiellement améliorer encore les résultats. En outre, bien que les modèles génératifs ne soient pas initialement conçus pour des tâches de classification, ils offrent des avantages tels que la capacité de générer des analyses prédictives et d'extraire des phrases clés. Avec ces tests, nous pouvons supposer que ces caractéristiques constituent potentiellement une raison majeure pour laquelle les chercheurs les sélectionnent pour les tâches de détection de problèmes de santé mentale. Par ailleurs, nos différents essais de prompt visent également à explorer l'impact de diverses stratégies sur le texte généré.

5.2 Recherche d'information (*Information Retrieval*)

Pour toutes les stratégies de prompt, à l'exception du prompt basique, nous exigeons que les modèles fournissent une analyse détaillée et progressive de leurs prédictions relatives aux humeurs dépressives. D'après les réponses générées, nous constatons que la dernière stratégie de prompt (PR + FT_GPT4o_DSM5) améliore le score F1 et l'exactitude pour les modèles génératifs, et la capacité de récupération et d'analyse des textes sont probablement élevées. De plus, cette stratégie réduit nettement les hallucinations textuelles et atteint un niveau élevé de cohérence linguistique. Nous sélectionnons au hasard les réponses (Voir [Table 3](#) et [Table 4](#) en Annexes) de Qwen à deux utilisateurs (la véritable étiquette de ces deux utilisateurs est "utilisateurs présentant des symptômes de dépression") sous diverses stratégies de prompt. Nous les considérons comme exemples pour analyser l'amélioration des capacités d'extraction d'informations et de génération de texte des modèles génératifs après l'intégration de prompts de recherche.

Bien qu'elles incluent plusieurs analyses, les réponses guidées par le prompt ciblé utilisent un langage général et les orientations sont assez vagues, sans fournir d'analyse spécifiques à l'appui. De plus, bien que le texte généré extrait certaines phrases des textes, nous avons observé dans l'analyse des données utilisateurs originales que ces textes ou descriptions extraits apparaissent souvent dans les 20% du début ou de la fin du texte. De plus, des erreurs ont été commises dans la prédiction des émotions dépressives. Nous pouvons donc déduire que si les indicateurs de dépression sont moins présents au milieu d'un texte long, le modèle peut ne pas être capable de les détecter directement, ce qui entraîne des erreurs de prédiction. En ce qui concerne le Prompt enrichi des connaissances, les réponses sont plus courtes et il n'y a pas eu de récupération d'informations, probablement sous l'influence des données d'ajustement. De même, les thèmes extraits et les descriptions fournies, comme avec le prompt ciblé, proviennent des 20% du début ou de la fin du texte. Les résultats montrent que le Prompt de recherche résout bien le problème des indicateurs de dépression non détectables au milieu de textes longs. De plus, l'utilisation d'une stratégie de l'ajustement similaire à la distillation de

grands modèles, combinée avec un prompt de recherche, peut considérablement améliorer l'efficacité de ce prompt. Dans les réponses générées, une haute précision est maintenue, et des descriptions des indicateurs potentiels de dépression sont fournies avec les numéros de texte correspondants. Après une vérification manuelle des textes entrés, la correspondance entre les numéros et les contenus extraits est confirmée, ce qui illustre davantage l'efficacité de notre stratégie.

6 Conclusions et limitations

Dans cette étude, nous avons utilisé 3 modèles génératifs chinois open source et gratuits pour évaluer leur capacité à détecter les émotions dépressives dans les données internet. Nous avons testé non seulement les modèles originaux mais aussi des versions ajustées avec deux ensembles de données spécialisées, en utilisant 5 stratégies de prompts différentes. L'objectif était d'explorer la performance potentielle des modèles génératifs chinois dans la détection des émotions dépressives, ainsi que leur capacité à analyser des textes et à capter des informations clés sur les indicateurs de dépression en utilisant différentes stratégies de prompts. Cette approche nous a permis de visualiser les résultats des prédictions et de fournir une évaluation explicative, palliant ainsi les lacunes des modèles supervisés qui, bien qu'ils classifient, ne fournissent pas d'explications. Les résultats montrent que la septième stratégie (utilisant les critères du DSM-5 pour le filtrage préalable de l'ensemble de formation par GPT-4o, suivi d'un affinage et l'utilisation d'un prompt de recherche) est la plus efficace. Cette stratégie a non seulement atteint une performance de classification comparable à celle des modèles d'apprentissage supervisé, mais a également montré des résultats significatifs dans la capture et l'explication des indicateurs de dépression dans les textes longs.

