Evaluating the Performance of GPT-assisted Identification and Classification of Eating Disorders with Text-based Chinese Social Media Data

Tianqiang Yan, Yucheng Zhang, Jiayi Han, Zhiyuan Liu, Wesley Barnhart, Shaojing Sun, Jianjun Zhou, Feng Ji, Jinbo He

Submitted to: Journal of Medical Internet Research
on: March 04, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Evaluating the Performance of GPT-assisted Identification and Classification of Eating Disorders with Text-based Chinese Social Media Data

Tianqiang Yan[1]; Yucheng Zhang[1]; Jiayi Han[1]; Zhiyuan Liu[2]; Wesley Barnhart[3]; Shaojing Sun[2]; Jianjun Zhou[4]; Feng Ji[5]; Jinbo He[1]

[1]The Chinese University of Hong Kong, Shenzhen Shenzhen CN
[2]Fudan University Shanghai CN
[3]Bowling Green State University Bowling Green US
[4]Shenzhen Yutong Yule Technology Co., Ltd. Shenzhen CN
[5]University of Toronto Toronto CA

Corresponding Author:
Jinbo He
The Chinese University of Hong Kong, Shenzhen
SR1-1901, Upper campus
Chinese University of Hong Kong Shenzhen
Shenzhen
CN

Abstract

Background: Eating disorders (EDs) are related to an array of negative health outcomes and have been a major public health concern globally, including in China. However, the rates of detection and treatment-seeking for EDs in China are low and the effective treatment is even lower. Thus, exploring new ways to detect and classify EDs has significant implications for EDs prevention and treatment in China.

Objective: This study aimed to evaluate the performance of large language models (LLMs), particularly OpenAI's GPT-4, on the identification and classification of EDs, utilizing real-world Chinese plain-text social media data.

Methods: We evaluated the performance of LLMs with two hierarchical tasks, including the Phase 1 task of judging whether a sample was ED-positive, and the Phase 2 task of inferring the specific ED subtypes for positive samples, including anorexia nervosa (AN), bulimia nervosa (BN), and binge-eating disorder (BED). GPT-4 was selected as the representative of state-of-the-art LLMs, tuned with natural language instructions in a manner of zero-shot Chain-of-Thought (CoT) prompting based on manually-edited ED criteria. The performance of GPT-4 was compared with three baseline schemes, including ERNIE 3.0, 1-gram Bag-of-Words (BoW), and 3-gram BoW. The performance was quantified through overall accuracy and linear accuracy.

Results: In the Phase 1 task of identifying ED-positive samples, GPT-4 showed the lowest overall accuracy of 0.768, compared with that of the baselines (0.810-0.818). However, in the Phase 2 task of classifying AN, BN, and BED, GPT-4 outperformed the others, with a linear accuracy of 0.943 (0.687-0.877 for baselines) and an overall accuracy of 0.887 (0.373-0.753 for baselines).

Conclusions: These findings suggest that GPT-4's zero-shot in-context learning capability may be better suited for classifying complex semantic capabilities such as ED subtypes (e.g., AN, BN, and BED). Also, conventional, non-LLM methods (ERNIE 3.0, 1-gram BoW, and 3-gram BoW) may be better suited for the initial identification of probable EDs.

(JMIR Preprints 04/03/2024:58054)
DOI: https://doi.org/10.2196/preprints.58054

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Evaluating the Performance of GPT-assisted Identification and Classification of Eating Disorders with Text-based Chinese Social Media Data

Abstract

Background: Eating disorders (EDs) are related to an array of negative health outcomes and have been a major public health concern globally, including in China. However, the rates of detection and treatment-seeking for EDs in China are low and the effective treatment is even lower. Thus, exploring new ways to detect and classify EDs has significant implications for EDs prevention and treatment in China.
Objective: This study aimed to evaluate the performance of large language models (LLMs), particularly OpenAI's GPT-4, on the identification and classification of EDs, utilizing real-world Chinese plain-text social media data.
Methods: We evaluated the performance of LLMs with two hierarchical tasks, including the Phase 1 task of judging whether a sample was ED-positive, and the Phase 2 task of inferring the specific ED subtypes for positive samples, including anorexia nervosa (AN), bulimia nervosa (BN), and binge-eating disorder (BED). GPT-4 was selected as the representative of state-of-the-art LLMs, tuned with natural language instructions in a manner of zero-shot Chain-of-Thought (CoT) prompting based on manually-edited ED criteria. The performance of GPT-4 was compared with three baseline schemes, including ERNIE 3.0, 1-gram Bag-of-Words (BoW), and 3-gram BoW. The performance was quantified through overall accuracy and linear accuracy.
Results: In the Phase 1 task of identifying ED-positive samples, GPT-4 showed the lowest overall accuracy of 0.768, compared with that of the baselines (0.810 −¿ 0.818). However, in the Phase 2 task of classifying AN, BN, and BED, GPT-4 outperformed the others, with a linear accuracy of 0.943 (0.687 −¿ 0.877 for baselines) and an overall accuracy of 0.887 (0.373 −¿ 0.753 for baselines).
Conclusions: These findings suggest that GPT-4's zero-shot in-context learning capability may be better suited for classifying complex semantic capabilities such as ED subtypes (e.g., AN, BN, and BED). Also, conventional, non-LLM methods (ERNIE 3.0, 1-gram BoW, and 3-gram BoW) may be better suited for the initial identification of probable EDs.

Keywords: Large Language Models; GPT; eating disorders; identification; classification; social media

Introduction

Eating disorders (EDs) are psychiatric conditions characterized by severe disturbances in eating behaviors and related thoughts and emotions [1,2]. EDs comprise several subtypes, such as anorexia nervosa (AN), bulimia nervosa (BN), and binge-eating disorder (BED) [1]. EDs are linked to an array of adverse health consequences, including depression, anxiety, breast cancer, high engagement in non-suicidal self-injury, high suicidality, high mortality rates, and reduced quality of life [3-10]. Moreover, ED is difficult to treat due to the unique and complex challenges that patients with EDs bring to treatment providers.

According to an epidemiological systematic review [11], EDs are prevalent worldwide (5.7% women with accurate ED diagnosis and 19.4% women with EDs as broad categories; 2.2% men with accurate ED diagnosis and 13.8% men with EDs as broad categories), with an increasing trend of prevalence (from 3.5% in the 2000-2006 period to 7.8% in the 2013-2018 period). Thus, EDs are an important global public health concern [12-17], including in China [18,19].

However, in China, despite its large population base and the relatively high estimates of EDs (e.g., 1.05 % for AN, 2.98 % for BN, and 3.53 % for BED in a large-scale epidemiological investigation in Chinese college women) [20], the rates of detection and treatment-seeking for EDs are low and the effective treatment is even lower[21]. One potential explanation for this dilemma is that EDs received relatively little publicity in Chinese media, leading to low public awareness of EDs in China [22]. Indeed, prior research suggests that the lack of recognition of EDs as a mental illness is a key barrier for not seeking for ED treatment among Chinese individuals with EDs [23]. Given that early detection of EDs has significant clinical implications (e.g., improvement in prognosis and decrease in morbidity and mortality) [24]. Thus, efforts to improve the identification of EDs in China are needed.

