electronics

MDPI

Article

Using Large Language Models to Infer Problematic Instagram Use from User Engagement Metrics: Agreement Across Models and Validation with Self-Reports

Davide Marengo * and Michele Settanni

Department of Psychology, University of Turin, Via Verdi 10, 10124 Torino, Italy; michele.settanni@unito.it
* Correspondence: davide.marengo@unito.it

Abstract

This study investigated the feasibility of using large language models (LLMs) to infer problematic Instagram use, which refers to excessive or compulsive engagement with the platform that negatively impacts users' daily functioning, productivity, or well-being, from a limited set of metrics of user engagement in the platform. Specifically, we explored whether OpenAI's GPT-4o and Google's Gemini 1.5 Pro could accurately predict self-reported problematic use tendencies based solely on readily available user engagement metrics like daily time spent on the platform, weekly posts and stories, and follower/following counts. Our sample comprised 775 Italian Instagram users (61.6% female; aged 18–63), who were recruited through a snowball sampling method. Item-level and total scores derived by querying the LLMs' application programming interfaces were correlated with self-report items and the total score measured via an adapted Bergen Social Media Addiction Scale. LLM-inferred scores showed positive correlations with both item-level and total scores for problematic Instagram use. The strongest correlations were observed for the total scores, with GPT-4o achieving a correlation of r = 0.414 and Gemini 1.5 Pro achieving a correlation of r = 0.319. In cross-validated regression analyses, adding LLM-generated scores, especially from GPT-4o, significantly improved the prediction of problematic Instagram use compared to using usage metrics alone. GPT-4o's performance in random forest models was comparable to models trained directly on Instagram metrics, demonstrating its ability to capture complex, non-linear relationships indicative of addiction without needing extensive model training. This study provides compelling preliminary evidence for the use of LLMs in inferring problematic Instagram use from limited data points, opening exciting new avenues for research and intervention.

Keywords: large language models (LLMs); machine learning; data mining; social-media; problematic Instagram use

Academic Editor: Hung-Yu Chien

Received: 20 April 2025
Revised: 19 June 2025
Accepted: 20 June 2025
Published: 24 June 2025

Citation: Marengo, D.; Settanni, M. Using Large Language Models to Infer Problematic Instagram Use from User Engagement Metrics: Agreement Across Models and Validation with Self-Reports. *Electronics* **2025**, *14*, 2548. https://doi.org/10.3390/electronics14132548

1. Introduction

Advances in artificial intelligence (AI), particularly the development of large language models (LLMs) such as Google's Gemini and OpenAI's GPT model families, present unprecedented opportunities for unobtrusively inferring psychological states. These models are poised to revolutionize mental health care by leveraging raw data to improve diagnostics, monitoring, prevention, and treatment [1]. Practically, they offer a scalable and low-burden method to detect risk patterns from minimal, non-invasive inputs, which is an approach that may be especially valuable in resource-limited contexts [2]. Theoretically,

their ability to extract latent psychological signals from raw behavioral data challenges traditional assumptions about assessment, potentially advancing real-time, unobtrusive screening in digital mental health. Notably, LLMs can process a broad spectrum of data types, including both structured inputs (e.g., demographic variables and rating scales) and unstructured clinical text, such as electronic health records, discharge summaries, or psychotherapy transcripts. As such, they have been used to predict clinical outcomes, such as treatment response and risk of violent behavior, and to identify social determinants of health embedded in medical records [3,4]. Multimodal implementations that integrate text with neuroimaging or physiological data further extend their applicability in health-related applications [5–7]. Beyond clinical settings, LLMs have shown promise in analyzing social media data to assess mental health risks, including suicide ideation and exposure to violence [8,9], offering continuous, scalable monitoring in digital environments where traditional assessment is not feasible [10]. Recent research has demonstrated the impressive capacity of LLMs, such as Openai's GPT-3.5 and GPT-4, to accurately infer psychological dispositions, such as Big Five personality traits, from social media content like Facebook status updates [11]. Peters and Matz [11] found that GPT-3.5 and GPT-4 could infer personality traits with an accuracy comparable to supervised machine learning models, achieving correlations of 0.39 between LLM-inferred and self-reported scores. Other applications have demonstrated their potential to infer mental health conditions from online text data [12] and to identify risky alcohol consumption patterns based on social media activity [13].

The ability to extract psychological insights from user-generated text has significant implications for research. Among the potential avenues of research for LLMs is their use in the study of problematic social media use. The ever-growing rise of social media platforms has prompted extensive research into the behavioral patterns linked to their use, particularly focusing on problematic usage behaviors that resemble addiction. Problematic social media use involves excessive and compulsive engagement that disrupts an individual's daily life, productivity, relationships, and well-being, often resulting in addiction-like symptoms [14,15]. The literature frequently overlaps terms like "problematic social media use" and "social media addiction," with the two often being used interchangeably to describe similar patterns of problematic behavior [16]. However, the term "social media addiction" is not widely accepted in clinical or diagnostic contexts due to the absence of an official diagnostic category in standard diagnostic manuals such as the DSM-5 or ICD-11 [17]. Montag et al. [17] and other scholars advocate caution when employing terminology suggestive of addiction, emphasizing the importance of clear distinctions to avoid misconceptions about clinical significance and diagnostic criteria. Here, we deliberately employ the broader and more neutral term of "problematic social media use" to acknowledge the existing debate and avoid implications of clinical addiction that currently lack formal recognition.

