Are LLMs effective psychological assessors? Leveraging adaptive RAG for interpretable mental health screening through psychometric practice

Federico Ravenda[1,2], **Seyed Ali Bahrainian**[3,4],
Andrea Raballo[1], **Antonietta Mira**[1,5], **Noriko Kando**[2]
federico.ravenda@usi.ch, bahrainian@brown.edu
andrea.raballo@usi.ch, antonietta.mira@usi.ch, kando@nii.ac.jp
[1]Università della Svizzera italiana, [2]National Institute of Informatics
[3]Brown University, [4]University of Tübingen, [5]Insubria University

Abstract

In psychological practice, standardized questionnaires serve as essential tools for assessing mental constructs (e.g., attitudes, traits, and emotions) through structured questions (aka items). With the increasing prevalence of social media platforms where users share personal experiences and emotions, researchers are exploring computational methods to leverage this data for rapid mental health screening. In this study, we propose a novel adaptive Retrieval-Augmented Generation (RAG) approach that completes psychological questionnaires by analyzing social media posts. Our method retrieves the most relevant user posts for each question in a psychological survey and uses Large Language Models (LLMs) to predict questionnaire scores in a zero-shot setting. Our findings are twofold. First we demonstrate that this approach can effectively predict users' responses to psychological questionnaires, such as the Beck Depression Inventory II (BDI-II), achieving performance comparable to or surpassing state-of-the-art models on Reddit-based benchmark datasets without relying on training data. Second, we show how this methodology can be generalized as a scalable screening tool, as the final assessment is systematically derived by completing standardized questionnaires and tracking how individual item responses contribute to the diagnosis, aligning with established psychometric practices [1].

1 Introduction

According to the World Health Organization (WHO), one in seven adolescents experiences a mental health disorder (Wiederhold, 2022), with depression, anxiety, and behavioral disorders leading among young people. Following COVID-19, mental health conditions surged, with depressive disorders increasing by 28% in 2020 (Kieling et al., 2011; Winkler et al., 2020). Given the extent of this need, the WHO Special Initiative for Mental Health prioritizes improving and expanding access to quality mental health interventions and services as a key strategic goal.[2]

Psychological questionnaires play a crucial role in describing mental states by measuring various psychological constructs, as outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM) (Hopwood et al., 2012). The conventional interpretation of data derived from psychometric scales assumes that obtained scores reflect the intensity of a respondent attitudes. Psychological questionnaires can be used to assess various constructs related to different mental disorders as screening methods to develop an initial clinical profile.

In this work, we focus on two widely used standardized psychological questionnaires, specifically the BDI-II (Beck et al., 1996) for depression screening and the Self-Harm Inventory (SHI) (Sansone and Sansone, 2010) for self-harm behaviour detection. Both are self-reported surveys where overall scores correspond to specific severity levels of the respective conditions.

Increasingly, people turn to social media as a space to discuss their feelings and experiences and to find support (Bucci et al., 2019; Naslund et al., 2016). Numerous initiatives have emerged to analyze social media content for health monitoring using NLP techniques, including CLPsych (Tsakalidis et al., 2022) and eRisk (Losada et al., 2017), through organized completion tasks. In this work, we use datasets from previous eRisk editions to validate our approach.

Recently, several studies have shown that closed-source LLMs struggle to achieve comparable results to state-of-the-art (SOTA) supervised methods in mental disorders classification tasks, both in zero-shot (Amin et al., 2023) and few-shot (XU

[1]Code available at the following link: github.com/Fede-stack/Adaptive-RAG-for-Psychological-Assessment

[2]https://www.who.int/publications/i/item/9789240049338

et al., 2023) scenarios. We argue that it is possible to mitigate this issue by introducing an intermediate step where we instruct an LLM to answer questions from a standardized psychological questionnaire rather than directly requesting a diagnosis.

To do that, we implement an adaptive RAG approach, combining retrieval and classification, to accurately predict users responses to psychological questionnaire items by analyzing their Reddit post history. Our goal is to test the capability of SOTA LLMs, both open- and closed-source, as potential annotators of standardized psychological questionnaires based on social media posts.

The main contributions of this paper are: **(1)** We explore various combinations of open- and closed-source LLMs together with dense retrievers for predicting psychological questionnaires, evaluating how performance varies with different combinations of LLMs, prompt strategies, and retrieval models. **(2)** We compare our approach's results with the best results obtained for the considered eRisk collections using primarily supervised models, showing how our unsupervised approach often outperforms the benchmarks. **(3)** We extend this paradigm to other mental disorders, introducing an interpretable and unsupervised method for predicting new behavioral disorders.

2 Related Works

Recent advancements in NLP have enabled the development of new and complex models across various areas, particularly in digital and mental health. In recent years, transformer-based models (Vaswani et al., 2017) have emerged as powerful tools for predicting mental health disorders, leveraging vast amounts of textual data available from social media platforms. (Pourkeyvan et al., 2024; Raj et al., 2024) employed different BERT models for depression detection on social media platforms, while (Ji et al., 2022) developed MentalBERT, a specialized BERT model pre-trained on mental health-related content from Reddit, specifically designed to capture language patterns associated with mental disorders. Regarding LLMs, (Yang et al., 2024a) introduced MentaLLaMA, an open source LLM specifically designed for interpretable mental health analysis on social media, building a multitask and multi-source instruction dataset to support model fine-tuning and evaluation. Furthermore, (Yang et al., 2024b) presented a comprehensive analysis on how LLMs can enhance mental health care through real-time feedback and early intervention.

Concerning the use of NLP methods to predict psychological questionnaires' responses, previous works (Elourajini and Aïmeur, 2022; Vu et al., 2020) used neural models to predict the 5 main personality traits (Goldberg, 2013) and the Myers-Briggs type indicator (MBTI) (Schweiger, 1985; Yang et al., 2021) based on user-generated textual posts and comments. (Rosenman et al., 2024) propose an approach in which an LLM impersonates the interviewee, completing structured questionnaires, and then these responses are encoded as features to train a Random Forest model to predict the scores of the questionnaire's items.

With regard to eRisk data, recent approaches have been used to predict BDI-II responses using advanced computational methods. (Pérez et al., 2023) implemented a computational approach capable of rapidly screening depressive disorders on social networks, using a retrieval pipeline and a combination of semi- and supervised models, implementing item-based classifiers to predict responses to different items. Recently, (Ravenda et al., 2025) proposed a probabilistic neural architecture that uses a retrieval pipeline to select the most relevant posts for each user as input and predicts the final score using a Poisson distribution to better handle the ordinal nature of Likert scales. These latter two works differ from earlier approaches by incorporating a retrieval pipeline to filter the most relevant social media posts for each user and item before prediction, whereas previous methods fixed the number of posts per model, not considering their relevance. Additionally, while MBTI is a questionnaire with binary responses (binary classification), BDI-II uses a 4-point Likert scale, introducing an additional layer of complexity due to its ordinal nature.