With the development of technology and the widespread use of social media, identification of individuals with EDs via social media data is promising [25-27]. To date, there have been many studies that explored the performance of using social media data to detect EDs. For instance, [28] found it effective to use machine learning methods like decision trees (i.e., ADTree) to distinguish pro-ED and non-pro-ED posts from social media platforms like Tumblr and Twitter. Also, [29] discovered that manual feature engineering (i.e., vocabulary extraction and topic modeling) based on raw text data helps enhance the identification, and proved his proposal by successfully applied manual feature engineering to social media posts, attaining promising results on detecting AN. Furthermore, [30] proposed an automated feature engineering scheme, through multiple machine learning approaches like Bag-of-Words (BoW) [31], TF-IDF [32], and Word2Vec [33], to preprocess social media posts from Reddit, and such preprocessing manner helped them achieve auto-detection of ED vs. not ED in social media samples with only 4% error rate. More recently, [25] proved that modern neural networks (i.e., convolutional neural networks [34] and recurrent neural networks [35]) are outstanding feature extractors and learners of social media data. In their study, both raw text representations and manually engineered features are utilized as the inputs of their classifier, and the results showcased 100% high-risk ED sample detection rate on Reddit posts.

Despite the above-mentioned progress in harnessing social media data, there is still plenty room for improvement. First, the majority of existing studies based on social media text data mainly investigated whether their methodologies could accurately identify an ED (e.g., AN) or EDs as a whole group, but did not further assess the performance of these methods on classifying ED subtypes (e.g., AN, BN, and BED). Second, the advent of large language models (LLMs) has revolutionized learning-based schemes with its exclusive, zero-shot in-context learning paradigm where no training is required [36,37]. To our knowledge, there still lacks research evaluating the effectiveness of this new technique on ED identification and classification and how it performs compared with conventional data-driven methods, namely traditional machine learning-based models (e.g., BoW, TF-IDF) and non-LLM deep learning-based models (BERT) that require supervised training [38,39]. Furthermore, ED-related social media contents vary across different linguistic and cultural contexts [22]. Despite this, existing studies have focused on English-oriented social media platforms, with little attention paid to social media platforms in the Chinese context.

To fill these gaps, the present study examined ED identification and classification via Chinese social media data with the state-of-the-art LLM GPT-4. In particular, we chose Zhihu, a widely-used Chinese social media platform as the social media platform to reflect our ED samples, and we leveraged the GPT-4's zero-shot in-context learning capability by prompting it, namely "instruction tuning." To fully examine the effectiveness and the performance of LLM-based ED identification and classification, we conducted two evaluation phases, including a preliminary "ED or not" binary classification task and a more fine-grained "AN, BN, or BED" multi-classification task. The performance of GPT-4's in-context learning was compared with that of three baseline data-driven NLP (ERNIE 3.0, 1-gram BoW, and 3-gram BoW) methods across two evaluation phases. In the following sub-sections of introduction, we reviewed the literature on using LLMs in detecting mental health issues and introduced the potentials of using LLMs in detecting EDs.

Use of Large Language Models in Detecting Mental Illnesses

The emergence and popularity of LLMs, such as OpenAI's GPT-4 [40], have paved the way for a wide range of disciplines where the requirement for direct text analysis is ubiquitous yet too complicated for conventional data-driven methods. Their transformative impact is particularly noticeable in psychology, reshaping traditional screening and therapeutic practices [41].

A significant aspect of studying psychological issues involves the modeling of natural language since people express themselves more frequently and heavily through words and sentences than through other modalities. Yet, linguistic cues are considered one of the most complicated modalities, owing to the underlying characteristics uniquely held by natural language, such as sparsity [42], diversity [43], uncertainty [44], and connotation [45]. Given the complexity of processing natural language, previous studies have proposed a substantial number of data-driven methods, where massive natural language corpora are utilized to train the models. However, in addition to the demands for computational resources, time, and storage, these NLP methods usually struggle with transferability; that is, if one of them is applied to a different field, the model would require retraining with a new, domain-specific dataset and such a procedure can be highly resource-intensive. LLMs can overcome these long-standing issues. Typically, a LLM employs a pre-training corpus that is not only voluminous but also extensively varied, covering a multitude of everyday situations and specialized domains. It even spans multiple language families, historical contexts, and cultural realms. Coupled with an LLM's billions of parameters, these factors grant its unparalleled semantic comprehension, reasoning, generation, and generalizability. These advantages help fulfill the aspirations of psychological research that intersects with natural language processing, providing novel perspectives in detecting and understanding a wide range of psychological issues [46].

Additionally, the scalability and accessibility of these models contribute to the promotion of mental health care's universal access. Especially, LLMs are designed to be deployed without the necessity of retraining, thanks to its rich and broad background knowledge as well as its intuitive and human-friendly natural language interaction logic. These inherent superiorities allow a quick, flexible, and low-cost implementation for various scenarios, enabling virtual consultations and automated mental screening, effectively widening the reach of psychological services beyond geographical and time constraints [47,48].

Potentials of Large Language Models in Identification and Classification of EDs

Eating Disorders (EDs), as the main focus of this study, are considered challenging with respect to the natural language-based content analysis, which is especially the case when dealing with social media data. In China, where diversities of regional, cultural, and educational backgrounds among the population exist, digitalization and networking services have been developed and popularized at a tremendous speed in recent years [49]. Hence, combined with the widespread connotations and the variances in the Chinese language [50,51], mental health screening for posts from individuals on Chinese social media platforms can be arduous. Identifying "ED or not" can still be straightforward given the fact that the negative output "0" and the positive output "1" are orthogonal, and data-driven models are skilled enough at extracting and fitting locally-distributed orthogonal features of the inputs, regardless of the global semantic-level complexity [52,53]. However, identifying whether a post reflects AN, BN, or BED is not orthogonal. For example, BN and BED share some similar symptoms (e.g., both BN and BED patients may experience low mood and feelings of losing control and guilty afterwards) and behaviors (e.g., both BN and BED are featured by binge eating; but unlike BN, BED does not involve regular unhealthy compensatory behaviors), resulting in frequent intersecting expressions between BN posts and BED posts that confuse even human experts [22,54,55]. Accordingly, semantic-level reasoning with sufficient cross-domain knowledge of both linguistics and EDs are of great importance for the identification of specific ED subtypes. Such capabilities may be achieved by LLMs [56].