Previous research has demonstrated the feasibility of using machine learning approaches to predict problematic smartphone and social media use by analyzing metrics of online activity, such as time spent on applications, posting frequency, and interaction metrics [18,19]. Such unobtrusive approaches could improve early detection and intervention strategies, particularly for populations at risk of social media addiction. Despite promising developments, the application of LLMs to infer addiction symptoms related to social media use based on limited usage metrics remains unexplored. This study investigates the feasibility of using LLMs to infer self-reported problematic Instagram use tendencies from a concise set of metrics of user engagement: daily time spent on the platform, weekly published posts and stories, and follower/following counts. By leveraging these data points, this research aims to evaluate how accurate LLM inferences (specifically, OpenAI's GPT-4o and Google's Gemini 1.5 Pro) are when compared to a traditional self-report as-

sessment using the Bergen Social Media Addiction Scale (BSMAS) [20,21], adapted for Instagram [22]. Data were collected from $n = 775$ adult Instagram users in Italy. Focusing on Instagram is particularly relevant given its sustained global popularity, with approximately 2 billion users as of February 2025, making it the third most used social media platform worldwide [23]. Given its broad reach, developing an automated and scalable method for detecting problematic usage patterns on Instagram holds substantial promise for both assessment and intervention. In this study, we describe the procedure used to prompt two large language models, GPT-4o and Gemini 1.5 Pro, to infer symptom scores related to problematic Instagram use based solely on platform engagement metrics. We present the distribution and internal consistency of the scores generated by each model, examine their associations with behavioral indicators and self-reported symptoms, and assess the level of agreement between the two models. Lastly, we evaluate the predictive value of these LLM-based inferences and consider their implications for scalable, unobtrusive approaches to digital mental health screening.

2. Materials and Methods

2.1. Procedure and Sample

Participants were recruited online through a snowball sampling approach. A survey link was shared via social media and private communications, with the initial seed consisting of six master's students in Psychology. To be eligible, participants had to be at least 18 years old, fluent in Italian, and active Instagram users. Anonymity was ensured by not recording personal information, such as names or IP addresses. The resulting sample comprised 775 adults aged 18 to 63 (M = 25.200, SD = 7.097), 61.6% of whom were female.

2.2. Measures

2.2.1. Problematic Instagram Use

We administered an adaptation of the Bergen Social Media Addiction Scale (BSMAS) [20,21], modified to refer specifically to Instagram [22]. The scale comprises six items addressing the following symptoms of problematic Instagram use: salience (e.g., "How often have you spent a lot of time thinking about or planning use of Instagram?"); mood modification (e.g., "How often have you used Instagram to forget about personal problems?"); tolerance (e.g., "How often during the last year have you felt an urge to use Instagram more and more?"); withdrawal (e.g., "How often have you become restless or troubled if you have been prohibited from using Instagram?"); conflict (e.g., "How often have you used Instagram so much that it has negatively impacted your job or studies?"); and relapse (e.g., "How often have you tried to cut down on your Instagram use without success?"). Each item was rated on a 5-point Likert scale (1 = very rarely, 5 = very often), and the total score was computed by summing the responses, with higher scores indicating more problematic use. Note that based on Cronbach's α [24], the scale demonstrated adequate internal consistency ($\alpha = 0.795$).

2.2.2. Instagram Usage Metrics

Participants provided their Instagram usage metrics by accessing the platform. The following metrics were collected: average daily time spent on Instagram (in minutes); the number of posts and stories published during the last week; and the total number of followers and accounts followed (i.e., the following count). On average, participants posted 0.58 (SD = 3.91) times per week and shared 4.36 (SD = 7.61) stories weekly on Instagram. Daily time spent on the platform averaged 95.13 min (SD = 130.00). Participants had a mean follower count of 675.24 (SD = 2048.53) and followed an average of 639.42 (SD = 577.04) accounts.

2.2.3. Leveraging Large Language Models for Inferring Problematic Instagram Use from Instagram Usage Metrics

To investigate the feasibility of using large language models (LLMs) for predicting problematic Instagram use, a structured, prompt-based approach was implemented. We used LLMs to analyze input data reflecting specific Instagram usage metrics and to generate symptom scores corresponding to the six addiction dimensions (i.e., salience, mood modification, tolerance, withdrawal, conflict, and relapse) outlined in the BSMAS, as adapted for Instagram. The process involved the use structured prompts to query OpenAI's GPT-4o (OpenAI, L.P., San Francisco, CA, USA) and Google's Gemini 1.5 Pro (Google LLC, Mountain View, CA, USA) via their respective application programming interfaces (APIs). These prompts guided the models to infer item-level scores on a 5-point Likert scale (1 = very rarely, 5 = very often). To ensure consistency and avoid potential confounds, all LLM predictions were generated using stateless prompting; each input was submitted in isolation, with no memory or context carried over between scoring operations. No examples, prior responses, or user-level history were retained or utilized during the inference process. Below is an example of the prompt used to operationalize the prediction task, including an example of user metric data submitted as part of the prompt:

> *"You are tasked with analyzing a dataset to infer a user's potential level of Instagram addiction. The dataset includes information about the average number of last week's published posts and stories, average daily time spent on the app, and current count of followers and following. Use this information to infer a symptom score on a 5-point scale (1 = Very rarely, 5 = Very often) for each of the six addiction dimensions:*
>
> *- Salience (Original Item: How often have you spent a lot of time thinking about or planned use of Instagram?)*
>
> *- Mood Modification (Original Item: How often have you used Instagram in order to forget about personal problems?)*
>
> *- Tolerance (Original Item: How often have you felt an urge to use Instagram more and more?)*
>
> *- Withdrawal (Original Item: How often have you become restless or troubled if you have been prohibited from using Instagram?)*
>
> *- Conflict (Original Item: How often have you used Instagram so much that it has had a negative impact on your job/study?)*
>
> *- Relapse (Original Item: How often have you tried to cut down on the use of Instagram without success?)*
>
> *USER DATA:*
>
> *- Weekly Posts: 0*
>
> *- Weekly Stories: 40*
>
> *- Time Spent on Instagram (average minutes/day): 165*
>
> *- Followers Count: 824*
>
> *- Following Count: 539*
>
> *Task: Generate the scores considering all input factors and output them in the specified format.*
>
> *Output Format:*
>
> *A comma-separated list of scores from 1 to 5 for each of the six symptoms."*

2.3. Data Analysis

To evaluate the psychometric properties and predictive utility of LLM-inferred scores, we conducted a series of analyses. Figure 1 illustrates the overall study design and the sequence of analyses performed. First, we assessed the internal consistency of the item-level scores generated by GPT-4o and Gemini 1.5 Pro using Cronbach's alpha. Total scores for each model were computed by summing the six symptom-specific ratings, following the same procedure used for the self-report scale. Descriptive statistics and score distributions were then examined for each item and total score, allowing for a comparison across GPT-4o, Gemini 1.5 Pro, and self-reports. To test systematic differences between LLM-inferred and self-reported scores, we conducted paired-sample t-tests for each symptom and the total score. Next, we assessed associations between Instagram usage metrics (e.g., time spent, posts, stories, and follower/following counts) and both LLM-inferred and self-reported scores using Pearson's correlations. This allowed us to evaluate the extent to which each score type was grounded in observable behavioral data. To examine the convergence between the two LLMs, we computed correlations between GPT-4o and Gemini 1.5 Pro for each item and total score. Additionally, we evaluated convergent validity by correlating each LLM's item-level and total scores with self-reported values on the adapted BSMAS.