Our work differs from previous approaches by focusing on a completely unsupervised scenario, leveraging LLMs in zero-shot contexts. The idea is to implement a retrieval pipeline to filter the most relevant posts and use LLMs to predict scores, linking the semantic content of posts to that of questionnaire items. To evaluate the effectiveness of LLMs in such tasks, we use two eRisk collections from the 2019 and 2020 editions (Losada et al., 2019, 2020), which contain the post-history of Reddit users alongside their completed BDI-II questionnaires. After demonstrating the effectiveness of this unsupervised approach, we extend it to differ-

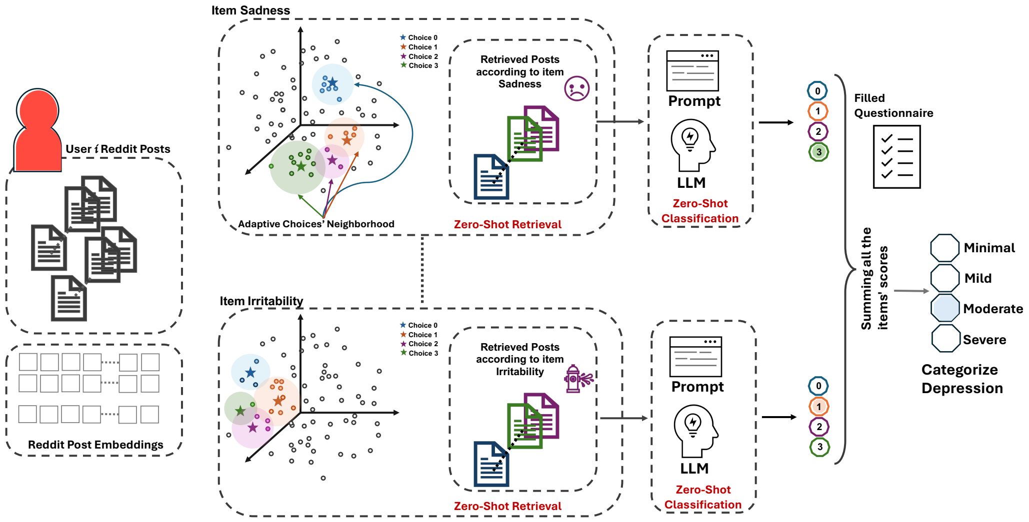

Figure 1: Pipeline of the main steps of our architecture. For each user, embeddings are created for each post and for each of the 4 choices of each item. The most relevant posts for each choice are retrieved and used as input for the LLM to generate the item score.

ent mental disorders, highlighting the benefits of constraining LLM predictions to individual psychological questionnaire responses.

3 Research Questions

In the following, we state the main **Research Questions** of our work:

(RQ1.) Is it possible to fill a psychological questionnaire based on a user's Reddit post history using a RAG-based approach in a completely unsupervised context? How does this approach compare in terms of performance to SOTA models for the considered datasets?

(RQ2.) How does the model effectiveness vary with changes in:

(RQ2a.) The LLM being used. To answer this question, we consider 4 LLMs: 2 open-source (`Phi-3-mini` and `Phi-3.5-mini`) and 2 closed-source (`Claude-3.5-Sonnet` and `gpt-4o-mini`).

(RQ2b.) The prompting strategy we employ. Specifically, we use both a direct prompting approach and Chain-of-Thought (CoT) (Wei et al., 2022) prompting.

(RQ2c.) The dense retrieval models. These are used to retrieve the most relevant posts for each user in response to each survey item.

(RQ3.) Can our approach—completing a standardized psychological questionnaire to obtain a "psychological explanation" of why a user is associated with a certain risk levels—be extended to other mental disorders where no ground truth is available?

(RQ4.) What are the benefits of our *psychological-guided* approach compared to an approach that relies exclusively on prompting?

Problem Definition. Computational approaches for predicting mental disorders from textual data, formulated as $f(\text{Text}) \rightarrow Y$ (Kim et al., 2020; Sekulić and Strube, 2019), where Text represents textual information like social media posts or interview transcriptions, face limitations in interpretability and generalizability across contexts (Paris et al., 2012; Friginal et al., 2017). These models, often neural, require large datasets to perform well. To address these limitations, we propose a method guided by psychological practice that leverages standardized questionnaires to assess the presence and/or intensity of mental disorders. The prediction is framed as a function correlating text and questionnaire items, $f(\text{Text}, \text{Item}_i) \rightarrow S_i$, where S_i represents the user's score for item i. The combination of these individual scores defines a final score used to diagnose symptoms as $\sum_i f(\text{Text}, \text{Item}_i) \rightarrow Y$. This approach not only provides an initial diagnosis but also enhances *outcome's interpretability* by linking the final score to specific questionnaire items, offering a clear reasoning for the prediction. The proposed method comprises two main steps, illustrated in Figure 1: **(1)** retrieving the most relevant documents for each item using an adaptive dense retrieval approach; **(2)** generating responses in a zero-shot setting with LLMs, using the retrieved documents as context.

Collection	Disorder	# of Users		# of Posts		Average # of Posts	
		Patient	Control	Patient	Control	Patient	Control
eRisk 2017	Depression	52	349	18’706	217’665	359.7	623.7
eRisk 2019	Self-Harm	41	299	6’927	163’506	169.0	546.8
eRisk 2020	Self-Harm	104	319	11’691	91’136	112.4	285.6

Table 1: Users statistics of “early detection” task for eRisk 2017 (depression) and 2019, 2020 (self-harm) editions.

Collection	*Minimal*	*Mild*	*Moderate*	*Severe*	# of Users	# of Posts
eRisk 2019	4	4	4	8	20	10’380
eRisk 2020	10	23	18	19	70	33’600

Table 2: Users summary statistics of “Measuring the severity of depression” task for eRisk 2019 and 2020 editions.

4 Methodology

4.1 Datasets

For the analyses in this study, we use datasets from the eRisk collections of 2017, 2019, and 2020 (Losada et al., 2017, 2019, 2020). The aim of this work is to evaluate the effectiveness of LLMs as assessors for mental disorders screening. For this purpose, we consider two different types of disorders: *depression* and *self-harm*, while focusing primarily on the former and using the latter as an additional proof-of-concept for our methodology. We use 5 datasets in total. For *depression*, we refer to task 1 of the 2017 edition and tasks 3 and 2 from the 2019 and 2020 editions. The first dataset includes social media posts and comments from users divided into two categories: those showing signs of depression and control users, belonging to the “*early detection*” task. The latter two datasets contain users post histories and responses to a standardized self-reported psychological questionnaire, the BDI-II, used for screening depressive disorders. These datasets belong to the “*Measuring the severity of depression*” task. Regarding *self-harm*, datasets contain social media posts and comments from users divided into two categories: users showing signs of self-harm and control users. Only posts preceding the user entry into the self-harm community are collected, in order to identify early signals and behavioral patterns that precede explicit help-seeking behavior, when intervention could be more effective and timely. The control group consists of both random users and users who actively participated in self-harm discussions (e.g., because they had a relative suffering from it). This dataset belongs to the “*early detection*” task. Table 1 shows some summary statistics for the datasets belonging to the early detection task, while Table 2 shows statistics for the Measuring the severity of depression task.