The Present Study

Overall, this study explored the possibility of utilizing GPT-4 to identify and classify EDs. We created the prompts to identify EDs guided by the definitions in the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) [1]. Then, we tuned the GPT-4 with the prompts to obtain the classification outcomes of ED subtypes. To better evaluate the performance of GPT-4, we chose and tested three data-driven baseline methods for comparison, including ERNIE 3.0, 1-gram Bag-of-Words (BoW), and 3-gram BoW. We assessed their performance with the same tasks. We chose these baselines because they are widely utilized as the preprocessing methods of text data and have been employed in conjunction with a downstream CatBoost classifier for both training and inference [57,58]. Also, due to the utilization of zero-shot in-context learning, GPT-4 does not require retraining and can perform inference directly on the validation data. This is fundamentally different from the conventional task-specific supervised learning paradigm required by the baselines. Fortunately, multiple recent studies reported that the language model's in-context learning demonstrates comparable or even better performance than the supervised learning-based models [36,59,60], which further encouraged us to explore the LLM's potential on ED screening.

Method

Dataset Description

Data used in the present study were from a project examining the perceived causes of EDs in Chinese Zhihu users with self-report EDs status. Specifically, data were extracted through the open API of Zhihu, a popular Chinese social media platform similar to Quora, where people share their experiences and knowledge in a Q&A format. We extracted the posts (i.e., answers) from the Zhihu users who responded to questions related to EDs [e.g., "□□□□□□□□□□□□" (How did you come to develop anorexia nervosa)]. We specifically focused on AN, BN, and BED because these are the main types of EDs discussed in Chinese social media [22]. A total of 5199 posts were obtained. Data were manually labeled by two groups of well-trained research assistants through carefully reading posts (Cohen's Kappa = 0.714-0.944), and inconsistency was resolved via discussions within the research team. Labeling was conducted in two stages. At the first stage, research assistants labeled the users as 1 "positive," indicating that the users self-reported as having an ED, and 0 "negative," indicating that the users did not self-report as having an ED (e.g., some users responded to the ED relevant questions to show their support or to share their friends' ED experiences). A total of 2098 data points were labeled as positive "1," and the remaining data were labeled as negative "0." At the second stage, based on the users' self-reported diagnosis and ED symptoms described in the posts, research assistants further classified the users into three categories, including AN, BN, and BED. However, 19 posts with unclear diagnosis information were removed at the second stage. Thus, a total of 2079 data entries remained (306 AN, 782 BN, and 991 BED). Then, we randomly selected 100 posts from the categories of non-ED, AN, BN, and BED (i.e., a total of 400 posts with an even distribution of each category). In the first binary classification task of our evaluation phase, all 5199 data with “ED or not” binary labels were included, where 400 (300 positives and 100 negatives) data were utilized as the validation set for both baselines and GPT-4, and the remaining 4799 data were utilized as the training set for baseline models requiring supervised training. In the second multi-classification task, the 2079 data from the second stage with “AN, BN, or BED” multi-class labels were included, where 300 (100 ANs, 100 BNs, and 100 BEDs) data were utilized as the validation set, and the remaining 1779 data were utilized to train the baseline models.

Instruction Tuning with GPT-4

The development of modern language models has revolutionized Natural Language Processing (NLP) and Natural Language Generation (NLG), enabling machines to learn and understand human language with remarkable accuracy. Starting with the Transformer model [61], the game-changing architecture shifted the focus from recurrent layers to self-attention mechanisms, fostering improvements in various NLP and NLG tasks. The OpenAI’s GPT (Generative Pretrained Transformers) series [36], especially the latest GPT-4 [40], is the state-of-the-art generative LLM, exploiting the power of unsupervised learning and scale to achieve human-like fluency and comprehension. The effectiveness of LLMs largely depends on the pivotal process called instruction tuning, which is a straightforward paradigm that emphasizes guiding the model’s behavior with explicit instructions.

The GPT-4, which employs a pure decoder-based Transformer framework, typically utilizes zero-shot or few-shot chain-of-thought (CoT) prompting mechanisms for instruction tuning [62]. The essence of CoT lies in facilitating optimal feedback from LLMs. This is achieved via meticulous prompt engineering, which may specifically involve transforming an abstract query into a more tangible interpretation, breaking down complex compound logic and offering supplementary resources, among other things. Such an approach steers the model towards conducting sequential reasoning guided by chained thoughts, which does not only reduce model divergence but also improves the consistency and accuracy of the output. Unlike supervised learning, the CoT mechanism does not focus on refining the sample space to enhance distribution learning of an LLM. Instead, it optimizes knowledge application strategies in a more consistent way with human interaction.

In our study, rather than utilize GPT’s innate knowledge base to classify EDs, which can be highly unreliable, we created a comprehensive knowledge base according to the definitions and diagnostic criteria of different EDs in the DSM-5 [1]. As described in (Figure 1), the “character definition” defines the “system” role attribute of the language model, serving as a psychological *domain constraint* by prompting the GPT to analyze and respond like a psychologist. “ED definition” as well as “task definition” are CoT prompts, considered as key instructions mainly consisting of the definitions of EDs and how different EDs can be identified. In addition, the semantic ambiguities are prevalent within modern social media posts, e.g., unofficial abbreviations and network words, so prompts for disambiguation are also injected for the LLM to be aware of such biases. Eventually, the combined screening instruction is subsequently fed to the GPT along with the sample to be analyzed and the description of our task. This approach steers GPT to make inferences utilizing cues from the knowledge base, thereby securing accurate and satisfactory classification outcomes.

Metrics and Parameter Settings

We employed overall accuracy and linear accuracy to assess identification and classification performance (Figure 2). Overall accuracy is a prevalent metric wherein each sample represents the smallest computational unit. Linear accuracy employs individual labels as the basic unit for the classification task. The unique feature of linear accuracy is that it transforms the actual value matrix and prediction matrix from multi-sample classification tasks in the validation set into two, one-dimensional arrays – each element representing a true or predicted label value (0 or 1). The average match rate, computed by correlating these two arrays post-flattening, equates to linear accuracy. As compared to standard accuracy measures, such an algorithm presents a more detailed mechanism for evaluating model performance. It should be noted that we do not use the evaluation metric of linear accuracy in every test phase listed in the next section. For a binary classification problem, since the positive and negative labels of samples are encoded by a single digit, linear accuracy and overall accuracy are considered equivalent (i.e., determining whether a sample has an ED), and we represent the accuracy metric in this context with overall accuracy.

Besides the metric setting, the “temperature” parameter of the GPT-4 is set to 0.2, of which the default value, 1.0, has shown to be unstable for multiple evaluation attempts. This is due to the fact that the temperature manipulates the associative ability of LLMs [40]. A too-large value, though allowing more creative feedback, significantly reduces the attention on the details of our instruction. We saw the decline of GPT-4’s performance on our task when the temperature was set too high. Additionally, it was observed that a near-zero “temperature” (e.g., 0.01, 0.02) also resulted in a performance loss. Chen stated that a too-low temperature degrades the LLM’s capability of dealing with out-of-domain (OOD) samples [63], and such samples occur widely in natural language tasks. As such, we finally settled on a temperature setting of 0.2, striking a sensible balance between performance and stability.