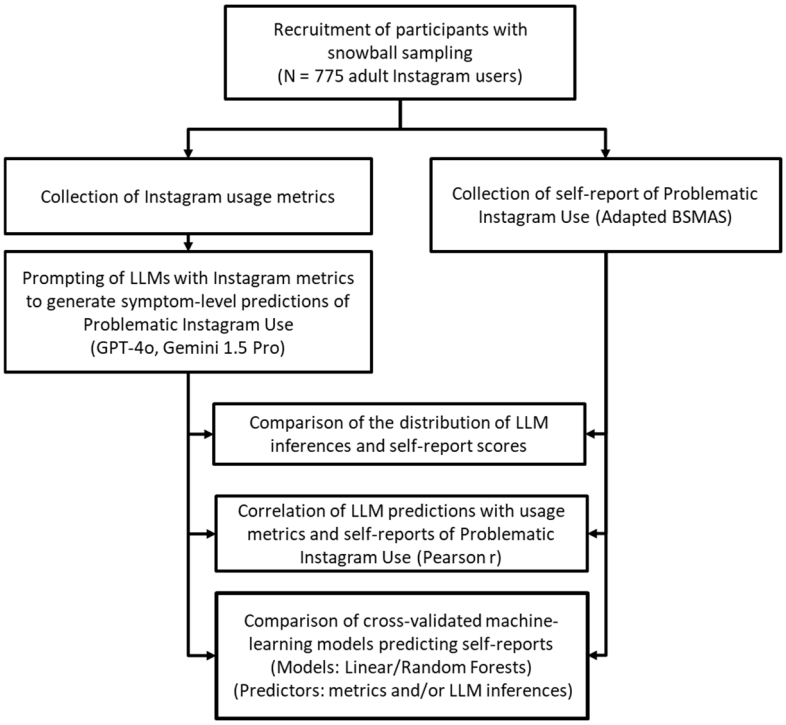

Figure 1. Overview of the study design and analytical procedures.

To further evaluate the LLMs, cross-validated regression and random forest models were applied to predict self-reported symptoms of problematic Instagram use based on a combination of both Instagram metrics and LLM-generated total scores for both GPT-4o and Gemini 1.5 Pro models. Specifically, 10-fold cross-validation was employed, with multiple linear regression models with no penalization and random forest [25] models based on 100 trees. The following combination of predictors were examined: (1) Instagram usage metrics; (2) Instagram usage metrics + GPT-4o total score; (3) Instagram usage metrics + Gemini 1.5 Pro total score; (4) Instagram usage metrics + GPT-4o total score + Gemini 1.5 Pro total score; (5) GPT-4o total score; (6) Gemini 1.5 Pro total score; (7) GPT-4o total score + Gemini 1.5 Pro total score. Model performance was evaluated based on the correlation (R) between self-reported and model-predicted scores, as well as the mean absolute error (MAE; i.e., the absolute difference in predicted and observed scores). These analyses were conducted using Weka [26].

3. Results

3.1. Distribution of LLM-Inferred and Self-Reported Scores for Problematic Instagram Use

Table 1 presents the distribution of item-level and total scores for problematic Instagram use, derived from prompts submitted to Gemini 1.5 Pro, GPT-4o, and self-reports. Preliminary analyses indicated excellent internal consistency for the item sets, with Cronbach's alpha values of 0.956 for items derived using GPT-4o and 0.953 for items derived using Gemini 1.5 Pro. To calculate total scores based on LLM-inferred item scores, we used a traditional summation method, simply adding up the individual item scores. Overall, GPT-4o consistently generated higher mean scores than Gemini 1.5 Pro for both item-level inferences and total scores. Self-reported scores, meanwhile, generally fell between the two LLM predictions, suggesting some alignment but also notable differences across scoring methods. Notably, all item and total-score comparisons between LLM-inferred and self-reported scores yielded statistically significant differences ($p < 0.05$), with two exceptions: *relapse*, as inferred by GPT-4o ($p = 0.570$), and *tolerance*, as inferred by Gemini ($p = 0.952$). Figure 2 provides a visualization of the distribution of the total scores for problematic Instagram use as obtained using self-reports and by prompting GPT-4o and Gemini 1.5 Pro.

Table 1. Descriptive statistics for LLM and self-reported scores for problematic Instagram use.

	GPT-4o			Gemini 1.5 Pro			Self-Report		
	M	**SD**	**Min–Max**	**M**	**SD**	**Min–Max**	**M**	**SD**	**Min–Max**
Items									
Salience	2.23	0.84	1–5	2.15	1.01	1–5	2.34	1.13	1–5
Tolerance	2.57	1.01	1–5	2.22	0.86	1–5	2.22	1.08	1–5
Mood modification	2.21	0.92	1–4	1.86	0.69	1–5	2.00	1.12	1–5
Relapse	1.83	0.80	1–5	1.24	0.66	1–5	1.81	1.07	1–5
Withdrawal	2.03	0.83	1–5	1.81	0.70	1–5	1.33	0.72	1–5
Conflict	1.44	0.57	1–4	1.33	0.73	1–5	1.73	1.01	1–5
Total Score	12.31	4.58	6–28	10.61	4.25	6–30	11.42	4.33	6–29