4.2 Adaptive Zero-Shot Retrieval Strategy

The number of Reddit posts per user can vary significantly, leading to two potential issues: exceeding the LLM token limit and reasoning degradation due to large input sizes (Fraga, 2024; Li et al., 2024). To manage this, we adopt a fully unsupervised posts retrieval approach using an embedding similarity approach, exploring 10 different dense retrieval models (see Table 3) and evaluating their effectiveness through LLM prediction accuracy.

To account for variability in relevant documents per user, we employ the ABIDE-ZS method (Ravenda et al., 2025). For each item, this approach identifies a neighborhood where the semantic meaning remains stable - i.e., the posts within this region share contextual relevance to the specific questionnaire’s item. Posts are retrieved based on semantic similarity, selecting the top k^* posts, where k^* is optimally determined by the ABIDE algorithm (Noia et al., 2024) for each item. This allows us to avoid the need to fix a priori the number of k documents to retrieve or specify a particular threshold (hence the name adaptive RAG for our approach).

In Figure 1, we illustrate the process of retrieving Reddit posts for a specific user according to each questionnaire item. For a given user i, embeddings are created for each post (scatters) and each choice (denoted by the $\star$ symbol) within an item. Let $\mathbf{P}_i = \{\mathbf{p}_{i1}, \ldots, \mathbf{p}_{im}\}$ represent the embeddings of the m posts for user i, and $\mathbf{iq}_j = \{\mathbf{iq}_{j1}, \mathbf{iq}_{j2}, \mathbf{iq}_{j3}, \mathbf{iq}_{j4}\}$ represent the embeddings of the 4 choices for item j. These item choices serve as queries (aka item-queries) to retrieve the most relevant posts from the user’s history. For each choice $\mathbf{iq}_{jl}$ (where $l = 1, 2, 3, 4$), the relevant Reddit posts $\mathbf{RRP}_{jl}$ (represented by colored dots in Figure 1) are retrieved by selecting the top k^* posts with the highest embedding similarity

Models	MiniLM-L6	MiniLM-L12	distilBERT-v4	T5	distilBERT-tas-b	all-mpnet	GIST	sf-e5	contriever	bge-large
Cosine	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗
Emb. Dim.	384	384	768	768	768	768	768	1024	768	1024
Avg. Docs retrieved	9	15	9	15	9	20	17	14	10	13

Table 3: Dense retrievers used are reported, along with how the similarity scores are calculated (*cosine similarity* - ✓ - or *dot product* - ✗), the dimension of the retriever embeddings, and the average number of documents retrieved per item with respect to the eRisk 2019 dataset.

scores: $\mathbf{RRP}_{ijl} = \{p_i\}_{p_i \in \aleph(iq_{jl})}$, where $\aleph(IC_{jl})$ represents the adaptive neighborhood of size k^* for item j and choice l. Specifically, for each item-query, we determine an adaptive neighborhood composed of k^* Reddit posts nearest neighbors, where k^* is not fixed but varies, from item to item, based on the local embeddings characteristics. Formally, we define two likelihood models: $M1$, which assumes different densities at the point q, representing a specific item-query, and its $(k+1)$-th neighbor, and $M2$, which assumes constant density. The log-likelihoods for these models are given by:
$L_{M1}(q,k) = klog((k^2)/(V_q, k * V_k + 1, k)) - 2k$
$L_{M2}(q,k) = 2klog((2k)/(V_q, k + V_k + 1, k)) - 2k$
where $V_{q,k}$ is the volume of the hypersphere centered at q with radius equal to the distance to its k-th neighbor. We then use a likelihood ratio test to determine the optimal $k^*(q)$: $k^*(q) = min(k : -2(L_{M2}(q,k) - L_{M1}(q,k)) \geq D_{thr})$ where D_{thr} is a threshold based on the desired confidence level. This adaptive approach allows us to identify a neighborhood size that is optimal for each query, adapting to local data densities and structures. In other words, the space identified by the adaptive neighborhood for each item-query can be viewed as a space where the semantics of the item-query remain constant, which is reflected in the region where the posts are semantically relevant to the item-query. Table 3 shows the statistics of the different dense retrieval models used, including the similarity measure adopted for each retriever (a more in depth discussion about retrievers used can be found in Appendix C). It can be observed that while the number of documents retrieved per item varies across models, no significant correlation exists between embedding dimension and the average number of retrieved documents (Pearson correlation, $p - value = 0.70$). The impact of k^* is discussed in Section D.1 of the Appendix.

4.3 Proposed Framework

The proposed workflow involves a two-step adaptive RAG pipeline. First, we retrieve the most relevant documents for each user with respect to each questionnaire item, and then we instruct the LLM to predict the corresponding score. In this work we explore various models for zero-shot retrieval and LLMs (both open- and closed-source) alongside different prompting techniques (Direct and CoT) to examine how the results vary across different combinations. All LLMs are used with temperature set to 0 to force deterministic outputs. For predictions, the closed-source models used include `gpt-4o-mini` and `Claude-3.5-Sonnet`, while the open-source models are `Phi-3-mini` and `Phi-3.5-mini`. For prompting, we use two techniques illustrated in Figure 2. The first technique generates the item scores directly based on relevant Reddit posts (Direct), while the second approach guides the LLM to reflect on intermediate steps (CoT), encouraging the LLM to go through reasoning steps before predicting the final score.

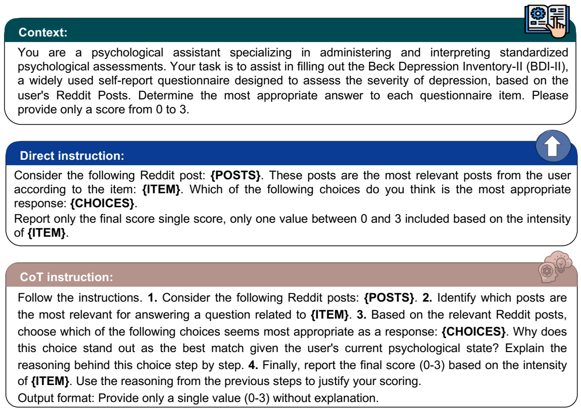

Figure 2: Prompt templates based on different prompting strategies: Direct and CoT.