Notre recherche comporte également certaines limitations. La complexité de l'annotation des données sur les plateformes sociales relatives aux états dépressifs fait que les ensembles de données en chinois de ce type sont rares. De plus, dans nos tests, pour faciliter le déploiement et les tests locaux, nous utilisons des modèles génératifs à faible nombre de paramètres. Nous n'avons également pas encore développé un système d'évaluation complet pour la recherche d'informations et la qualité des textes générés. Dans nos travaux futurs, nous envisageons de collecter davantage de données pour vérification, ou d'appliquer la même stratégie à des ensembles de données multilingues pour tester leur efficacité. Nous prévoyons aussi d'expérimenter avec des modèles plus grands, combinant différentes stratégies de prompt et les comparant aux performances des modèles supervisés. Finalement, nous projetons de mettre en place un système d'évaluation par des médecins pour évaluer la qualité des textes générés, la pertinence de la recherche d'informations et la précision des analyses prédictives.

Références

- ALHAMED F., IVE J. & SPECIA L. (2024). Classifying social media users before and after depression diagnosis via their language usage : A dataset and study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 3250–3260.
- BAI J., BAI S., CHU Y., CUI Z., DANG K., DENG X. *et al.* (2023). Qwen technical report. *arXiv preprint arXiv :2309.16609*.
- BATBAATAR E., LI M. & RYU K. H. (2019). Semantic-emotion neural network for emotion recognition from text. *IEEE access*, **7**, 111866–111878.
- BECK A. T., STEER R. A., BALL R. & RANIERI W. F. (1996). Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *Journal of personality assessment*, **67**(3), 588–597.
- BITTAR A., VELUPILLAI S., ROBERTS A. & DUTTA R. (2019). Text classification to inform suicide risk assessment in electronic health records. In *MedInfo*, p. 40–44.
- BREIMAN L. (2001). Random forests. *Machine learning*, **45**, 5–32.
- CAI Y., WANG H., YE H., JIN Y. & GAO W. (2023). Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*, **217**, 119538.
- CHANCELLOR S. & DE CHOUDHURY M. (2020). Methods in predictive techniques for mental health status on social media : a critical review. *NPJ digital medicine*, **3**(1), 43.
- CHEN T. & GUESTRIN C. (2016). Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, p. 785–794.
- CORRIGAN P. W. & WATSON A. C. (2002). Understanding the impact of stigma on people with mental illness. *World psychiatry*, **1**(1), 16.
- CORTES C. & VAPNIK V. (1995). Support-vector networks. *Machine learning*, **20**, 273–297.
- CUI Y., CHE W., LIU T., QIN B., WANG S. & HU G. (2020). Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv :2004.13922*.
- DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4171–4186.
- DU Z., QIAN Y., LIU X., DING M., QIAN Y., LIU X., QIU J., YANG Z. & TANG J. (2022). Glm : General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 320–335, Dublin, Ireland : Association for Computational Linguistics.
- FERNANDES A. C., DUTTA R., VELUPILLAI S., SANYAL J., STEWART R. & CHANDRAN D. (2018). Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Scientific reports*, **8**(1), 7426.
- GRAVES A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv :1308.0850*.
- HONG J., LEE N. & THORNE J. (2024). Reference-free monolithic preference optimization with odds ratio. *arXiv preprint arXiv :2403.07691*.
- HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2021). Lora : Low-rank adaptation of large language models. *arXiv preprint arXiv :2106.09685*.

- JI S., ZHANG T., ANSARI L., FU J., TIWARI P. & CAMBRIA E. (2022). Mentalbert : Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 7184–7190, Marseille, France : European Language Resources Association.
- JI Z., LEE N., FRIESKE R., YU T., SU D., XU Y., ISHII E., BANG Y. J., MADOTTO A. & FUNG P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, **55**(12), 1–38.
- KUPFER D. J., FIRST M. B. & REGIER D. A. (2008). A research agenda for dsm v.
- LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, **1**(1), 127–165.
- LAI T., SHI Y., DU Z., WU J., FU K., DOU Y. & WANG Z. (2023). Psy-llm : Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv :2307.11991*.
- LAN X., CHENG Y., SHENG L., GAO C. & LI Y. (2024). Depression detection on social media with large language models. *arXiv preprint arXiv :2403.10750*.
- LE GLAZ A., HARALAMBOUS Y., KIM-DUFOR D.-H., LENCA P., BILLOT R., RYAN T. C., MARSH J., DEVYLDER J., WALTER M., BERROUIGUET S. *et al.* (2021). Machine learning and natural language processing in mental health : systematic review. *Journal of medical Internet research*, **23**(5), e15708.
- LEIS A., RONZANO F., MAYER M. A., FURLONG L. I. & SANZ F. (2019). Detecting signs of depression in tweets in spanish : behavioral and linguistic analysis. *Journal of medical Internet research*, **21**(6), e14199.
- LI M., XU H., LIU W. & LIU J. (2022). Bidirectional lstm and attention for depression detection on clinical interview transcripts. In *2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN)*, p. 638–643 : IEEE.
- MENDES A. R. & CASELI H. (2024). Identifying fine-grained depression signs in social media posts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 8594–8604.
- MITCHELL A. J., YADEGARFAR M., GILL J. & STUBBS B. (2016). Case finding and screening clinical utility of the patient health questionnaire (phq-9 and phq-2) for depression in primary care : a diagnostic meta-analysis of 40 studies. *BJPsych open*, **2**(2), 127–138.
- PETERSON L. E. (2009). K-nearest neighbor. *Scholarpedia*, **4**(2), 1883.
- SALEM A. (1986). Segments répétés et analyse statistique des données textuelles. *Histoire & mesure*, p. 5–28.
- SALTON G. & BUCKLEY C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, **24**(5), 513–523.
- SHEN Y., YANG H. & LIN L. (2022). Automatic depression detection : An emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6247–6251 : IEEE.
- SKAIK R. & INKPEN D. (2020). Using social media for mental health surveillance : a review. *ACM Computing Surveys (CSUR)*, **53**(6), 1–31.
- TAUSCZIK Y. R. & PENNEBAKER J. W. (2010). The psychological meaning of words : Liwc and computerized text analysis methods. *Journal of language and social psychology*, **29**(1), 24–54.
- TEAM G., ANIL R., BORGEAUD S., WU Y., ALAYRAC J.-B., YU J., SORICUT R., SCHALKWYK J., DAI A. M., HAUTH A. *et al.* (2023). Gemini : a family of highly capable multimodal models. *arXiv preprint arXiv :2312.11805*.

- TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F. *et al.* (2023). Llama : Open and efficient foundation language models. *arXiv preprint arXiv :2302.13971*.
- VALETTE M. (2016). Analyse statistique des données textuelles et traitement automatique des langues. une étude comparée. In D. MAYAFFRE, C. POUDAT, L. VANNI, V. MAGRI & P. FOLLETTE, Éd., *Proceedings of 13th International Conference on Statistical Analysis of Textual Data*, volume II, p. 697–706, Nice, France : 7-10 Juin 2016.
- VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.
- WANG Y., INKPEN D. & GAMAARACHCHIGE P. K. (2024). Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, p. 108–126.
- WANG Y., WANG Z., LI C., ZHANG Y. & WANG H. (2022). Online social network individual depression detection using a multitask heterogenous modality fusion approach. *Information Sciences*, **609**, 727–749.
- WILLIAM D. & SUHARTONO D. (2021). Text-based depression detection on social media posts : A systematic literature review. *Procedia Computer Science*, **179**, 582–589.
- XU J., MAGISTRY P. & VALETTE M. (2024a). Détection des humeurs dépressives sur les réseaux sociaux chinois à partir d’une combinaison de plongements lexicaux et de méthodes textométriques. In *JADT 2024 - 17es Journées internationales d’Analyse statistique des Données Textuelles*.
- XU X., YAO B., DONG Y., GABRIEL S., YU H., HENDLER J., GHASSEMI M., DEY A. K. & WANG D. (2024b). Mental-llm : Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, **8**(1), 1–32.
- YANG K., JI S., ZHANG T., XIE Q., KUANG Z. & ANANIADOU S. (2023). Towards interpretable mental health analysis with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- YAO B., SHI C., ZOU L., DAI L., WU M., CHEN L., WANG Z. & YU K. (2022). D4 : A chinese dialogue dataset for depression-diagnosis-oriented chat. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 2438–2459, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics.
- YATES A., COHAN A. & GOHARIAN N. (2017). Depression and self-harm risk assessment in online forums. In *Proceedings of the 14th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 2968–2978, Copenhagen, Denmark.
- YU H.-F., HUANG F.-L. & LIN C.-J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, **85**, 41–75.
- YU X., PENG B., GALLEY M., GAO J. & YU Z. (2023). Teaching language models to self-improve through interactive demonstrations. *arXiv preprint arXiv :2310.13522*.
- ZHANG Y. & WALLACE B. (2015). A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv :1510.03820*.

Annexes

1.Prompts :

1)Prompt Basique (PB) :

Prompt en chinois :

请仔细阅读以下多条文本,然后回答问题: 基于这些文本, 这个用户是否存在抑郁情绪? 用“是”或者“否”回答

重要

你必须用以下格式回答问题

答案: “是”或“否”

Traduction en français :

Veillez lire attentivement les multiples textes suivants, puis répondre à la question : Basé sur ces textes, cet utilisateur présente-t-il des symptômes de dépression ?

Répondez par "Oui" ou "Non".