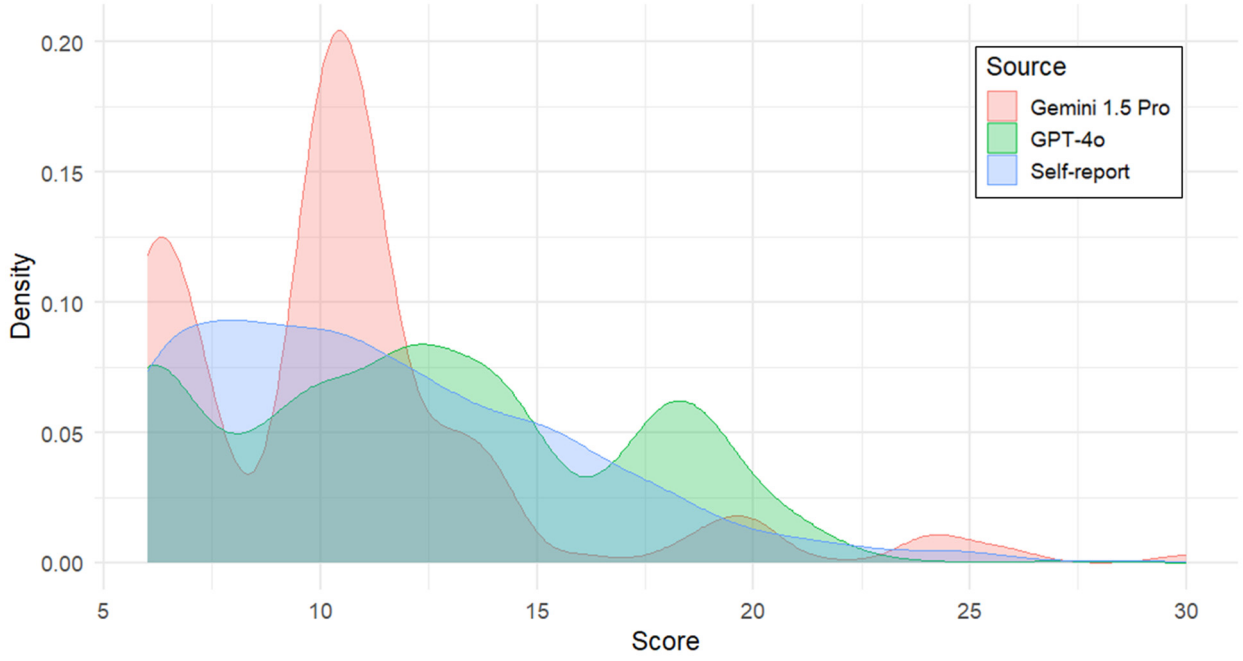

Figure 2. Comparison of score distributions between self-reported and LLM-based measures of problematic Instagram use.

3.2. Associations Between Instagram Usage Metrics and Self-Reported and Problematic Use Scores

The results of the correlation between Instagram usage metrics and LLM-inferred item and total scores for problematic Instagram use are reported in Table 2. Correlations were examined separately for GPT-4o, Gemini 1.5 Pro, and self-reported scores, across the five indicators used in the prompts: number of weekly posts, stories, time spent on Instagram, followers, and followings. As expected, LLM-inferred scores were strongly associated with the input usage metrics, particularly time spent on Instagram, which showed high correlations with total scores (r = 0.660 for GPT-4o; r = 0.808 for Gemini) and individual symptom dimensions (e.g., Withdrawal: r = 0.699 for GPT-4o; Conflict: r = 0.804 for Gemini). Moderate associations also emerged for the number of stories, while correlations with posts, followers, and following counts were generally lower. In contrast, self-reported scores showed much weaker associations with the same behavioral indicators. The strongest correlation was observed for time spent on Instagram (total score: r = 0.240).

Table 2. Correlation between Instagram metrics and item-level and total scores for problematic Instagram use obtained using LLMs and self-report.

		Number of Weekly Posts	Number of Weekly Stories	Time Spent on Instagram	Followers	Following
GPT-4o	Salience	**0.129**	**0.486**	**0.659**	**0.183**	**0.244**
	Tolerance	0.052	**0.387**	**0.594**	**0.141**	**0.272**
	Mood modification	−0.007	**0.460**	**0.447**	**0.113**	**0.241**
	Relapse	0.039	**0.460**	**0.610**	**0.144**	**0.257**
	Withdrawal	0.022	**0.390**	**0.699**	**0.125**	**0.212**
	Conflict	0.019	**0.473**	**0.661**	**0.134**	**0.241**
	Total Score	0.047	**0.479**	**0.660**	**0.153**	**0.268**
GEMINI 1.5 Pro	Salience	**0.185**	**0.400**	**0.716**	**0.208**	**0.219**
	Tolerance	**0.103**	**0.380**	**0.677**	**0.159**	**0.184**
	Mood modification	**0.123**	**0.403**	**0.739**	**0.213**	**0.204**
	Relapse	**0.169**	**0.379**	**0.757**	**0.231**	**0.119**
	Withdrawal	**0.111**	**0.376**	**0.750**	**0.210**	**0.197**
	Conflict	**0.073**	**0.315**	**0.804**	**0.156**	**0.083**
	Total Score	**0.142**	**0.413**	**0.808**	**0.214**	**0.188**
Self-report	Salience	0.068	**0.244**	**0.232**	**0.093**	**0.141**
	Tolerance	0.063	**0.223**	**0.200**	0.038	**0.101**
	Mood modification	0.032	**0.088**	**0.164**	0.040	0.041
	Relapse	0.025	0.030	**0.141**	−0.013	0.031
	Withdrawal	**0.098**	**0.169**	**0.099**	0.022	**0.111**
	Conflict	0.043	**0.094**	**0.154**	0.022	**0.109**
	Total Score	**0.074**	**0.200**	**0.240**	0.050	**0.124**

Note. Correlations in bold are significant at $p < 0.05$.

3.3. Cross-Model Agreement in LLM-Inferred Scores for Problematic Instagram Use

Correlations between the item-level and total scores generated by GPT-4o and Gemini 1.5 Pro indicated a substantial degree of agreement between the two LLMs in their assessments of problematic Instagram use based on the provided usage metrics. Notably, the strongest agreement was observed between the scores for the items assessing salience (r = 0.792, $p < 0.001$), tolerance (r = 0.795, $p < 0.001$), and withdrawal (r = 0.761, $p < 0.001$) symptoms, while substantially lower (but still strong) linear agreement was observed for conflict (r = 0.617, $p < 0.001$), mood modification (r = 0.597, $p < 0.001$), and relapse (r = 0.526, $p < 0.001$). The summed total scores were highly convergent across the two models (r = 0.803, $p < 0.001$).