5 Results

5.1 Predicting Psychological Questionnaire Scores

For evaluating the effectiveness of our approach in predicting responses to the BDI-II questionnaire, we use the official eRisk benchmark metrics (Losada et al., 2019) which assess performance at two distinct levels (for all the considered metrics, the higher the value, the better):

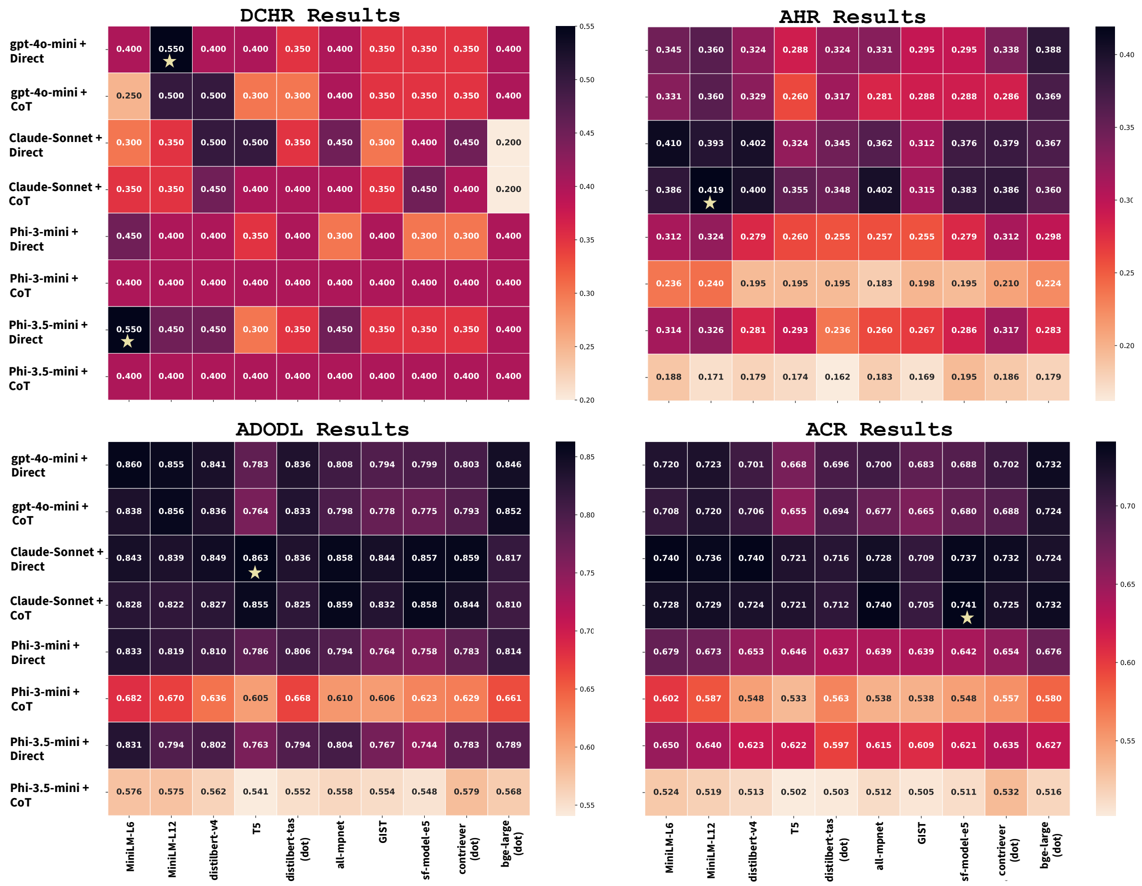

Figure 3: Results associated with the use of different combinations of LLMs, prompting strategies, and retrieval models across various metrics for the eRisk 2019 collection. The best combinations for each metric are highlighted with $\star$. The darker the color, the better the score.

At *questionnaire level*, we examine Depression Category Hit Rate (DCHR) which measures the accuracy in estimating depression severity levels (minimal: **0-9**, mild: **10-18**, moderate: **19-29**, and severe: **30-63**), and the Average Difference Between Overall Depression Levels (ADODL) which evaluates the overall BDI-II score estimations.

At *item level*, we employ Average Hit Rate (AHR) to evaluate prediction accuracy for individual symptoms, and Average Closeness Rate (ACR) to measure how close predictions are to actual values for each symptom.

We use datasets from the 2019 and 2020 eRisk editions, which contain users' post histories and their responses to the BDI-II questionnaire.

The idea is to explore different combinations of LLMs, dense retrievers, and prompting strategies for the eRisk 2019 dataset, and using the best combinations for the eRisk 2020 collection. To answer **(RQ.2)**, Figure 3 shows the results as heatmaps for each combination across the 4 metrics considered for this task for the eRisk 2019 dataset. On average, we observe that using a direct prompting technique leads to better results compared to a Chain of Thought (CoT) approach. Overall, we notice that closed-source LLMs often outperform open-source ones. Specifically, the worst results are obtained when using open-source models combined with CoT prompting techniques.

Table 4 shows the results of the best combinations for each metric for the corresponding eRisk edition, along with benchmark models. For comparison, we consider the best prior works for each metric for the eRisk 2019 and 2020 collections. We refer the reader to the corresponding overview for more in-depth details (Losada et al., 2019, 2020). In Section C of the Appendix we further discuss all the models used as benchmarks. After filtering the best performing combinations from the Figure 3, these are also used for the 2020 edition. We observe that, even among the benchmark models used for comparison, there is no single model that outperforms others across all metrics. Regarding the 2019 edition, the proposed approaches outper-

Collection	Model + Prompt Strategy	Retrieval	Questionnaire Metrics		Item Metrics	
			DCHR	ADODL	AHR	ACR
eRisk 2019	CAMH		45.00%	81.03%	23.81%	57.06%
	UNSLC (Burdisso et al., 2019)		40.00%	78.02%	41.43%	69.13%
	UNSLE (Burdisso et al., 2019)		35.00%	80.48%	40.71%	71.27%
	`gpt-4o-mini` + **Direct**	MiniLM-L12	**55.00%**	85.48%	35.95%	72.30%
	`Claude Sonnet` + **Direct**	T5	50.00%	**86.27%**	32.38%	72.14%
	`Claude Sonnet` + **CoT**	MiniLM-L12	35.00%	82.22%	**41.90%**	72.86%
	`Claude Sonnet` + **CoT**	sf-model-e5	45.00%	85.79%	38.33%	**74.12%**
eRisk 2020	BioInfo (Oliveira, 2020)		30.00%	76.01%	38.30%	69.21%
	ILab (Martínez-Castaño et al., 2020)		27.14%	81.70%	37.07%	69.41%
	Relai (Maupomé et al., 2020)		34.29%	83.15%	36.39%	68.32%
	Sense2vec (Pérez et al., 2022a)		37.14%	82.61%	38.97%	70.10%
	(Pérez et al., 2023) Recall		50.00%	85.24%	35.44%	67.23%
	(Pérez et al., 2023) Voting		47.14%	**85.33%**	35.24%	67.41%
	`gpt-4o-mini` + **Direct**	MiniLM-L12	41.43%	84.01%	36.60%	71.59%
	`Claude Sonnet` + **Direct**	T5	47.14%	83.92%	39.52%	73.31%
	`Claude Sonnet` + **CoT**	MiniLM-L12	32.86%	81.59%	**41.90%**	72.56%
	`Claude Sonnet` + **CoT**	sf-model-e5	42.86%	84.17%	41.77%	73.83 %
	`LLMs Ensemble`	-	**52.86%**	84.63%	39.52%	**74.10%**

Table 4: Model performance comparison on eRisk 2019 and 2020 collection w.r.t. questionnaire metrics (DCHR, ADODL) and item metrics (AHR, ACR). Bold values represent the best results for each collection. For all the considered metrics, the higher the score, the better.