Settings of the Evaluation Tasks

We reported the performance of instruction-tuned GPT-4 in the Phase 1 task of ED identification (a binary classification task) and the Phase 2 task of ED classification (a multi-class classification task) based on purely textual social media data. The same prompts were used for both classification tasks to better examine and showcase the generalizable in-context learning capability of GPT-4. We compared the performance of GPT-4 with three non-LLM, conventional learning-based NLP approaches, including ERNIE 3.0, 1-gram Bag-of-Words (BoW), and 3-gram BoW, followed by a downstream CatBoost classifier. ERNIE is a pre-training framework representing a series of modern word2vec and sentence2vec models, and ERNIE 3.0 is one of the latest series that leverages more complex Transformer framework, brand new pre-training tasks, and knowledge enhancing techniques compared to its predecessors. Bag-of-Words, though much earlier proposed, is still among the most welcomed NLP methods in a wide range of research fields for its strong interpretability and distinguished generality. CatBoost is a new and popular gradient boosting tool and is widely implemented for classification tasks. Compared to XGBoost [64], an earlier proposed gradient boosting scheme receiving even broader acceptance, CatBoost is claimed to possess better generalization ability and more efficient multi-processor and GPU training [57,58]. Except GPT-4 employing the zero-shot instruction tuning, which requires no training, the data-driven baseline models were trained under supervision on samples from the dataset (excluding those for validation). (Table 1) shows baseline settings of our evaluation tasks.

Tasks	No. of samples for train/ Evaluation	Baseline upstream (vectorization)	Baseline downstream (classification)	Accuracy metric
binary (ED or not) classification	4799/400	ERNIE 3.0	CatBoost	overall accuracy
		1-gram BoW		
		3-gram BoW		
multiclass (AN,	1798/300	ERNIE 3.0		overall &

BN, BED) classification		1-gram BoW 3-gram BoW		linear accuracy

Table 1. Baseline settings of the evaluation tasks.[a]

[a]ED = Eating Disorder, AN = Anorexia Nervosa, BN = Bulimia Nervosa, BED = Binge-Eating Disorder.

Results

Binary ED Identification Performance

In the Phase 1 task, we compared the accuracy of GPT-4 and baseline NLP methods in determining the presence of ED based on the text content for 400 samples (300 positives and 100 negatives) of the validation set, where the numerical result is demonstrated in (Table 2). It can be observed that, in this stage, GPT-4 unexpectedly achieved the least desirable result (overall accuracy: 0.768), while classifiers based on BoW and ERNIE 3.0 showed better performance on the validation set (overall accuracy: 0.818 for ERNIE 3.0 and 3-gram BoW, 0.810 for 1-gram BoW). Statistics on the proportion of true positives (TP) and true negatives (TN) achieved by each method revealed that GPT-4 tended to classify samples as positive (having ED), with a positive prediction accuracy reaching 1 (all predictions correct), whereas the negative prediction accuracy was only 0.410. On the other side, the non-LLM methods did not show a significant difference between the proportions of TN and TP, and in most cases, the proportion of TN was slightly higher than that of TP (0.850 vs. 0.807 for ERNIE 3.0, 0.830 vs. 0.803 for 1-gram BoW, 0.810 vs. 0.820 for 3-gram BoW).

Table 2. Result of baselines and GPT-4 performing the “ED or not” binary classification.[a]

Model	TN/TP	Overall accuracy
ERNIE 3.0	**0.850**/0.807	**0.818**
1-gram BoW	0.830/0.803	0.810
3-gram BoW	0.810/0.820	**0.818**
GPT-4 (zero-shot CoT)	0.410/**1.000**	0.768

[a]ED = Eating Disorder, AN = Anorexia Nervosa, BN = Bulimia Nervosa, BED = Binge-Eating Disorder. TN = True Negative, TP = True Positive.

Multiclass ED Classification Performance

In the Phase 2 task, we employed 300 positive samples out of a total of 400 in the validation set, where each of them was labeled as AN, BN, or BED to assess the ability of GPT-4 in making detailed predictions on these samples compared to the baseline methods. The number (100) and proportion (1/3) of samples corresponding to each label are consistent. During data preprocessing, all labels are one-hot encoded for the convenience of training, evaluation, and analysis. The numerical result is shown in (Table 3). The accuracy of AN, BN, and BED is calculated binarily by $(TN+TP)/n_{subtype}$, where $n_{subtype}=100$ for every ED subtype in our setup.

Table 3. Result of baselines and GPT-4 performing the “AN, BN, BED” multi-classification.[a]

Model	AN accuracy	BN accuracy	BED accuracy	Linear accuracy	Overall accuracy
ERNIE 3.0	0.867	0.813	0.827	0.877	0.753
1-gram BoW	0.815	0.710	0.736	0.815	0.630

3-gram BoW	0.601	0.569	0.576	0.687	0.373
GPT-4 (zero-shot CoT)	**0.943**	**0.897**	**0.933**	**0.943**	**0.887**

[a]ED = Eating Disorder, AN = Anorexia Nervosa, BN = Bulimia Nervosa, BED = Binge-Eating Disorder..

In contrast to the result from the previous evaluation, GPT-4 demonstrated a considerable increase in accuracy for the multi-classification task across three subtypes of EDs, achieving more than 0.8 out of 1 on all indicators regarding accuracy, more than half number of which even exceeds 0.9 (0.943 linear accuracy and AN accuracy, 0.933 BED accuracy). Compared to the baseline methods, GPT-4's overall accuracy improved by 17.8% over ERNIE 3.0, 40.8% over 1-gram BoW, and 137.8% over 3-gram BoW. In terms of linear accuracy, it also showed enhancements of 7.5% over ERNIE 3.0, 15.7% over 1-gram BoW, and 37.3% over 3-gram BoW. In terms of predicting outcomes for each ED subtype, all methods consistently achieved the highest accuracy in identifying positive samples as AN, followed by BED. However, BN samples had the lowest proportion of accurate predictions.

Discussion

Generally speaking, the zero-shot instruction tuning method we adopted did not enable GPT-4 to provide optimal accuracy in the "ED or not" binary prediction task, but for the finer-grained task, like predicting the specific subtypes of ED, prompted GPT-4 managed to attain state-of-the-art performance.