3.4. Concurrent Validity Between LLM-Inferred and Self-Reported Scores of Problematic Instagram Use

Table 3 displays the correlations between LLM-inferred item-level and total scores based on Gemini 1.5 Pro and GPT-4o outputs and the corresponding self-reported problematic Instagram use scores. The correlations between self-reported and LLM-inferred scores obtained from both models were small to moderate and positive, ranging from 0.169 (withdrawal) to 0.414 (total score) for GPT4o and from 0.111 (relapse) to 0.319 (total score) for Gemini 1.5 Pro.

Table 3. Correlations between self-reported and LLM-inferred symptoms and total score for problematic Instagram use (N = 775).

Self-Report	Gemini 1.5 Pro	GPT-4o
Items		
Salience	0.303 ($p < 0.001$)	0.387 ($p < 0.001$)
Tolerance	0.269 ($p < 0.001$)	0.336 ($p < 0.001$)
Mood modification	0.228 ($p < 0.001$)	0.206 ($p < 0.001$)
Relapse	0.111 ($p = 0.001$)	0.206 ($p < 0.001$)
Withdrawal	0.145 ($p < 0.001$)	0.169 ($p < 0.001$)
Conflict	0.129 ($p < 0.001$)	0.243 ($p < 0.001$)
Total Score	0.319 ($p < 0.001$)	0.414 ($p < 0.001$)

3.5. Incremental Validity of LLM Inferences over Instagram Usage Metrics in Predicting Self-Reported Problematic Instagram Use

Figure 3 presents cross-validation metrics for linear regression and random forest models that predict problematic Instagram use. These models compare the predictive value of usage metrics alone, LLM-inferred inferences from GPT-4o and Gemini 1.5 Pro, and their combinations. In linear regression models, the usage metrics on their own resulted in modest predictive performance, with a correlation of 0.189 between predicted and self-reported scores and relatively high error values (MAE = 3.401). Adding GPT-4o inferences to the model significantly enhanced prediction accuracy, increasing the correlation to 0.348 and reducing errors (MAE = 3.189). Gemini 1.5 Pro inferences also improved performance but to a lesser extent, with a correlation of 0.243 and slightly higher error metrics (MAE = 3.345). Combining inferences from both LLMs showed a marginally lower correlation (R = 0.344) compared to using GPT-4o alone, with similar error values. Notably, when used independently, GPT-4o predictions outperformed all configurations, achieving the best results, with a correlation of 0.407 between predicted and self-reported scores, and the lowest error metrics (MAE = 3.149). Gemini 1.5 Pro predictions alone produced weaker performance, with a correlation of 0.309 and more errors (MAE = 3.314).

Random forest models provided a different picture. When using only usage metrics, the models achieved stronger predictive performance compared to linear regression, with a correlation of 0.391 and improved error metrics (MAE = 3.173). Adding GPT-4o or Gemini 1.5 Pro inferences to these models did not lead to noticeable improvement. Correlations remained similar, ranging from 0.379 to 0.391, with minimal changes in MAE and RMSE values. Using GPT-4o predictions alone in random forest models resulted in moderate performance, with a correlation of 0.333 and slightly higher error metrics (MAE = 3.251). Similarly, Gemini 1.5 Pro predictions alone yielded weaker results, with a correlation of 0.317 and even higher error metrics (MAE = 3.282). Overall, the linear regression model utilizing GPT-4o inferences alone demonstrated superior performance, achieving the highest correlation (r = 0.407) and lowest error rate (MAE = 3.149) in predicting problematic Instagram use compared to all other configurations.