Important

Vous devez répondre à la question en utilisant le format suivant : Réponse : "Oui" ou "Non"

2)Prompt Ciblé (PC) :

Prompt en chinois :

请仔细阅读以下多条文本, 关注用户的情绪表达、对未来的态度、社交参与度、兴趣水平以及任何关于身体症状或睡眠/食欲变化、自我价值、个人感受的表述。回答问题: 基于这些文本, 这个用户是否存在抑郁情绪? 用“是”或者“否”回答, 再给出你一步一步的分析过程。

重要

你必须用以下格式回答问题

答案: “是”或“否”

分析: 你一步一步的分析过程, 解释你为什么给出这个答案。

Traduction en français :

Veillez lire attentivement les différents textes ci-dessous, en prêtant attention à l'expression émotionnelle de l'utilisateur, son attitude envers l'avenir, son degré de participation sociale, son niveau d'intérêt, ainsi que toute mention de symptômes physiques ou de changements dans le sommeil/l'appétit, l'estime de soi, et les sentiments personnels. Répondez à la question : Sur la base de ces textes, cet utilisateur présente-t-il des symptômes de dépression ? Répondez par "oui" ou "non", puis donnez votre analyse étape par étape.

Important

Vous devez répondre à la question en utilisant le format suivant :

Réponse : "Oui" ou "Non"

Analyse : Votre processus d'analyse étape par étape, expliquant pourquoi vous avez donné cette réponse.

3) Prompt Enrichi de Connaissances (Supervision faible) (PEC) :

Prompt en chinois :

请仔细阅读以下多条推文，关注用户的情绪表达、对未来的态度、社交参与度、兴趣水平以及任何关于身体症状或睡眠/食欲变化、自我价值、个人感受的表述。回答问题：基于这些推文，这个用户是否存在抑郁情绪？用“是”或者“否”回答，再给出你一步一步的分析过程。

你可以利用以下信息帮助你判断

抑郁发作有不同的类型，包括：

单次发作的抑郁障碍，意指病人第一次也是唯一一次发作；复发性抑郁障碍，这意味着病人有至少两次抑郁发作史；以及双相情感障碍，意味着抑郁发作与躁狂症状交替出现，包括情绪高涨或易怒、活动增加或精力旺盛，以及其他症状，如健谈、思绪奔涌、自尊心增强、睡眠需求减少、分心和冲动鲁莽行为。

一般来说以下类型的词或者表达会在抑郁病人微博文本中出现：

1. 负面情绪。例如：（想哭，累了，烦，不开心，抑郁，痛苦，难受，害怕，崩溃，焦虑，孤独，压抑，怕，眼泪，悲伤，恐惧，低落，伤心）
2. 负面心理状态：悲观，自我价值缺失，负罪感，自杀自残想法。例如：（讨厌自己，我的错，自杀，死亡，自残，折磨，死去，解脱，跳楼，遗书，想死，绝望，煎熬）
3. 临床描述：食欲变化，睡眠障碍，身体异常和难受。例如：（厌食，失眠，噩梦，累，晕，痛，病，疼，恶心，头疼，想吐，幻听）
4. 社会关系变化：婚姻感情失败，家庭破碎，工作失败，学业失败。例如：（分手，离婚，原生家庭，休学，裁员，退学）
5. 医疗相关：抗抑郁药物名称，疾病名，医疗专业名词。例如：（舍曲林，文拉法辛，抗抑郁药，抑郁症，焦虑症，躁郁症，双相情感障碍，住院，心理医生，心理咨询，医院）
6. 第一人称代词“我”+否定词或者消极表达。例如：（我不知道，我害怕，我想死，我不配，我好累）

重要

你必须用以下格式回答问题

答案：“是”或“否”

分析：你一步一步的分析

Traduction en français :

""Veuillez lire attentivement les différents tweets ci-dessous, en prêtant attention à l'expression des émotions de l'utilisateur, à son attitude envers l'avenir, à son niveau de participation sociale, à son intérêt, ainsi qu'à toute mention de symptômes physiques ou de changements dans le sommeil/l'appétit, l'estime de soi et les sentiments personnels. Répondez à la question : Sur la base de ces tweets, cet

utilisateur présente-t-il des symptômes de dépression ? Répondez par "oui" ou "non", puis donnez votre analyse étape par étape.

Vous pouvez utiliser les informations suivantes pour vous aider à juger

Il existe différents types de crises dépressives, y compris :

Trouble dépressif à épisode unique, signifiant que le patient a une seule occurrence de crise ; Trouble dépressif récurrent, ce qui signifie que le patient a eu au moins deux crises ; et Trouble bipolaire, signifiant que les crises dépressives alternent avec des symptômes de manie, incluant une humeur élevée ou irritabilité, une augmentation de l'activité ou de l'énergie, et d'autres symptômes tels que la logorrhée, le flot de pensées, une estime de soi accrue, une diminution du besoin de sommeil, la distractibilité et un comportement impulsif et téméraire.