From the perspective of the learning paradigm, the unique zero-shot in-context learning capability inherent to LLMs such as GPT-4, enabling them to directly infer on the samples from the validation set, is distinctly different from the traditional paradigm that data-driven models typically rely on because they are trained on a specified dataset for completing a task in the same domain. Fundamentally, a data-driven model (i.e., any baseline models implemented in our paper) is optimized over a specific dataset and task, thus it exhibits a stronger capacity for fitting lower-level features and shows greater performance to such data and task. In contrast, with a considerably larger number of parameters, GPT-4 is pretrained on a massive corpus of cross-domain textual samples and diverse semantic downstream tasks, empowering its high-level semantic attribution capabilities. The preliminary "ED or not" binary classification task may gain some advantages by identifying certain (sometimes biased) feature, leading their performance to potentially surpass that of the more generalization-oriented GPT-4. Although they work well in this dataset, nonetheless, it does not imply that these models will exhibit the same level generalizability over a wider range of datasets. As opposed to this coarse-grained binary classification tasks, when it comes to the more intricate ED subtype classification, the limited combinations of low-level features become insufficient for capturing the real-world distribution of samples. In this case, the in-context learning that GPT possess is capable of accessing higher-level semantic features. It incorporates logical and strategic reasoning trajectories, as well as an abundance of supportive background knowledge. This combination allows the large language model not to be bounded by the constraints of data limitation, and to provide self-explanatory inferences. As a result, GPT-4 can be better when tackling complex reasoning tasks requiring more contexts and higher levels of granularity. In addition, the primary framework of GPT-4, Transformer, which focuses on capturing the global semantic features of the text, may sometimes lead to the loss of critical local semantic information [65].

In general, the results speak to the performance and potential of the instruction-tuned GPT-4 utilizing in-context learning in the identification of eating disorders given purely textual social media data. The results combined with our analysis suggest that GPT-4 can serve as a powerful auxiliary tool for subtype analysis and the identification of eating disorders, which usually requires substantial background knowledge and solid professional competence. When there is limited resources of collecting data and training specialized machine learning models for a preliminary screening (i.e. ED or not), we can also consider utilizing the zero-shot feature of GPT-4 while intervention and guidance from human experts remain critical at the current stage. If a fully automated screening process is attempted to achieve, we recommend combining mainstream non-LLM NLP methods with LLMs like GPT-4, which tends to strike a balance between reliability, effectiveness, and efficiency.

The exceptional performance of LLMs like GPT-4 in complex text analysis tasks is largely attributed to its high-level abstract reasoning which aligns with human intuition and a hyper-rational approach that enables it not only to think with human-like strategy and logic, but also to navigate cognitive biases that frequently trap ordinary people[66-68].

The utilization of large language models offers several advantages. First and foremost is the low-cost and efficient local deployment. Notably, open-access large language models like OpenAI's GPT provide remote API interfaces. Put in other words, this eliminates the need for users to download voluminous model parameters, saving both time and network resources. Furthermore, there is no requirement for locally deployed high-performance computing devices for additional training or inference. Additionally, the high-level API functionality allows end-users to implement end-to-end task flows with minimal code. Second is human-level natural language comprehension: large language models, meticulously pre-trained on vast and diverse corpora across various domains, exhibit exceptional generalization capabilities. Consequently, they excel in downstream tasks such as analyzing textual expressions of psychological symptoms. These tasks, characterized by their complexity and specialized nature, pose significant challenges for conventional NLP models. Third is the the inherent interpretability of large language models in their innate chain-like thinking and text generation abilities. These qualities enable them to provide self-explanatory analysis results. By addressing the issues of interpretability and reliability that have long plagued the field of deep learning, large language models significantly outperform traditional machine learning algorithms, which, despite possessing some degree of interpretability, often struggle to yield acceptable performance on intricate text analysis tasks. Moreover, large language models demonstrate outstanding domain adaptation capabilities. When confronted with specific downstream tasks, users typically only need to provide task-relevant definitions or prompts. The models autonomously extract key guiding information from these prompts. Leveraging their superior understanding and reasoning abilities, large language models establish associations and mappings among domain knowledge, tasks, samples, and results.

It is important to acknowledge the limitations and related future research directions of this study. Firstly, we did not assess the credibility of each user entry, which is a pervasive constraint in research within this field. Addressing this limitation is crucial, and in our future work, we intend to explore methods such as anomaly detection or leveraging large language models to quantify the credibility of samples. Furthermore, considering the relatively high token usage cost of GPT-4, we carefully selected a small yet representative sample to ensure a fair assessment of model performance while keeping costs manageable. Nevertheless, it remains uncertain whether the same prompt and model would yield similarly high identification accuracy when applied to out-of-domain data. We recognize the need for improvement in this aspect and will also incorporate it into our future endeavors. We invite and encourage eating disorder researchers to extend the methodology presented in this paper to a wider spectrum of social media data and tasks. Notably, for platforms with English-language content, such as Reddit, research thus far has relied on content analysis strategies without employing language models for studying eating disorder text data [69,70]. Given that state-of-the-art models are often reported to perform optimally in analysis and question-answering benchmarks on English texts, we anticipate the emergence of enhanced analytical and screening capabilities when these large language models are applied to English eating disorder text datasets.

Conclusion

In summary, our study provided novel discussions and insights on an emerging trend of natural language processing for eating disorders by integrating large language models with instruction tuning and in-context learning. This emerging technology points to new paths for research in various domains within psychology, not limited to eating disorders. It empowers researchers to conduct robust automated analysis directly on abundant textual data, liberating them from the constraints of traditional information collection and metric quantification methods. With lower costs and more efficient workflows, researchers can achieve performance levels that rival or even surpass those of human experts across a broad spectrum of objectives. We anticipate a future where automated analysis powered by large language models plays a vital role in assisting mental health professionals and researchers in diagnosing and understanding eating disorders and other psychological conditions. This technology has the potential to enhance accessibility, efficiency, and patient care, ultimately leading to improved outcomes in the field of mental health.

Acknowledgements

Jinbo He was supported by the National Natural Science Foundation of China (Grant Number: 72204208) and the 2022 Shenzhen College Stability Support Plan (2022 □□□□□□□□□□□□□□□) from Shenzhen Science and Technology Innovation Commission (Grant Number: 2023SC0028).

Tianqiang Yan: Formal analysis, Writing – original draft, Writing – review & editing.
Yuchen Zhang: Data curation, Investigation, Writing – original draft, Writing – review & editing.
Jiayi Han: Writing – review & editing.
Zhiyuan Liu: Investigation, Writing – review & editing.
Wesley R. Barnhart: Writing — review & editing.
Shaojing Sun: Writing – review & editing.
Jianjun Zhu: Supervision, Writing – review & editing.
Feng Ji: Conceptualization, Supervision, Writing – review & editing.
Jinbo He: Conceptualization, Supervision, Funding acquisition, Project administration, Writing – original draft, Writing – review & editing.

Conflicts of Interest

The authors have no conflict(s) of interest to declare.