Collection	Model + Prompt Strategy	Retrieval	DCHR BDI-II	DCHR BDI
eRisk 2019	`gpt-4o-mini` + **Direct**	MiniLM-L12	55.00%	55.00%
	`Claude Sonnet` + **Direct**	T5	45.00%	50.00%
	`Claude Sonnet` + **CoT**	MiniLM-L12	**50.00%**	35.00%
	`Claude Sonnet` + **CoT**	sf-model-e5	**55.00%**	45.00%
eRisk 2020	`gpt-4o-mini` + **Direct**	MiniLM-L12	**42.86%**	41.43%
	`Claude Sonnet` + **Direct**	T5	**48.57%**	47.14%
	`Claude Sonnet` + **CoT**	MiniLM-L12	**50.00%**	32.86%
	`Claude Sonnet` + **CoT**	sf-model-e5	**54.29%**	42.86%

Table 5: Performance comparison regarding the correct categorization of depressive state intensity considering the BDI-II true reparametrization.

form the benchmarks across all considered metrics, except for AHR, where only one combination manages to outperform that edition's best model. For the 2020 edition, the comparison is more challenging because all the considered comparisons use the 2019 edition data for training. In particular, the recent work by (Pérez et al., 2023), in a specific configuration, achieves SOTA results for questionnaire metrics, despite having low scores at the item level (AHR and ACR), where our combinations outperform the benchmarks. For depression category accuracy (DCHR), we obtain the best result with a voting-regressor ensemble of the top 3 models based on DCHR from 2019. This ensemble represents the rounded average of each prediction from the three models.

Interestingly, although the distribution of different categories changes between the two editions, the results, in terms of metrics, seem to show consistency between editions. No substantial performance drops are observed, except for DCHR. This negative result may be due to how overall scores are categorized into the 4 severity categories of the BDI-II questionnaire. Within the context of the challenge, the authors of the eRisk workshop use the BDI parameterization, the version preceding BDI-II. In BDI-II, new ranges are introduced that change from those of the previous test, especially regarding the minimum level of depression. Specifically, minimal or absent depression is identified as **0-13**, Mild as **14-19**, Moderate as **20-28**, and severe as **29-63**. Table 5 shows how DCHR changes when using the correct reparameterization, obtaining excellent and completely counterintuitive results, especially for the two models using Claude and the CoT strategy, compared to those obtained with the previous questionnaire parameterization in the 2020 dataset.

5.2 Identifying Signs of Self-Harm through questionnaire

We approach the identification of self-harm behaviors using the Self-Harm Inventory (SHI), a 22-item, yes/no self-report questionnaire, that screens for lifetime history of self-harm behaviors. A score of 5 or more "yes" responses on the SHI indicates potential mild forms of Deliberate Self-Harm (DSH) (Latimer et al., 2009). To answer **(RQ.3)**, the methodology follows the same approach used

Collection	Model	P	R	F1
eRisk 2019	UNSL	**0.71**	0.41	0.52
	iLab	0.68	**0.66**	**0.67**
	OURS	0.65	0.61	0.62
eRisk 2020	iLab	0.83	**0.69**	**0.75**
	OURS	**0.90**	0.52	0.67

Table 6: Performance of our approach compared to the best approach in each of the two eRisk editions for the early detection of self-harm task. Precision, Recall, and F1 metrics are reported.

Collection	Model	τ	P	R	F1
eRisk 2017	`gpt-4o-mini`	-	0.32	0.74	0.45
	PHQ-9	10	**0.69**	0.45	0.55
	DASS-42 (subscale)	20	0.36	0.64	0.46
	BDI-II	20	0.42	**0.76**	0.54
	Questionnaire Ensemble		0.53	0.69	**0.60**

Table 7: Performance comparison of our approach applied to different depression screening questionnaires (BDI-II, DASS-42, PHQ-9) and instructing `gpt-4o-mini` to only classify users into depressed/non-depressed categories.

for BDI-II, with the key distinction that SHI is structured as a binary questionnaire, unlike BDI-II Likert scale. The process involves retrieving the most relevant documents using SHI questions as queries, followed by LLM-generated responses. For this specific task, we chose the combination of gpt-4o-mini with Direct Prompt and MiniLM-L12-v3 as the retrieval model, as this configuration demonstrated the best trade-off between performance, computational costs, and processing speed in previous tests. The dataset provides each user complete post history, associated with a binary label indicating the presence or absence of self-harm behaviors. While the original task aims to identify self-harm cases as early as possible using the minimum number of posts, our approach considers the entire history to maximize prediction accuracy. For the 2019 edition, we compared our approach with two reference models: UNSL (Burdisso et al., 2019), which analyzes only a subset of posts, and iLab (Martínez-Castaño et al., 2020), which uses BERT fine-tuned in a fully supervised context. iLab approach was optimized to maximize F1-score and trained on a custom dataset built from self-harm subreddit posts. To ensure a fair comparison, we considered the version of iLab that has access to the complete post history for both the 2019 and 2020 editions. Table 6 shows that while the fine-tuned BERT-based model, benefiting from an extensive training corpus, generally outperforms our unsupervised approach, the performance gap is remarkably narrow, particularly for the 2019 edition. Notably, our model achieves superior precision in the 2020 edition. These results are particularly significant considering that our approach is completely unsupervised and requires no training data, yet achieves competitive performance compared to fully supervised models that leverage extensive domain-specific training. This suggests that our unsupervised approach based on adaptive RAG can offer a viable alternative in scenarios where labeled training data is scarce or unavailable.

5.3 The Importance of Questionnaire for Screening Procedure

For this task, we used data from the 2017 eRisk edition to address **(RQ4.)**. As in the previous subsection, we have access to users post histories along with corresponding labels indicating whether each user exhibited clinical signs of depression.

We tested the same adaptive RAG approach that proved effective for previous tasks, using the same combination of methods but guided by different types of standardized depression screening questionnaires. Specifically, we compared results obtained from the Beck Depression Inventory-II (BDI-II), Patient Health Questionnaire-9 (PHQ-9), Depression Anxiety Stress Scales-42 (DASS-42) (specifically focusing on its depression subscale), and simply instructing the LLM to determine if the patient was suffering from depression.

Each of these three questionnaires has established optimal cut-off scores for identifying clinically significant depression, shown as τ in Table 7, which also presents the results from the different approaches. The worst results were obtained through direct instruction to `gpt-4o-mini` and through the use of the DASS-42 depression subscale. The best results in terms of precision were achieved using PHQ-9 (0.69), while BDI-II showed the highest recall (0.76). The highest F1 score (0.60) was achieved using an ensemble classifier combining all three questionnaires. To answer **(RQ4.)**, our findings suggest that structured psychological assessment tools, can enhance the effectiveness of LLM-based mental health assessments compared to direct questioning approaches.