Généralement, les types de mots ou expressions suivants apparaîtront dans les textes Weibo des patients dépressifs :

Émotions négatives. Par exemple : (envie de pleurer, fatigué, agacé, triste, déprimé, souffrance, inconfort, peur, effondrement, anxiété, solitude, oppression, peur, larmes, tristesse, terreur, abattement, chagrin)

États psychologiques négatifs : pessimisme, manque de valeur personnelle, culpabilité, pensées de suicide ou d'automutilation. Par exemple : (je me déteste, c'est ma faute, suicide, mort, automutilation, torture, décès, libération, sauter d'un bâtiment, lettre d'adieu, envie de mourir, désespoir, agonie)

Descriptions cliniques : changements d'appétit, troubles du sommeil, anomalies et malaises corporels. Par exemple : (anorexie, insomnie, cauchemars, fatigue, vertiges, douleur, maladie, douleur, nausée, mal de tête, envie de vomir, hallucinations auditives)

Changements dans les relations sociales : échecs amoureux, familles brisées, échecs professionnels, échecs scolaires. Par exemple : (rupture, divorce, famille d'origine, abandon scolaire, licenciement, déscolarisation)

Médicaux : noms de médicaments antidépresseurs, noms de maladies, termes professionnels médicaux. Par exemple : (sertraline, venlafaxine, médicaments antidépresseurs, dépression, anxiété, trouble bipolaire, hospitalisation, psychiatre, conseil psychologique, hôpital)

Premier pronom personnel "je" + mot de négation ou expression négative. Par exemple : (je ne sais pas, j'ai peur, je veux mourir, je ne le mérite pas, je suis très fatigué)

Important

Vous devez répondre à la question en utilisant le format suivant

Réponse : "Oui" ou "Non"

Analyse : Votre processus d'analyse étape par étape

4) Prompt de Recherche (Retrieval) (PR) :

Prompt en chinois :

请逐条阅读以下用户的文本，关注用户的情绪表达、对未来的态度、社交参与度、兴趣水平以及任何关于身体症状或睡眠/食欲变化、自我价值、个人感受的表述。回答问题：基于这些文本，这个用户是否存在抑郁情绪？用“是”或者“否”回答，再给出你一步一步的分析过程，指出哪条或哪几条文本让你得出这个回答。

重要

如果用户在文本中明确提到自己得到了抑郁的诊断，请直接判断为“是”，并在分析中指出相关文本。

你必须用以下格式回答问题

答案：“是”或“否”

分析：你一步一步的分析过程，指出哪条或哪几条文本让你得出这个回答。

Traduction en français :

Veillez lire attentivement chaque texte de l'utilisateur, en prêtant attention à l'expression des émotions, à l'attitude envers l'avenir, au niveau de participation sociale, au niveau d'intérêt, ainsi qu'à toute mention de symptômes physiques ou de changements dans le sommeil/l'appétit, l'estime de soi et les sentiments personnels. Répondez à la question : Sur la base de ces textes, cet utilisateur présente-t-il des symptômes de dépression ? Répondez par "oui" ou "non", puis donnez votre analyse étape par étape, en indiquant quel texte ou quels textes vous ont mené à cette réponse.

Important

Si l'utilisateur mentionne explicitement dans un texte qu'il a été diagnostiqué avec une dépression, jugez directement "oui" et indiquez dans votre analyse le texte concerné.

Vous devez répondre à la question en utilisant le format suivant :

Réponse : "Oui" ou "Non"

Analyse : Votre processus d'analyse étape par étape, en indiquant quel texte ou quels textes vous ont mené à cette réponse.

5) Prompt Ciblé de l'Inverse (PCI) :

Prompt en chinois :

请仔细阅读以下多条文本，关注用户的情绪表达、对未来的态度、社交参与度、兴趣水平以及任何关于身体症状或睡眠/食欲变化、自我价值、个人感受的表述。基于这些文本，进行一步一步的分析，然后回答问题：这个用户是否存在抑郁情绪？用“是”或者“否”回答，再给出你一步一步的分析过程。

重要

你必须用以下格式回答问题

分析：你一步一步的分析过程，解释你为什么给出这个答案。

答案：“是”或“否”

Traduction en français :

Veillez lire attentivement les textes suivants, en prêtant attention à l'expression des émotions de l'utilisateur, à son attitude envers l'avenir, à son niveau de participation sociale, à son niveau d'intérêt, ainsi qu'à toute mention de symptômes physiques ou de changements dans le sommeil/l'appétit, l'estime de soi et les sentiments personnels. Sur la base de ces textes, procédez à une analyse étape par étape, puis répondez à la question : Cet utilisateur présente-t-il des symptômes de dépression ? Répondez par "oui" ou "non", et fournissez votre processus d'analyse étape par étape.

Important

Vous devez répondre à la question en utilisant le format suivant :

Analyse : Votre processus d'analyse étape par étape, expliquant pourquoi vous avez donné cette réponse.