Abbreviations

AN: Anorexia Nervosa
API: Application Programming Interface
BED: Binge-Eating Nervosa
BN: Bulimia Nervosa
BoW: Bag-of-Words
CoT: Chain-of-Thought
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, 5th Edition
ED: Eating Disorder
ERNIE: Enhanced Language RepresentatioN with Informative Entities

GPT: Generative Pretrained Transformers
GPU: Graphic Processing Unit
LLM: Large Language Model
NLG: Natural Language Generation
NLP: Natural Language Processing
TF-IDF: Term Frequency-Inverse Document Frequency

References

1. American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders: DSM-5. Washington, DC: American Psychiatric Association.
2. Vögele, C., Lutz, A. P., & Gibson, E. L. (2018). Mood, emotions and eating disorders. The Oxford Handbook of Eating Disorders, 155-186.
3. Sansone, R. A., Levitt, J. L., & Sansone, L. A. (2005). Psychotropic medications, self-harm behavior, and eating disorders. In Self-Harm Behavior and Eating Disorders (pp. 263-276). Routledge.
4. Franko, D. L., & Keel, P. K. (2006). Suicidality in eating disorders: occurrence, correlates, and clinical implications. Clinical Psychology Review, 26(6), 769-782. https://doi.org/10.1016/j.cpr.2006.04.001
5. Berkman, N. D., Lohr, K. N., & Bulik, C. M. (2007). Outcomes of eating disorders: a systematic review of the literature. International Journal of Eating disorders, 40(4), 293-309. https://doi.org/10.1002/eat.20369
6. Steinhausen, H. C. (2009). Outcome of eating disorders. Child and adolescent psychiatric clinics of North America, 18(1), 225-242. https://doi.org/10.1016/j.chc.2008.07.013
7. Ginty, A. T., Phillips, A. C., Higgs, S., Heaney, J. L., & Carroll, D. (2012). Disordered eating behaviour is associated with blunted cortisol and cardiovascular reactions to acute psychological stress. Psychoneuroendocrinology, 37,715–724. https://doi.org/10.1016/j.psyneuen.2011.09.004
8. O’Brien, K. M., Whelan, D. R., Sandler, D. P., & Weinberg, C. R. (2017). Eating disorders and breast cancer. Cancer Epidemiology Biomarkers & Prevention, 26, 206–211. https://doi.org/10.1158/1055-9965.EPI-16-0587
9. Kärkkäinen, U., Mustelin, L., Raevuori, A., Kaprio, J., & Keski-Rahkonen, A. (2018). Do disordered eating behaviours have long-term health-related consequences?. European Eating Disorders Review, 26(1), 22-28. https://doi.org/10.1002/erv.2568
10. Radunz, M., Keegan, E., Osenk, I., & Wade, T. D. (2020). Relationship between eating disorder duration and treatment outcome: Systematic review and meta-analysis. International Journal of Eating Disorders, 53(11), 1761-1773. https://doi.org/10.1002/eat.23373
11. Galmiche, M., Déchelotte, P., Lambert, G., & Tavolacci, M. P. (2019). Prevalence of eating disorders over the 2000–2018 period: A systematic literature review. American Journal of Clinical Nutrition, 109, 1402–1413. https://doi.org/10.1093/ajcn/nqy342
12. Hoek, H. W. (2016). Review of the worldwide epidemiology of eating disorders. Current opinion in psychiatry, 29(6), 336-339. doi: 10.1097/YCO.0000000000000282
13. Erskine, H. E., Whiteford, H. A., & Pike, K. M. (2016). The global burden of eating disorders. Current Opinion in Psychiatry, 29(6), 346-353. https://doi.org/10.1097/YCO.0000000000000276
14. Thomas, J. J., Lee, S., & Becker, A. E. (2016). Updates in the epidemiology of eating disorders in Asia and the Pacific. Current Opinion in Psychiatry, 29(6), 354-362. https://doi.org/10.1097/YCO.0000000000000288
15. Wu, J., Liu, J., Li, S., Ma, H., & Wang, Y. (2020). Trends in the prevalence and disability-

adjusted life years of eating disorders from 1990 to 2017: results from the Global Burden of Disease Study 2017. Epidemiology and Psychiatric Sciences, 29, e191. https://doi.org/10.1017/S2045796020001055

16. Udo, T., & Grilo, C. M. (2018). Prevalence and correlates of DSM-5–defined eating disorders in a nationally representative sample of US adults. Biological Psychiatry, 84(5), 345-354. https://doi.org/10.1016/j.biopsych.2018.03.014
17. Santomauro, D. F., Melen, S., Mitchison, D., Vos, T., Whiteford, H., & Ferrari, A. J. (2021). The hidden burden of eating disorders: an extension of estimates from the Global Burden of Disease Study 2019. The Lancet Psychiatry, 8(4), 320-328. https://doi.org/10.1016/S2215-0366(21)00040-7
18. Li, Z., Wang, L., Guan, H., Han, C., Cui, P., Liu, A., & Li, Y. (2021). Burden of eating disorders in China, 1990-2019: an updated systematic analysis of the global burden of disease study 2019. Frontiers in psychiatry, 12, 632418.
19. Wu, J., Lin, Z., Liu, Z., He, H., Bai, L., & Lyu, J. (2022). Secular trends in the incidence of eating disorders in China from 1990 to 2017: A joinpoint and age–period–cohort analysis. Psychological medicine, 52(5), 946-956. https://doi.org/10.1017/S0033291720002706
20. Tong, J., Miao, S., Wang, J., Yang, F., Lai, H., Zhang, C., ... & Hsu, L. G. (2014). A two-stage epidemiologic study on prevalence of eating disorders in female university students in Wuhan, China. Social Psychiatry and Psychiatric Epidemiology, 49, 499–505. https://doi.org/10.1007/s00127-013-0694-y
21. Kong, Q. (2018). Interpretation of the guideline for prevention and treatment of eating disorders Chinese. Chinese Journal of Psychiatry, 51(6), 355-358. https://doi.org/10.3760/cma.j.issn.1006-7884.2018.06.003
22. Sun, S., He, J., Fan, X., Chen, Y., & Lu, X. (2020). Chinese media coverage of eating disorders: Disorder representations and patient profiles. International Journal of Eating Disorders, 53(1), 113–122. https://doi. org/10.1002/eat.23154
23. Ma, R., Zhang, M., Oakman, J. M., Wang, J., Zhu, S., Zhao, C., ... & Buchanan, N. T. (2021). Eating disorders treatment experiences and social support: Perspectives from service seekers in mainland China. *International Journal of Eating Disorders*, *54*(8), 1537-1548. https://doi.org/10.1002/eat.23565
24. Kalindjian, N., Hirot, F., Stona, A. C., Huas, C., & Godart, N. (2022). Early detection of eating disorders: a scoping review. Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity, 27, 21-68. https://doi.org/10.1007/s40519-021-01164-x
25. Tébar, B., & Gopalan, A. (2021, December). Early Detection of Eating Disorders using Social Media. In 2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) (pp. 193-198). IEEE. https://doi.org/10.1109/CHASE52844.2021.00042
26. Aragon, M. E., Lopez-Monroy, A. P., Gonzalez-Gurrola, L. C. G., & Montes, M. (2021). Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression. IEEE Transactions on Affective Computing. https://doi.org/10.1109/TAFFC.2021.3075638
27. Benítez-Andrades, J. A., Alija-Pérez, J. M., García-Rodríguez, I., Benavides, C., Alaiz-Moretón, H., Vargas, R. P., & García-Ordás, M. T. (2021, June). BERT Model-Based Approach For Detecting Categories of Tweets in the Field of Eating Disorders (ED). In 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS) (pp. 586-590). IEEE. https://doi.org/10.1109/CBMS52027.2021.00105
28. He, L., & Luo, J. (2016, December). “What makes a pro eating disorder hashtag”: Using hashtags to identify pro eating disorder tumblr posts and Twitter users. In 2016 IEEE International Conference on Big Data (Big Data) (pp. 3977-3979). IEEE.