6 Conclusions

Our work demonstrates that LLMs, when guided by standardized psychological questionnaires through our novel adaptive RAG approach, can effectively support mental health screening tasks in a completely unsupervised setting. The results show the goodness of this approach in automatically completing questionnaires **(RQ1.)**, proving effective not only for depression screening but also extending successfully to other conditions like self-harm detection, and potentially other behavioral disorders **(RQ3.)**, across different combinations of LLMs, prompts, and retrievers **(RQ2.)**. We also show how our approach improves upon simpler methods for screening procedure that rely on direct prompting about the presence or absence of a depressive disorder **(RQ4.)**.

7 Limitations

The proposed methodology offers several advantages in terms of its implementation and performance. Despite these, it is important to address the limitations of this approach.

Although the results are particularly promising given the available data, a limitation of this work is the relatively small number of users. Furthermore, although the method can be easily extended to other types of questionnaires, there is no guarantee that similar results will be replicated across different questionnaires or various types of mental disorders.

Additionally, while the BDI-II is considered one of the most reliable tool for depression assessment, it has some limitations. As with all self-report measures, it can be influenced by the patient subjectivity and should not replace a comprehensive clinical diagnosis. Instead, it should be used as a screening tool in conjunction with other clinical evaluation methods for a complete and accurate diagnosis.

Furthermore, in this work we use cut-off scores to define different risk thresholds for specific disorders as reported in the original works based on psychometric criteria. However, these cut-off score guidelines are typically provided with the recommendation that thresholds should be adjusted according to sample characteristics and the intended purpose of the questionnaire.

8 Ethical Considerations

The proposed methodology for mental health support and assessment, while novel, raises several ethical considerations that must be addressed to ensure responsible deployment.

There is potential for AI to be misused as a clinical tool. Without proper safeguards, these models could exhibit harmful or biased behaviors. It is crucial to emphasize that this approach should not be viewed as a substitute for specialized medical professionals, but rather as a method to screen for potential subjects at risk of depression.

Ethical considerations extend to privacy and data protection, ensuring the confidentiality and security of users social media data. In the case of this study, informed consent has been given from users for research practice, with users fully aware of how their data is being used and for what purpose.

Acknowledgements. This research has partly been founded by the SNSF (Swiss National Science Foundation) grant 200557.

References

Mostafa M Amin, Erik Cambria, and Björn W Schuller. 2023. Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt. *IEEE Intelligent Systems*, 38(2):15–23.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Ramiro N Barros, Kristen K Arguello, and Jônatas Wehrmann. 2024. Anchor your embeddings through the storm: Mitigating instance-to-document semantic gap. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Aaron T Beck, Robert A Steer, Gregory K Brown, et al. 1996. Beck depression inventory.

Sandra Bucci, Matthias Schwannauer, and Natalie Berry. 2019. The digital revolution and its impact on mental health care. *Psychology and Psychotherapy: Theory, Research and Practice*, 92(2):277–297.

Sergio Gastón Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. 2019. Unsl at erisk 2019: a unified approach for anorexia, self-harm and depression detection in social media. In *CLEF (Working Notes)*.

Fahed Elourajini and Esma Aïmeur. 2022. Aws-ep: A multi-task prediction approach for mbti/big5 personality tests. In *2022 IEEE International Conference* *on Data Mining Workshops (ICDMW)*, pages 1–8. IEEE.

Natanael Fraga. 2024. Challenging llms beyond information retrieval: Reasoning degradation with long context windows.

Eric Friginal, Oksana Waugh, and Ashley Titak. 2017. Linguistic variation in facebook and twitter posts. In *Studies in corpus-based sociolinguistics*, pages 342–362. Routledge.

Lewis R Goldberg. 2013. An alternative "description of personality": The big-five factor structure. In *Personality and Personality Disorders*, pages 34–47. Routledge.

Christopher J Hopwood, Katherine M Thomas, Kristian E Markon, Aidan GC Wright, and Robert F Krueger. 2012. Dsm-5 personality traits and dsm–iv personality disorders. *Journal of abnormal psychology*, 121(2):424.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190.

Ekaterina Khramtsova, Shengyao Zhuang, Mahsa Baktashmotlagh, Xi Wang, and Guido Zuccon. 2023. Selecting which dense retriever to use for zero-shot search. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 223–233.

Christian Kieling, Helen Baker-Henningham, Myron Belfer, Gabriella Conti, Ilgi Ertem, Olayinka Omigbodun, Luis Augusto Rohde, Shoba Srinath, Nurper Ulkuer, and Atif Rahman. 2011. Child and adolescent mental health worldwide: evidence for action. *The lancet*, 378(9801):1515–1525.

Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific reports*, 10(1):11846.

Shane Latimer, Tanya Covic, Steven R Cumming, and Alan Tennant. 2009. Psychometric analysis of the self-harm inventory using rasch modelling. *BMC psychiatry*, 9:1–9.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*.

David E Losada, Fabio Crestani, and Javier Parapar. 2017. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, pages 346–360. Springer.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pages 340–357. Springer.

David E Losada, Fabio Crestani, and Javier Parapar. 2020. erisk 2020: Self-harm and depression challenges. In *European conference on information retrieval*, pages 557–563. Springer.

Rodrigo Martínez-Castaño, Amal Htait, Leif Azzopardi, and Yashar Moshfeghi. 2020. Early risk detection of self-harm and depression severity using bert-based transformers: ilab at clef erisk 2020.

Diego Maupomé, Maxime D Armstrong, Raouf Moncef Belbahar, Josselin Alezot, Rhon Balassiano, Marc Queudot, Sébastien Mosser, and Marie-Jean Meurs. 2020. Early mental health risk assessment through writing styles, topics and neural models. In *CLEF (Working Notes)*.

John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences*, 25(2):113–122.

Antonio Di Noia, Iuri Macocco, Aldo Glielmo, Alessandro Laio, and Antonietta Mira. 2024. Beyond the noise: intrinsic dimension estimation with optimal neighbourhood identification. *Preprint*, arXiv:2405.15132.

Luıs Oliveira. 2020. Bioinfo@ uavr at erisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases. In *Proceedings of the CEUR Workshop Proceedings, Thessaloniki, Greece*, pages 22–25.

Cécile Paris, Paul Thomas, and Stephen Wan. 2012. Differences in language and style between two social media communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 539–542.

Anxo Pérez, Javier Parapar, and Álvaro Barreiro. 2022a. Automatic depression score estimation with word embedding models. *Artificial Intelligence in Medicine*, 132:102380.

Anxo Pérez, Javier Parapar, Álvaro Barreiro, and Silvia Lopez-Larrosa. 2023. Bdi-sen: A sentence dataset for clinical symptoms of depression. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2996–3006.

Anxo Pérez, Neha Warikoo, Kexin Wang, Javier Parapar, and Iryna Gurevych. 2022b. Semantic similarity models for depression severity estimation. *arXiv preprint arXiv:2211.07624*.

Alireza Pourkeyvan, Ramin Safa, and Ali Sorourkhah. 2024. Harnessing the power of hugging face transformers for predicting mental health disorders in social networks. *IEEE Access*, 12:28025–28035.