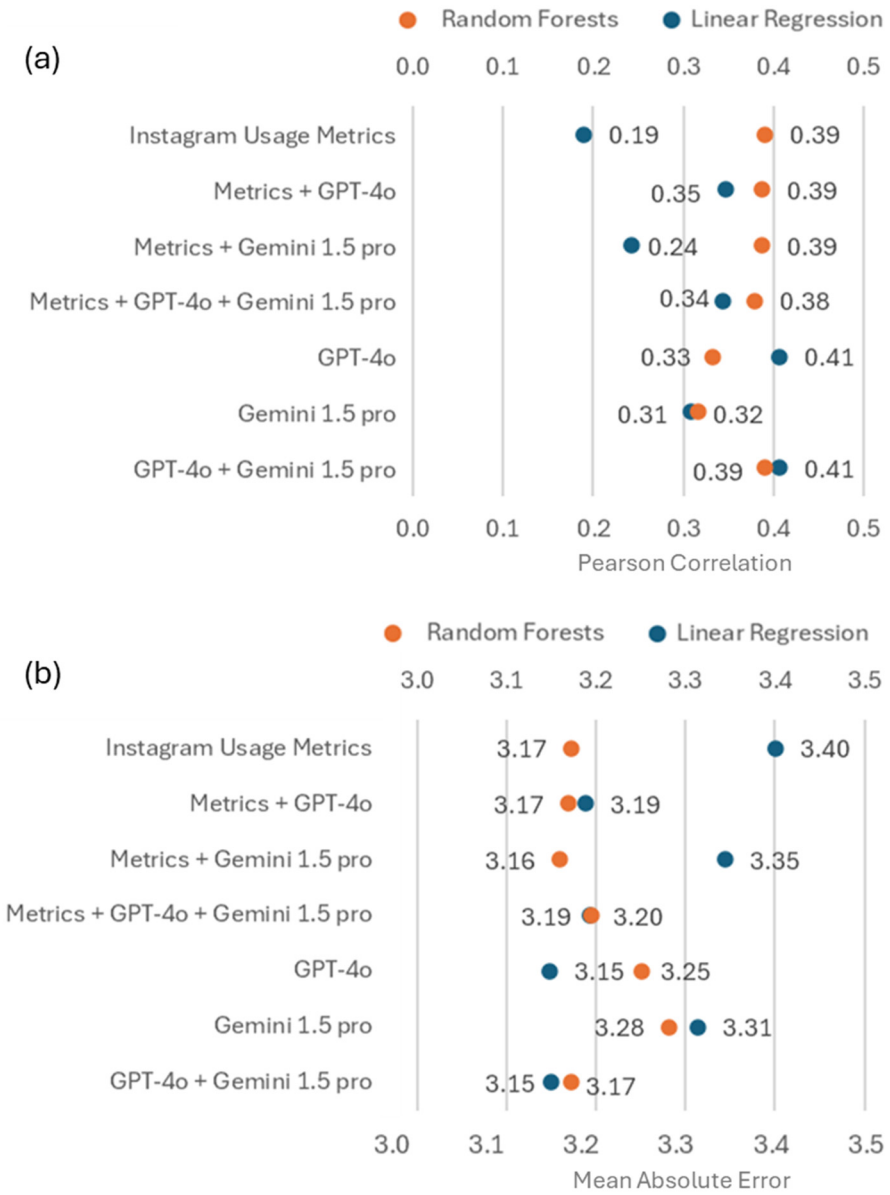

Figure 3. Cross-validation metrics for linear regression and random forest models in predicting problematic Instagram use; (**a**) Pearson correlations between predicted and observed scores across 10-fold cross-validation; (**b**) Mean absolute error (MAE) of predictions across 10-fold cross-validation.

4. Discussion

This study explored the capacity of two large language models (LLMs), GPT-4o and Gemini 1.5 Pro, to infer tendencies toward problematic Instagram use based on a limited set of Instagram user metrics. The results provide compelling evidence for the potential of LLMs to infer psychological states, as represented by items assessing symptoms of problematic Instagram use, from limited indicators of user engagement in the platform. Both LLMs demonstrated substantial agreement in their assessments of problematic Instagram use, with correlations ranging from moderate to strong. Notably, the strongest agreement was observed between the scores for the items assessing salience, tolerance, and withdrawal symptoms, while substantially lower, although still strong, agreement was found for conflict, mood modification, and relapse. Additionally, item-level inferences by both GPT-4o and Gemini 1.5 Pro demonstrated high internal consistency ($\alpha \geq 0.90$), providing preliminary support for the feasibility of generating total scores. The resulting total scores also demonstrated a high level of convergence across the two models ($r = 0.80$).

This strong inter-LLM agreement suggests a consistent interpretation of usage patterns across different models.

Of note, significant discrepancies were observed in the score distributions for LLM inferences and self-report. Specifically, GPT-4o tended to generate higher overall scores, while Gemini produced more conservative estimates. Self-reported problematic Instagram use scores generally fell between the two LLM outputs. These differences may reflect the distinct nature of the data sources that each method relies on. LLMs derive their inferences solely from observable behavioral patterns, such as posting frequency or time spent on the platform, without access to subjective experiences, contextual factors, or biases that shaped participants' self-reported assessments.

In spite of the aforementioned discrepancies, both LLM-inferred scores showed positive correlations with self-reported symptoms of problematic Instagram use. The strongest correlations were observed for the total scores, with inferences made by GPT-4o and Gemini 1.5 Pro achieving correlations of $r = 0.414$ and $r = 0.318$, respectively, with self-reported problematic Instagram use. These findings reveal a meaningful relationship between LLM-inferred scores and users' self-reported Instagram use behaviors, particularly in capturing overall addiction tendencies. While these correlations are not strong, they indicate a meaningful relationship between LLM inferences and users' own perceptions of their tendency toward problematic Instagram use. Notably, the observed correlation of $r = 0.414$ between GPT-4o predictions and self-reported scores falls at the lower bound of test–retest reliability estimates for the BSMAS, which range from 0.42 to 0.53 over a one-year period [27]. It is, however, lower than the stability reported over a shorter three-month interval ($r = 0.72$) [28]. Note that the BSMAS is designed to capture generalized problematic social media use rather than Instagram-specific symptoms, as in the adaptation used in our study [22]; still, this comparison underscores the practical relevance of our findings, suggesting that LLM-based inferences may approach the psychometric reliability of established self-report instruments, especially when applied as scalable, indirect screening tools. Moreover, the emerging correlations are on par with those emerging from previous studies employing machine learning approaches to predict problematic social media use from Facebook activity data [19]; the observed effect sizes are also in line with the recent study by Peters and Matz (2024) [11] that leveraged LLMs to infer the psychological dispositions of social media users from social media data (i.e., a correlation ranging from 0.2 to 0.4 between observed and predicted personality scores).

Next, cross-validated predictive analyses revealed that incorporating LLM-generated scores significantly enhanced the prediction of self-reported problematic use beyond what could be achieved with objective usage metrics alone. When using regression, the inclusion of LLM-generated scores, particularly those derived from GPT-4o, significantly improved the prediction of self-reported addiction scores beyond that obtained using the set of Instagram metrics as sole predictors. Moreover, the linear regression model that utilized GPT-4o inferences alone demonstrated superior performance, achieving the highest correlation and lowest error rate in predicting problematic Instagram use compared to all other configurations. This finding is particularly noteworthy given that the LLM scores were generated entirely from the same five metrics. This suggests that LLMs can extract and process information relevant to addiction in a way that goes beyond what is captured by a traditional linear regression analysis of the raw metrics. Notably, in this context, the similarity in performance between the random forest models, one using only raw usage metrics and the other using GPT-4o inferences derived from the same metrics, suggests that the GPT-4o inferences capture non-linearities and complex patterns that are otherwise only accessible through extensive training on the data. This highlights how pretrained models like GPT-4o can encapsulate domain-specific non-linearities and interactions without the need

for additional training, making them highly valuable in contexts in which computational or data resources are limited for model training. However, when comparing these results to the weaker performance observed with Gemini's inferences, it becomes evident that not all pretrained models are equally adept at capturing and encoding these complex non-linear relationships. This discrepancy may stem from differences in the training objectives, data scales, or architectures of the two models.