https://doi.org/10.1109/BigData.2016.7841081
29. Mayans Yern, M. (2018). Early detection of eating disorders in reddit. http://hdl.handle.net/10230/35695
30. Yan, H., Fitzsimmons-Craft, E. E., Goodman, M., Krauss, M., Das, S., & Cavazos-Rehg, P. (2019). Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention. International Journal of Eating Disorders, 52(10), 1150-1156. https://doi.org/10.1002/eat.23148
31. Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics, 1, 43-52. https://doi.org/10.1007/s13042-010-00001-0
32. Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, No. 1, pp. 29-48).
33. Rong, X. (2014). word2vec parameter learning explained. arXiv preprint arXiv:1411.2738. https://doi.org/10.48550/arXiv.1411.2738
34. Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60, 84-90.
35. Schuster, M., & Paliwal, K.K. (1997). Bidirectional recurrent neural networks. IEEE Trans. Signal Process., 45, 2673-2681. https://doi.org/10.1109/78.650093.
36. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.
37. Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work?. arXiv preprint arXiv:2202.12837. https://doi.org/10.48550/arXiv.2202.12837
38. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.
39. Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3), 210-229. https://doi.org/10.1147/rd.33.0210
40. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
41. Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: a review of the psychiatric landscape. The Canadian Journal of Psychiatry, 64(7), 456-464. https://doi.org/10.1177/0706743719828977
42. Zamani, H., Dehghani, M., Croft, W. B., Learned-Miller, E., & Kamps, J. (2018, October). From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In Proceedings of the 27th ACM international conference on information and knowledge management (pp. 497-506). https://doi.org/10.1145/3269206.3271800
43. Henderson, J., & Brill, E. (1999). Exploiting Diversity in Natural Language Processing: Combining Parsers. In 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.
44. Xiao, Y., & Wang, W. Y. (2019, July). Quantifying uncertainties in natural language processing tasks. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 7322-7329). https://doi.org/10.1609/aaai.v34i07.6999
45. Carnap, R. (1955). Meaning and synonymy in natural languages. Philosophical studies, 6, 33-47. https://doi.org/10.1007/BF02330951
46. Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and

implementable ML systems. ACM Transactions on Computer-Human Interaction (TOCHI), 27(5), 1-53. https://doi.org/10.1145/3398069

47. 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. https://doi.org/10.48550/arXiv.2307.11991
48. He, T., Fu, G., Yu, Y., Wang, F., Li, J., Zhao, Q., ... & Yang, B. X. (2023). Towards a Psychological Generalist AI: A Survey of Current Applications of Large Language Models and Future Prospects. arXiv preprint arXiv:2312.04578. https://doi.org/10.48550/arXiv.2312.04578
49. Ito, A. (2019). Digital China: A fourth industrial revolution with Chinese characteristics?. Asia-Pacific Review, 26(2), 50-75. https://doi.org/10.1080/13439006.2019.1691836
50. Shi, W. J., & Jiang, Y. H. (2015). Comparison and contrast between English and Chinese idioms from cultural connotation perspective. Studies in Literature and Language, 10(1), 102-113. http://dx.doi.org/10.3968/6338
51. Zhang, J., & Lu, X. (2013). Variability in Chinese as a foreign language learners' development of the Chinese numeral classifier system. The Modern Language Journal, 97(S1), 46-60. https://doi.org/10.1111/j.1540-4781.2012.01423.x
52. Lin, M., Chen, Q., & BYan, S. (2013). Network in network. arXiv preprint arXiv:1312.4400. https://doi.org/10.48550/arXiv.1312.4400
53. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539
54. Keel, P. K., & Brown, T. A. (2010). Update on course and outcome in eating disorders. International Journal of Eating Disorders, 43(3), 195-204. https://doi.org/10.1002/eat.20810
55. Swanson, S. A., Crow, S. J., Le Grange, D., Swendsen, J., & Merikangas, K. R. (2011). Prevalence and correlates of eating disorders in adolescents: Results from the national comorbidity survey replication adolescent supplement. Archives of General Psychiatry, 68(7), 714-723. https://doi. org/10.1001/archgenpsychiatry.2011.22
56. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
57. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.
58. Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363. https://arxiv.org/abs/1810.11363
59. Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019, November). Language Models as Knowledge Bases?. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 2463-2473).
60. Rubin, O., Herzig, J., & Berant, J. (2022, July). Learning To Retrieve Prompts for In-Context Learning. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 2655-2671).
61. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
62. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.
63. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., ... & Zaremba, W.