Anuraag Raj, Zain Ali, Shonal Chaudhary, Kavitesh Kumar Bali, and Anuraganand Sharma. 2024. Depression detection using bert on social media platforms. In *2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pages 228–233. IEEE.

Federico Ravenda, Seyed Ali Bahrainian, Noriko Kando, Antonietta Mira, Andrea Raballo, and Fabio Crestani. 2025. Tailoring adaptive-zero-shot retrieval and probabilistic modelling for psychometric data. In *The 40th ACM/SIGAPP Symposium on Applied Computing (SAC '25)*, SAC '25. Association for Computing Machinery.

Gony Rosenman, Lior Wolf, and Talma Hendler. 2024. Llm questionnaire completion for automatic psychiatric assessment. *arXiv preprint arXiv:2406.06636*.

Randy A Sansone and Lori A Sansone. 2010. Measuring self-harm behavior with the self-harm inventory. *Psychiatry (Edgmont)*, 7(4):16.

David M Schweiger. 1985. Measuring managerial cognitive styles: On the logical validity of the myers-briggs type indicator. *Journal of Business Research*, 13(4):315–328.

Ivan Sekulić and Michael Strube. 2019. Adapting deep learning methods for mental health prediction on social media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 322–327.

Aivin V Solatorio. 2024. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *Preprint*, arXiv:2104.08663.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, et al. 2022. Overview of the clpsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Huy Vu, Suhaib Abdurahman, Sudeep Bhatia, and Lyle Ungar. 2020. Predicting responses to psychological questionnaires from participants' social media posts and question text embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1512–1524.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Brenda K Wiederhold. 2022. The escalating crisis in adolescent mental health.

Petr Winkler, T Formanek, K Mlada, A Kagstrom, Z Mohrova, P Mohr, and L Csemy. 2020. Increase in prevalence of current mental disorders in the context of covid-19: analysis of repeated nationwide cross-sectional surveys. *Epidemiology and psychiatric sciences*, 29:e173.

XUHAI XU, BINGSHENG YAO, YUANZHE DONG, HONG YU, JAMES HENDLER, ANIND K DEY, and DAKUO WANG. 2023. Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol*, 1(1).

Feifan Yang, Tao Yang, Xiaojun Quan, and Qinliang Su. 2021. Learning to answer psychological questionnaire for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1131–1142.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024a. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.

Minqiang Yang, Yongfeng Tao, Hanshu Cai, and Bin Hu. 2024b. Behavioral information feedback with large language models for mental disorders: Perspectives and insights. *IEEE Transactions on Computational Social Systems*, 11(3):3026–3044.

A Data Availability

The datasets supporting the conclusions of this article from eRisk collections are available for research purposes under signing user agreements.

B Dense Retrieval Models

We use a pool of different dense retrieval models: *msmarco-MiniLM-L-6-v3*[3], *msmarco-MiniLM-L-12-v3*[4], *msmarco-distilbert-base-v4*[5], *sentence-T5-base*[6], *msmarco-distilbert-base-tas-b*[7], *all-mpnet-base-v2*[8], *GIST*[9], *sf-model-e5*[10], *contriever-msmarco*[11], *bge-large*[12]. The selected models are a diverse and representative sample of the SOTA for dense retrieval and have already been tested in the literature (Khramtsova et al., 2023; Barros et al., 2024; Solatorio, 2024).

The chosen models, mostly pre-trained on MS MARCO (Bajaj et al., 2018), allow for an in-depth analysis of generalization capabilities in zero-shot scenarios. From an empirical validation perspective, many of the selected models are present in the BEIR leaderboard (Thakur et al., 2021), thus providing established benchmarks for performance evaluation. The standardization of implementation through the Hugging Face platform ensures uniformity in evaluations and facilitates the reproducibility of results.

C Benchmark Models

As benchmark models for "*Measuring the Severity of depression*" task, we use a combination of top-performing models from the eRisk competition and SOTA models from the literature on the considered datasets, specifically regarding the task of measuring depression severity.

For eRisk 2019, two notable approaches were developed. The first is CAMH (Losada et al., 2019), which represented users through LIWC features. Then, for each BDI questionnaire item, it matches a vectorial representation of the user against vectorial representations of possible responses. The second approach, UNSL (Burdisso et al., 2019), converts textual indicators from user posts into a standardized clinical depression score (0-63). It maps linguistic analysis into 4 clinical severity categories using various statistical and text processing techniques to complete the 21-question diagnostic questionnaire.

For eRisk 2020, several methods emerges. BioInfo (Oliveira, 2020) and Relai (Maupomé et al., 2020) methods obtained their own datasets to perform standard ML classifiers using engineered features as linguistic markers.

We also refer to recent works (Pérez et al., 2022b, 2023). The two approaches aim to estimate depression severity from Reddit posts using BDI-II symptom-based classifiers. While the first approach (Pérez et al., 2022b) uses word embeddings to compare BDI-II options and user texts, (Pérez et al., 2023) leverages expert-annotated "golden" sentences (738 in total) as queries to identify semantically similar "silver" sentences through RoBERTa embeddings, achieving better performance through Accumulative Voting and Recall aggregation methods.

D Ablation Study

D.1 The Impact of the Adaptive k^* Dense Retrieval Approach

Figure 4 shows how performance metrics change as we vary the number of retrieved documents k in the eRisk 2019 dataset. We analyze the performance of different metrics using `gpt-4o-mini` as the LLM and `MiniLM-L12-v3` as the retrieval model, while varying the parameter k across {5, 10, 15, 20, 30, 40, 50}, where $k = 15$ represents the mean value of k^*. We test this scenario on both the 2019 and 2020 eRisk editions. We observe that using k^* (horizontal dashed lines), which corresponds to adaptive RAG, often yields the best results.

For the 2019 edition, the best values are achieved with k^* for the questionnaire metrics (DCHR, ADODL), while $k = 10$ produces the best metric in terms of AHR, and $k = 15$ performs best for ACR (item metrics) - both values being close to the mean k^* value.

The 2020 edition shows slightly different results: for DCHR, the best value is obtained with the mean k^*, while the best ADODL and AHR corresponds

[3] https://huggingface.co/sentence-transformers/msmarco-MiniLM-L-6-v3

[4] https://huggingface.co/sentence-transformers/msmarco-MiniLM-L-12-v3

[5] https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v4

[6] https://huggingface.co/sentence-transformers/sentence-t5-base

[7] https://huggingface.co/sentence-transformers/msmarco-distilbert-base-tas-b

[8] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[9] https://huggingface.co/avsolatorio/GIST-Embedding-v0