Réponse : "Oui" ou "Non"

6) Prompt de Filtrage des Données de *fine-tuning* avec GPT-4o pour le Septième Test

Prompt en chinois :

请逐条阅读以下用户的推文，关注用户的情绪表达、对未来的态度、社交参与度、兴趣水平以及任何关于身体症状或睡眠/食欲变化、自我价值、个人感受的表述。回答问题：基于这些推文以及DSM-5诊断标准，这个用户是否存在抑郁情绪？用“是”或者“否”回答，再给出你一步一步的分析过程，指出哪条或哪几条文本让你得出这个回答。

重要

如果用户在推文中明确提到自己觉得患上了抑郁，请直接判断为“是”，并在分析中指出相关文本。

你必须用以下格式回答问题

答案：“是”或“否”

分析：你一步一步的分析过程，指出哪条或哪几条文本让你得出这个回答。（你只需要提供文本的编号，不需要提供具体的文本内容）

Traduction en français :

Veillez lire attentivement chaque tweet de l'utilisateur, en portant une attention particulière à l'expression de ses émotions, à son attitude envers l'avenir, à son niveau de participation sociale, à son niveau d'intérêt, ainsi qu'à toute mention de symptômes physiques ou de changements dans le sommeil/l'appétit, l'estime de soi et les sentiments personnels. Répondez à la question : Basé sur ces tweets et les critères de diagnostic du DSM-5, cet utilisateur présente-t-il des symptômes de dépression ? Répondez par "oui" ou "non", et fournissez votre processus d'analyse étape par étape, en indiquant quels tweets vous ont mené à cette conclusion.

Important

Si l'utilisateur mentionne explicitement dans un tweet qu'il pense souffrir de dépression, jugez immédiatement "oui" et indiquez dans votre analyse le tweet concerné.

Vous devez répondre à la question en utilisant le format suivant :

Réponse : "Oui" ou "Non"

Analyse : Votre processus d'analyse étape par étape, en indiquant le numéro des tweets qui vous ont amené à cette conclusion (vous n'avez pas besoin de fournir le contenu spécifique des tweets).

2. Exemples de réponses en chinois de Qwen et leurs traductions en français pour la section 5.2 :

Prompt	PC	PEC	PR	PR_FT_GPT4o_DSM5
user_3150	<p>答案：“否”分析：根据推文内容，用户的情绪表达多样，既有积极的如分享音乐、期待明天、享受美食和自然美景，也有消极的比如失眠、身体不适、情绪困扰、对未来的担忧等。但没有直接提及抑郁症状，如持续的悲伤、失去兴趣、自我价值感下降、严重的睡眠或食欲问题等。虽然提到了一些负面感受，但整体上用户并未表现出明显的抑郁情绪，而是通过幽默和自嘲来应对生活中的困难。因此，基于这些推文，不能断定这个用户存在抑郁症。</p>	<p>无抑郁症状：近日，发现一条消极微博，但未见连续多条，且该用户在其他方面表现得较为积极，故暂不构成抑郁症诊断。建议继续关注。</p>	<p>答案：“是”分析：用户在文本36中明确提到“我已经是一个重度抑郁症患者了”，这表明存在抑郁情绪。这条文本表达的是负面的自我评价和生活状态，与抑郁症的症状相符。其他文本虽然提及了一些日常生活的细节、情绪变化以及对未来的期待，但没有直接提及抑郁。因此，基于文本36，可以判断这个用户存在抑郁情绪。</p>	<p>答案：“是”分析：1. 文本36中明确提到“我已经是一个重度抑郁症患者了”，这是用户自述的诊断信息。2. 从文本40和44可以看出用户多次提到自己快要自杀、距离越来越近等内容，这表明有严重自杀倾向。3. 文本54中提到“不管熬夜到多晚都要戴面膜”，这是一种应对压力的行为，但同时也表现出睡眠问题和焦虑情绪。这些文本都指向了抑郁症的诊断特征，因此判断为“是”。</p>
user_680	<p>答案：“否”分析：这个用户的推文主要包含了多种情绪表达，但并没有直接显示抑郁情绪的特征。虽然提到了一些生活中的困扰和挑战（如减肥、情感问题、没有共同话题等），但这些内容更多的是日常生活琐事，并非抑郁症的典型症状。1. 用户提到对减肥和社交关系的抱怨，可能反映了压力或者不满足，但并不强烈到显示抑郁的程度。2. 通过“珍惜”、“支持”这类积极情绪词汇，可以看出用户在关注自己关心的人或事物，情感相对稳定。3. “期待新戏”、“喜欢孩子”等表明用户对生活保持乐观和兴趣，也不太符合抑郁症的消极态度。4. 尽管部分推文中提到了自我价值感低落（如“拖了社会的后腿”），但整体上并没有持续的自我贬低或否定。5. 没有提到明显的睡眠/食欲变化、身体症状或其他生理不适，这些通常是抑郁情绪的表现。综上所述，虽然推文中含有挑战和困扰，但缺乏深度的抑郁情绪表达，因此不能断定这个用户存在抑郁情绪。</p>	<p>否无抑郁症状：10岁就开始担心吃辣条长胖，可见有些焦虑。翻新闻感叹生命脆弱，可见存在悲伤情绪。并无其他抑郁相关症状。建议合理饮食，释放对辣条的思念，放松心情，多与朋友交流。</p>	<p>答案：“否”分析：用户的文本中并未直接提到抑郁情绪。在文本1、23、46、50、57、68和72中，用户提到了一些负面情绪或生活中的困扰，如减肥、担心落后社会、失眠、没有共同话题等，但这些并不足以明确表明存在抑郁情绪。例如，用户提到的“饿了么天天领红包”可能是在表达日常生活的乐趣，并非抑郁症的症状。在文本72中，“怎么突然就抑郁了呢我该怎么办”是用户在询问自己为何会有这样的感受，但并未表现出持续的抑郁状态。其他多数文本内容更偏向于日常生活、情感互动或者对他人行为的评论，没有直接指向抑郁情绪。因此，综合来看，没有足够的证据表明这个用户存在抑郁情绪。</p>	<p>答案：“是”分析：1. 文本22中的“突然没反应过来，刚才服务器好像一度瘫痪了”，文本30的“31岁生日快乐！李三岁长大也可爱”以及文本36的“佛系粉丝。不抢不抢。迟来的祝福福哥哥，生日快乐”中包含了用户对于自己情绪状态的描述和表达。2. 文本72中的“怎么突然就抑郁了呢我该怎么办”，文本73中的“怎么突然就抑郁了呢我该怎么办”，连续两次提到“抑郁”，这是明显的抑郁情绪症状的表述。3. 在这些文本中，用户没有明确提及自己是否得到了诊断，但他们的描述和表达符合抑郁症的一些常见症状，如持续的情绪低落、失去兴趣或快乐感、感到无助或无价值等。综合以上分析，这个用户存在抑郁情绪。</p>