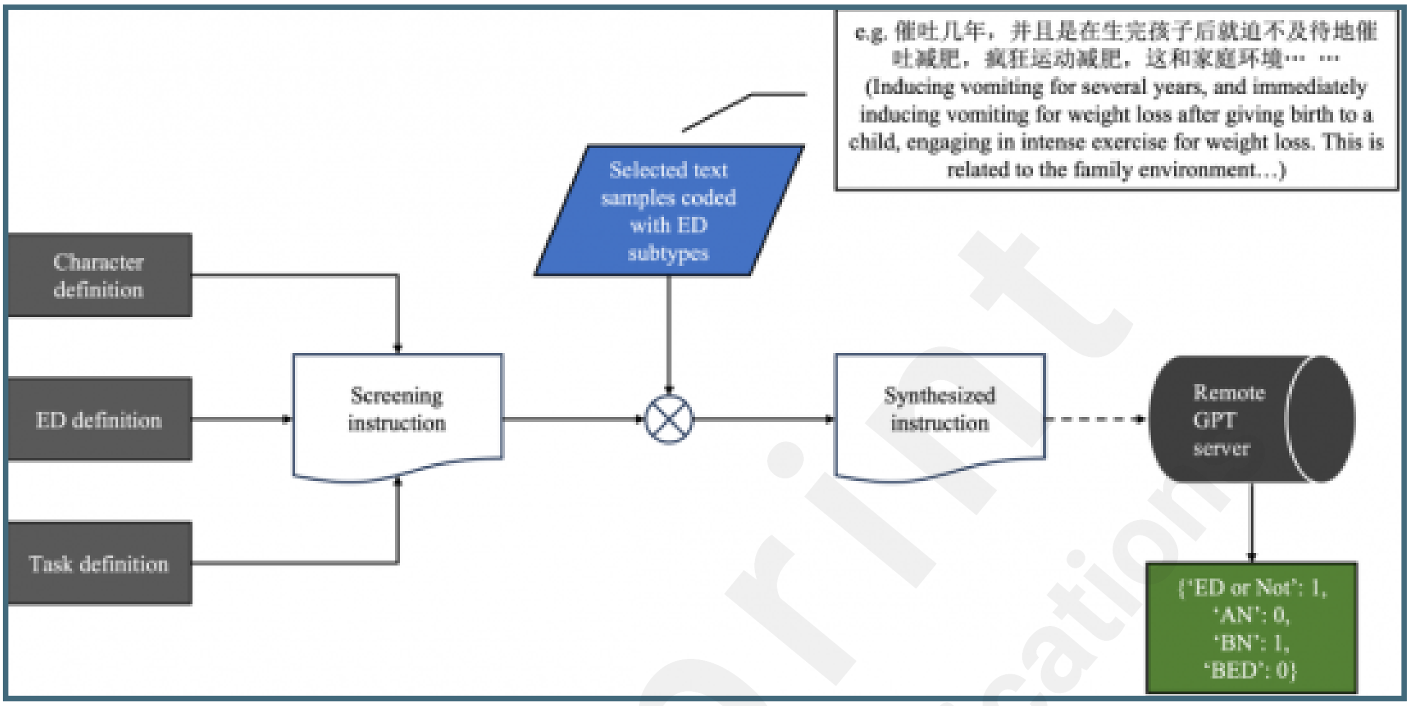
(2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374. https://doi.org/10.48550/arXiv.2107.03374

64. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785
65. Azad, R., Heidari, M., Wu, Y., & Merhof, D. (2022, September). Contextual attention network: Transformer meets u-net. In International Workshop on Machine Learning in Medical Imaging (pp. 377-386). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-21014-3_39
66. Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nature Computational Science, 1-6. https://doi.org/10.1038/s43588-023-00527-x
67. Collins, K. M., Wong, C., Feng, J., Wei, M., & Tenenbaum, J. (2022). Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 44, No. 44).
68. Christiansen, J. G., Gammelgaard, M., & Søgaard, A. (2023, November). Large language models partially converge toward human-like concept organization. In NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations.
69. Sowles, S. J., McLeary, M., Optican, A., Cahn, E., Krauss, M. J., Fitzsimmons-Craft, E. E., ... & Cavazos-Rehg, P. A. (2018). A content analysis of an online pro-eating disorder community on Reddit. Body Image, 24, 137-144. https://doi.org/10.1016/j.bodyim.2018.01.001
70. Fettach, Y., & Benhiba, L. (2019, December). Pro-eating disorders and pro-recovery communities on Reddit: text and network comparative analyses. In Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services (pp. 277-286). https://doi.org/10.1145/3366030.3366058

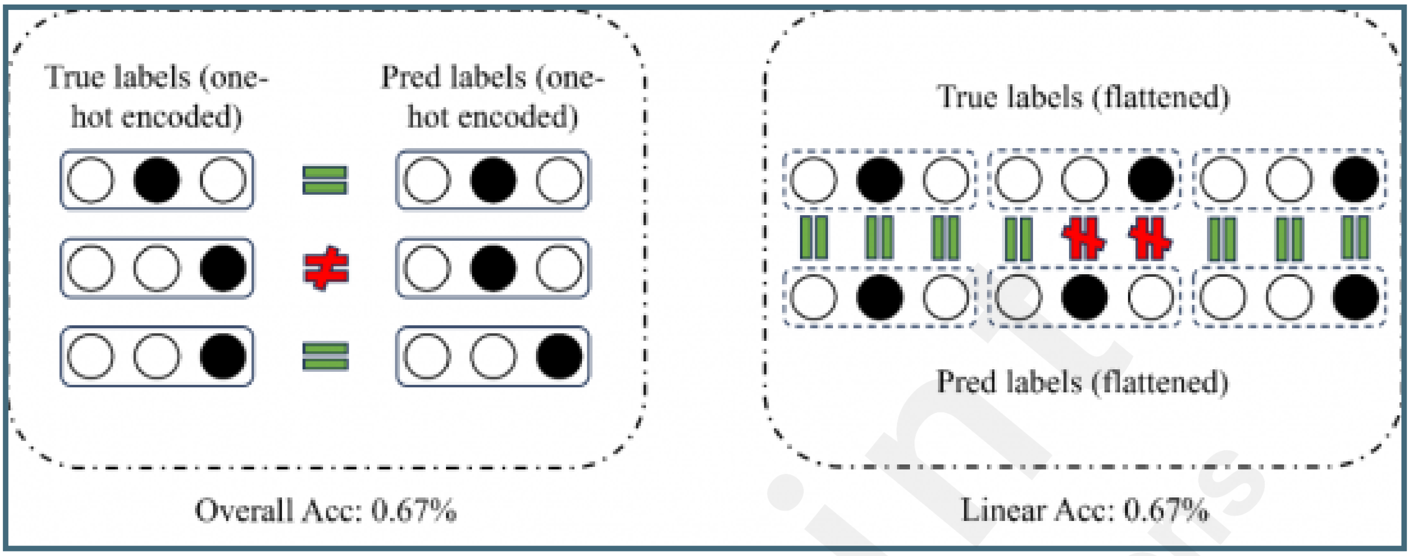
Supplementary Files

Figures

The schematic diagram of our proposed LLM-based text analysis scheme for eating disorders in a finite-state machine (FSM) style. Note: ED = Eating Disorder, AN = Anorexia Nervosa, BN = Bulimia Nervosa, BED = Binge-Eating Disorder.


e.g. 催吐几年，并且是在生完孩子后就迫不及待地催吐减肥，疯狂运动减肥，这和家庭环境… …
(Inducing vomiting for several years, and immediately inducing vomiting for weight loss after giving birth to a child, engaging in intense exercise for weight loss. This is related to the family environment…)
Selected text samples coded with ED subtypes
Character definition
ED definition
Task definition
Screening instruction
Synthesized instruction
Remote GPT server
{'ED or Not': 1,
'AN': 0,
'BN': 1,
'BED': 0}

A graphical illustration of how overall and linear accuracy are calculated and how they are different from each other. Acc. = Accuracy.


True labels (one-hot encoded)
Pred labels (one-hot encoded)
Overall Acc: 0.67%
True labels (flattened)
Pred labels (flattened)
Linear Acc: 0.67%