[10] https://huggingface.co/jamesgpt1/sf_model_e5

[11] https://huggingface.co/facebook/contriever-msmarco

[12] https://huggingface.co/BAAI/bge-large-en

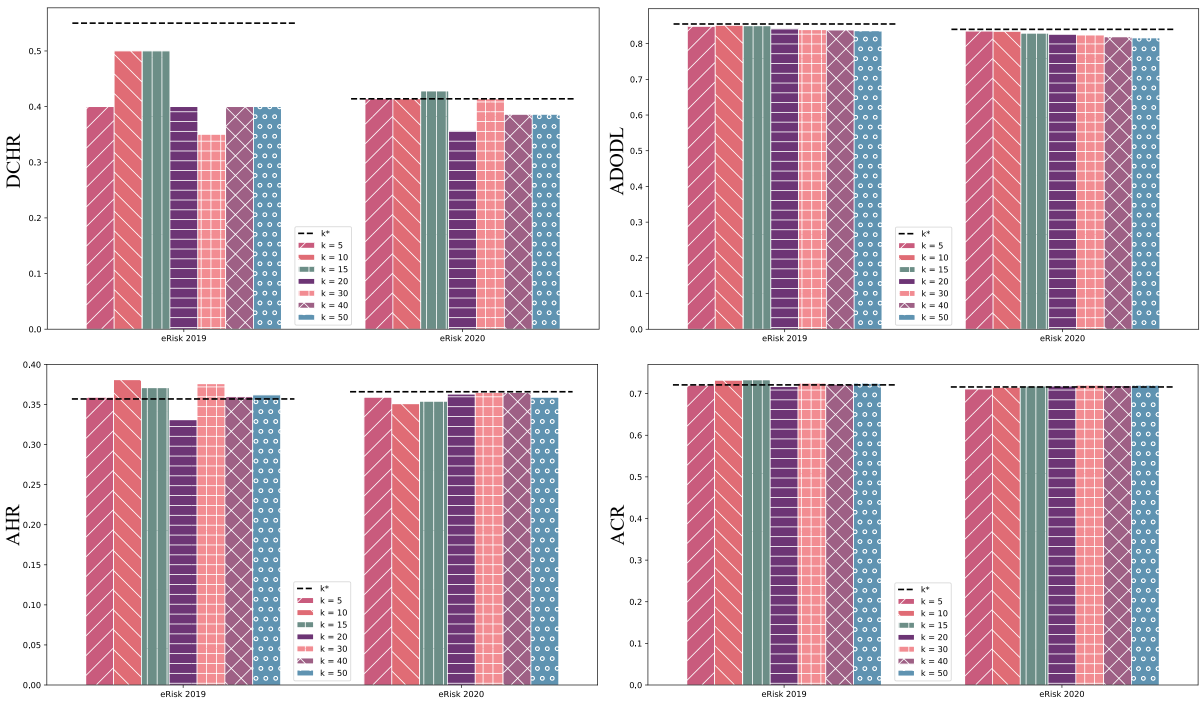

Figure 4: Comparison of the performance of the gpt-4o-mini + Direct Prompt combination with MiniLM-L12-v3 retrieval model on the eRisk 2019 and 2020 datasets, varying by k value used to select the number of documents to retrieve. k^* represent our adaptive method, and $k = 15$ the average number of k^*.

to k^*. The best ACR results are obtained with $k = 30, 50$.

Overall, while the adaptive RAG strategy may not always lead to optimal results across all metrics, it allows us to achieve consistently strong performance without the need to set any parameters a priori or explore different choices of k. This automated approach to determining k offers a robust and efficient solution that removes the need for manual parameter tuning while maintaining competitive performance levels across different evaluation scenarios and metrics.

LLMs	**DCHR**	**ADODL**	**AHR**	**ACR**
$\mu_{direct} > \mu_{CoT}$				
`gpt-4o-mini`		†	†	†
`Claude Sonnet`		†		
`Phi-3-mini`		‡	‡	‡
`Phi-3.5-mini`		‡	‡	‡

Table 8: The significance of the LLMs prediction goodness w.r.t. the two prompting techniques used, Direct and CoT, is shown (according to the eRisk 2019 collection). Metrics with no significant difference are marked in red, while † denotes a significant difference according to the t-test, and ‡ denotes significance according to the Mann-Whitney U test as well.

D.2 The Impact of Different Prompting Strategies

In Table 8, we evaluate whether the use of direct prompting is statistically better than CoT across different metrics w.r.t. eRisk 2019 dataset collection. We perform both parametric, t-test, and non-parametric, Mann-Whitney U test (in both cases, we test whether one population mean is statistically greater than the other, $\alpha = 0.05$). Open-source models perform well only in direct prompt contexts, while they perform poorly when the prompting technique is CoT, tending to overestimate questionnaire scores. Regarding closed-source LLMs, we observe that there isn't always a significant difference between the two prompting approaches. For most metrics, with respect to gpt-4o-mini, the difference is significant only when using t-test.

D.3 The Impact of Different Retrieval Approaches

We also observe that some dense retrieval models perform better globally compared to others (see Figure 3), while some perform better only with respect to a subset of LLMs. In Figure 5, we examine the distribution of LLMs scores across different retrieval approaches and their rankings for each metric w.r.t. eRisk 2019 dataset collection. While no single approach consistently outperforms the others across all metrics, we can observe that some retrievers (*msmarco-MiniLM-L-6-v3*, *msmarco-MiniLM-L-12-v3*, *distillbert-v4*, and *bge-large*) generally achieve better rankings based on median values.

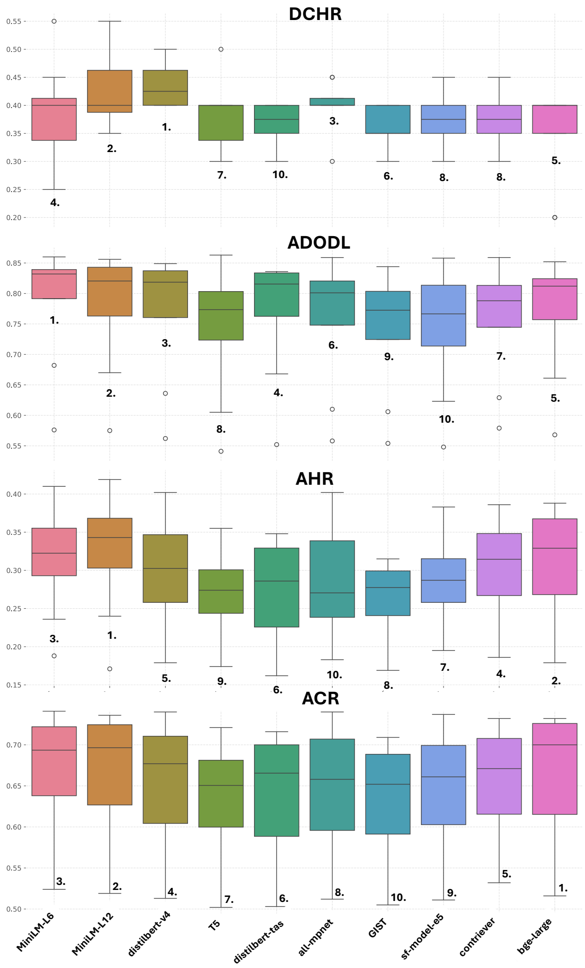

Figure 5: Distribution of adaptive RAG scores conditioned on different retrieval approaches and their rankings for each metric on the eRisk 2019 dataset. Rankings are based on median scores, with mean scores used as a tiebreaker in case of equal medians.