Note that in selecting benchmark models for comparison with LLM-generated scores, our intent was not to demonstrate that classical machine learning algorithms can match the performance of LLMs but rather to assess whether LLM inferences, obtained using a zero-shot approach, without fine-tuning, could approximate the predictive accuracy of models explicitly trained on outcome data. Our goal was not to determine which model performs best but to evaluate the viability of LLMs as lightweight, scalable alternatives to trained predictors in contexts in which training data may be limited or unavailable. To this end, we employed linear regression and random forest models, which are widely adopted in behavioral science and social media analytics; linear regression captures linear relationships and is highly transparent, while random forest can account for non-linear interactions and is robust to overfitting, even in small- to medium-sized datasets [29]. We acknowledge, however, that LLMs operate differently from these traditional algorithms, relying on massive pretraining to implicitly encode semantic and behavioral associations [30]. As such, our comparison represents a conservative test: it evaluates whether pretrained LLMs can generate psychologically meaningful predictions without task-specific training. Including models such as recurrent neural network architectures (e.g., Long Short-Term Memory models) would be a valuable direction for future work.

Finally, the use of LLMs to infer psychological states from behavioral data also raises important ethical considerations. Chief among these are concerns about privacy, the potential misuse of predictions, and the risk of stigmatization. As noted by Malgaroli et al. (2025) [1], the application of LLMs in mental health contexts calls for a robust ethical framework that ensures transparency, explainability, and user autonomy. In this study, ethical risks were mitigated by using fully anonymized data with no personally identifiable information or timestamps, as well as by excluding any textual content. These measures reduced the likelihood of re-identification and behavioral surveillance. Moving forward, ethical LLM use in psychological contexts should be guided by strict data governance, informed consent procedures, and clear boundaries around how predictions can be applied.

Several limitations should be considered. The reliance on self-reported usage data, while common in this field, introduces the potential for inaccuracies. The reliance on a snowball sampling method may limit generalizability. Indeed, while effective in reaching a large number of respondents quickly, this approach may introduce sampling bias due to the tendency of participants to recruit others within their own social or demographic circles [31]. Future research should consider employing stratified or quota sampling strategies to ensure more diverse and representative samples across demographic and behavioral dimensions. Furthermore, the cross-sectional nature of the study prevents conclusions about causality. Future research should incorporate more objective behavioral measures and longitudinal designs to assess changes in usage patterns and addiction symptoms over time and delve deeper into the mechanisms by which LLMs derive their predictions. Exploring these mechanisms could shed light on the underlying processes that contribute to problematic social media use. Finally, consistent with previous research examining the use of social media data to infer psychological dispositions (e.g., [11]), we employed a zero-shot prompting strategy to evaluate the baseline capabilities of LLMs in detecting symptoms of problematic Instagram use. This approach allowed us to assess model performance without providing annotated examples or task-specific fine-tuning, thereby offering a con-

servative and broadly generalizable benchmark. However, the absence of guidance may have constrained model accuracy, particularly for items requiring nuanced interpretation. Future research could investigate the benefits of enhanced prompting techniques, such as few-shot prompting [2], which may improve inference quality.

5. Conclusions

Despite these limitations, this study offers compelling preliminary evidence for the utility of LLMs in inferring problematic Instagram use. The findings suggest that LLMs, particularly GPT-4o, can provide valuable insights beyond simple usage metrics, thereby offering a potentially scalable and unobtrusive method for assessing and addressing problematic social media use. Among the key contributions of this study are the demonstration that LLMs can generate symptom-level inferences with strong internal consistency, the computation of total scores that show promising convergence with self-reports, and the finding that LLM-derived scores, especially from GPT-4o, significantly enhance the prediction of self-reported problematic use beyond behavioral metrics alone.

However, our analyses also revealed systematic differences between LLM-inferred and self-reported scores, underscoring important limitations in how LLMs interpret behavioral cues in the absence of subjective context. These findings emphasize the need for continued investigation into how different LLMs process and represent behavioral patterns. In particular, future research should consider prompting strategies that allow LLMs to infer problematic usage patterns more autonomously (e.g., chain of thought prompting [32]), potentially uncovering how models internally conceptualize behavioral dysfunction beyond predefined symptom structures.

In sum, the approach presented here holds promise for advancing the understanding of online behavior and developing targeted interventions for individuals at risk. Pretrained LLMs could serve as lightweight, low-burden tools for inferring behavioral health risks in digital contexts, particularly in settings in which annotated data are unavailable. Future research might focus on refining these models and validating their accuracy in diverse populations and social media platforms.

Author Contributions: Conceptualization, D.M. and M.S.; methodology, D.M. and M.S.; formal analysis, D.M.; data curation, D.M.; validation, M.S.; writing—original draft preparation, D.M.; writing—review and editing, D.M. and M.S.; visualization, D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data can be shared by the authors upon reasonable request.

Acknowledgments: During the preparation of this work, the authors employed OpenAI's ChatGPT to enhance language and readability. The author(s) reviewed and edited the output as necessary and assume full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Malgaroli, M.; Schultebraucks, K.; Myrick, K.J.; Loch, A.A.; Ospina-Pinillos, L.; Choudhury, T.; Kotov, R.; De Choudhury, M.; Torous, J. Large language models for the mental health community: Framework for translating code to care. *Lancet Digit. Health* **2025**, *7*, e282–e285. [CrossRef] [PubMed]
2. Volkmer, S.; Meyer-Lindenberg, A.; Schwarz, E. Large language models in psychiatry: Opportunities and challenges. *Psychiatry Res.* **2024**, *339*, 116026. [CrossRef] [PubMed]
3. Han, S.; Zhang, R.F.; Shi, L.; Richie, R.; Liu, H.; Tseng, A.; Quan, W.; Ryan, N.D.; Brent, D.A.; Tsui, F. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *J. Biomed. Inform.* **2022**, *127*, 103984. [CrossRef]