TABLE 3 – Exemples de réponses

Prompt	PC	PEC	PR	PR_FT_GPT4o_DSM5
user_3150	Réponse : 'Non' Analyse : Selon le contenu des tweets, les expressions émotionnelles de l'utilisateur sont variées, à la fois positives comme partager de la musique, anticiper demain, profiter de la nourriture et des beaux paysages naturels, et négatives comme l'insomnie, les maux physiques, les troubles émotionnels, les inquiétudes pour l'avenir, etc. Mais il n'est pas directement fait mention de symptômes dépressifs, comme une tristesse persistante, une perte d'intérêt, une baisse de l'estime de soi, des problèmes graves de sommeil ou d'appétit, etc. Bien que certains sentiments négatifs soient mentionnés, dans l'ensemble, l'utilisateur ne manifeste pas d'humeur dépressive évidente, mais fait plutôt face aux difficultés de la vie avec humour et autodérision. Par conséquent, sur la base de ces tweets, on ne peut pas affirmer que cet utilisateur souffre de dépression.	Pas de symptômes dépressifs : Récemment, un tweet négatif a été découvert, mais pas vu plusieurs tweets consécutifs, et cet utilisateur s'est comporté de manière assez positive dans d'autres domaines, donc ne constitue pas encore un diagnostic de dépression pour le moment. Il est recommandé de continuer à suivre.	"Oui" Analyse : L'utilisateur mentionne explicitement dans le texte 36 "Je suis déjà un patient souffrant de dépression sévère", ce qui indique la présence d'humeur dépressive. Ce texte exprime une auto-évaluation négative et un état de vie, en accord avec les symptômes de la dépression. Bien que les autres textes mentionnent certains détails de la vie quotidienne, des changements d'humeur et des attentes pour l'avenir, ils ne font pas directement référence à la dépression. Par conséquent, sur la base du texte 36, on peut juger que cet utilisateur présente une humeur dépressive.	"Oui" Analyse : 1. Dans le texte 36, il est explicitement mentionné "Je suis déjà un patient souffrant de dépression sévère", ce qui est une information de diagnostic décrite par l'utilisateur lui-même. 2. On peut voir dans les textes 40 et 44 que l'utilisateur mentionne à plusieurs reprises qu'il est sur le point de se suicider, s'en rapproche de plus en plus, etc., ce qui indique une grave tendance suicidaire. 3. Dans le texte 54, il est mentionné "Peu importe jusqu'à quelle heure je reste éveillé, je dois me faire un masque", ce qui est un comportement pour faire face au stress, mais révèle également des problèmes de sommeil et des émotions anxieuses. Ces textes pointent tous vers les caractéristiques diagnostiques de la dépression, c'est pourquoi le jugement est "Oui".
user_680	Non" Analyse : Les tweets de cet utilisateur contiennent principalement diverses expressions émotionnelles, mais ne montrent pas directement les caractéristiques de l'humeur dépressive. Bien que des difficultés et défis de la vie soient mentionnés (comme la perte de poids, les problèmes relationnels, le manque de sujets de conversation communs, etc.), ces contenus relèvent davantage des tracas de la vie quotidienne et ne sont pas des symptômes typiques de la dépression. 1. Les plaintes de l'utilisateur concernant la perte de poids et les relations sociales peuvent refléter du stress ou de l'insatisfaction, mais pas au point de montrer une dépression. 2. À travers des mots à connotation positive comme "chérir", "soutenir", on peut voir que l'utilisateur porte attention aux personnes ou aux choses qu'il apprécie, ses émotions sont relativement stables. 3. "Attendre une nouvelle pièce", "Aimer les enfants", etc. montrent que l'utilisateur reste optimiste et intéressé par la vie, ce qui ne correspond pas vraiment à l'attitude négative de la dépression. 4. Bien qu'une faible estime de soi soit mentionnée dans certains tweets (comme "Être un poids pour la société"), dans l'ensemble il n'y a pas de précipitation ou de déni de soi continu. 5. Aucun changement flagrant de sommeil/appétit, symptôme physique ou autre malaise physiologique n'est mentionné, alors que ce sont généralement des manifestations de l'humeur dépressive. En résumé, bien que les tweets contiennent des défis et des difficultés, il manque une expression profonde de l'humeur dépressive, donc on ne peut pas conclure que cet utilisateur présente une humeur dépressive.	Non, pas de symptômes de dépression : À 10 ans, déjà inquiet de grossir en mangeant des bâtonnets épicés, on peut voir une certaine anxiété. En lisant les nouvelles et en déplorant la fragilité de la vie, on peut voir l'existence d'émotions tristes. Pas d'autres symptômes liés à la dépression. Il est recommandé d'avoir une alimentation raisonnable, de libérer la nostalgie des bâtonnets épicés, de détendre l'humeur, et d'interagir davantage avec les amis.	"Non" Analyse : Dans les textes de l'utilisateur, l'humeur dépressive n'est pas directement mentionnée. Dans les textes 1, 23, 46, 50, 57, 68 et 72, l'utilisateur mentionne certaines émotions négatives ou difficultés de la vie, comme la perte de poids, l'inquiétude d'être à la traîne dans la société, l'insomnie, le manque de sujets de conversation communs, etc., mais cela n'est pas suffisant pour indiquer clairement l'existence d'une humeur dépressive. Par exemple, lorsque l'utilisateur mentionne "Recevoir des coups rouges tous les jours sur Ele.me", cela peut exprimer les plaisirs de la vie quotidienne, et non les symptômes de la dépression. Dans le texte 72, "Comment suis-je soudainement devenu déprimé, que dois-je faire ?" montre que l'utilisateur s'interroge sur ses propres sentiments, mais ne manifeste pas un état dépressif persistant. La plupart des autres contenus sont davantage orientés vers la vie quotidienne, les interactions émotionnelles ou les commentaires sur le comportement d'autrui, sans pointer directement vers une humeur dépressive. Par conséquent, dans l'ensemble, il n'y a pas suffisamment de preuves pour affirmer que cet utilisateur présente une humeur dépressive.	"Oui" Analyse : 1. Dans le texte 22 "Soudain je n'ai pas réagi, il semble qu'il y a eu un moment de panne du serveur", le texte 30 "Joyeux 31ème anniversaire ! Lee Sansai est également adorable en grandissant" et le texte 36 "Fan bouddhiste. Pas de disputes ni de courses. Retardé, mais félicitations à Hot Brother, joyeux anniversaire", l'utilisateur décrit et exprime son propre état émotionnel. 2. Dans le texte 72 "Comment suis-je soudainement devenu déprimé, que dois-je faire ?" et le texte 73 "Comment suis-je soudainement devenu déprimé, que dois-je faire ?", "déprimé" est mentionné deux fois de suite, ce qui est une expression évidente de symptômes d'humeur dépressive. 3. Dans ces textes, l'utilisateur ne mentionne pas clairement s'il a reçu un diagnostic, mais ses descriptions et expressions correspondent à certains symptômes courants de la dépression, tels qu'une humeur déprimée persistante, une perte d'intérêt ou de plaisir, un sentiment d'impuissance ou de manque de valeur, etc. Sur la base de l'analyse ci-dessus, cet utilisateur présente une humeur dépressive.

TABLE 4 – Exemples de réponses traduites en français