4. Mosteiro, P.; Rijcken, E.; Zervanou, K.; Kaymak, U.; Scheepers, F.; Spruit, M. Machine learning for violence risk assessment using Dutch clinical notes. *arXiv* **2022**, arXiv:2204.13535. [CrossRef]
5. Jeong, J.; Tian, K.; Li, A.; Hartung, S.; Adithan, S.; Behzadi, F.; Calle, J.; Osayande, D.; Pohlen, M.; Rajpurkar, P. Multimodal image-text matching improves retrieval-based chest X-ray report generation. *arXiv* **2023**, arXiv:2303.17579.
6. Jiang, L.Y.; Liu, X.C.; Nejatian, N.P.; Nasir-Moin, M.; Wang, D.; Abidin, A.Z.; Eaton, K.; Riina, H.A.; Laufer, I.; Punjabi, P.P.; et al. Health system-scale language models are all-purpose prediction engines. *Nature* **2023**, *619*, 357–362. [CrossRef]
7. Tiu, E.; Talius, E.; Patel, P.; Langlotz, C.P.; Ng, A.Y.; Rajpurkar, P. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* **2022**, *6*, 1399–1406. [CrossRef]
8. Al-Garadi, M.A.; Kim, S.; Guo, Y.; Warren, E.; Yang, Y.-C.; Lakamana, S.; Sarker, A. Natural language model for automatic identification of intimate partner violence reports from Twitter. *Array* **2022**, *15*, 100217. [CrossRef]
9. Xu, X.; Yao, B.; Dong, Y.; Gabriel, S.; Yu, H.; Hendler, J.; Ghassemi, M.; Dey, A.K.; Wang, D. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2024**, *8*, 1–32. [CrossRef]
10. Elyoseph, Z.; Levkovich, I. Beyond human expertise: The promise and limitations of ChatGPT in suicide risk assessment. *Front. Psychiatry* **2023**, *14*, 1213141. [CrossRef]
11. Peters, H.; Matz, S.C. Large language models can infer psychological dispositions of social media users. *PNAS Nexus* **2024**, *3*, 231. [CrossRef] [PubMed]
12. Settanni, M.; Quilghini, F.; Toscano, A.; Marengo, D. Assessing the Accuracy and Consistency of Large Language Models in Triaging Social Media Posts for Psychological Distress. *Psychiatry Res.* **2025**, *351*, 116583. [CrossRef] [PubMed]
13. Marengo, D.; Quilqhini, F.; Settanni, M. Leveraging social media and large language models for scalable alcohol risk assessment: Examining validity with AUDIT-C and post recency effects. *Addict. Behav.* **2025**, *168*, 108375. [CrossRef]
14. Griffiths, M.D.; Kuss, D.J.; Demetrovics, Z. Social networking addiction: An overview of preliminary findings. In *Behavioral Addictions*; Rosenberg, K.P., Feder, L.C., Eds.; Academic Press: Cambridge, MA, USA, 2014; pp. 119–141. [CrossRef]
15. Kuss, D.J.; Griffiths, M.D. Social networking sites and addiction: Ten lessons learned. *Int. J. Environ. Res. Public Health* **2017**, *14*, 311. [CrossRef]
16. Sun, Y.; Zhang, Y. A review of theories and models applied in studies of social media addiction and implications for future research. *Addict. Behav.* **2021**, *114*, 106699. [CrossRef] [PubMed]
17. Montag, C.; Demetrovics, Z.; Elhai, J.D.; Grant, D.; Koning, I.; Rumpf, H.-J.; Spada, M.M.; Throuvala, M.; Van den Eijnden, R. Problematic social media use in childhood and adolescence. *Addict. Behav.* **2024**, *153*, 107980. [CrossRef] [PubMed]
18. Marengo, D.; Sariyska, R.; Schmitt, H.S.; Messner, E.-M.; Baumeister, H.; Brand, M.; Kannen, C.; Montag, C. Exploring the associations between self-reported tendencies toward smartphone use disorder and objective recordings of smartphone, instant messaging, and social networking app usage: A correlational study. *J. Med. Internet Res.* **2021**, *23*, e27093. [CrossRef]
19. Marengo, D.; Montag, C.; Mignogna, A.; Settanni, M. Mining digital traces of Facebook activity for the prediction of individual differences in tendencies toward social networks use disorder: A machine learning approach. *Front. Psychol.* **2022**, *13*, 830120. [CrossRef]
20. Andreassen, C.S.; Torsheim, T.; Brunborg, G.S.; Pallesen, S. Development of a Facebook addiction scale. *Psychol. Rep.* **2012**, *110*, 501–517. [CrossRef]
21. Monacis, L.; De Palo, V.; Griffiths, M.D.; Sinatra, M. Social networking addiction, attachment style, and validation of the Italian version of the Bergen Social Media Addiction Scale. *J. Behav. Addict.* **2017**, *6*, 178–186. [CrossRef]
22. Marengo, D.; Mignogna, A.; Elhai, J.D.; Settanni, M. Distinguishing high engagement from problematic symptoms in Instagram users: Associations with big five personality, psychological distress, and motives in an Italian sample. *Cyberpsychol. J. Psychosoc. Res. Cyberspace* **2024**, *18*, 4. [CrossRef]
23. Statista Research Department. Most Popular Social Networks Worldwide as of February 2025, by Number of Monthly Active Users. Available online: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed on 19 June 2025).
24. Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika* **1951**, *16*, 297–334. [CrossRef]
25. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
26. Frank, E.; Hall, M.A.; Witten, I.H. *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed.; Morgan Kaufmann: Burlington, MA, USA, 2016.
27. Gomez, R.; Zarate, D.; Brown, T.; Hein, K.; Stavropoulos, V. The Bergen–Social Media Addiction Scale (BSMAS): Longitudinal measurement invariance across a two-year interval. *Clin. Psychol.* **2024**, *28*, 185–194. [CrossRef]
28. Chen, I.H.; Strong, C.; Lin, Y.C.; Tsai, M.C.; Leung, H.; Lin, C.Y.; Pakpour, A.H.; Griffiths, M.D. Time invariance of three ultra-brief internet-related instruments: Smartphone application-based addiction scale (SABAS), Bergen social media addiction scale (BSMAS), and the nine-item internet gaming disorder scale-short form (IGDS-SF9) (study Part B). *Addict. Behav.* **2020**, *101*, 105960. [CrossRef]

29. Fife, D.A.; D'Onofrio, J. Common, uncommon, and novel applications of random forest in psychological research. *Behav. Res. Methods* **2023**, *55*, 2447–2466. [CrossRef]
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
31. Atkinson, R.; Flint, J. Accessing hidden and hard-to-reach populations: Snowball research strategies. *Soc. Res. Update* **2001**, *33*, 1–4.
32. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.