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]Euler Institute, Università della Svizzera italiana, [2]National Institute of Informatics
[3]Brown University, [4]University of Tübingen, [5]Insubria University

Abstract

In psychological practices, standardized questionnaires serve as essential tools for assessing mental health through structured, clinically-validated questions (i.e., items). While social media platforms offer rich data for mental health screening, computational approaches often bypass these established clinical assessment tools in favor of black-box classification. We propose a novel questionnaire-guided screening framework that bridges psychological practice and computational methods through adaptive Retrieval-Augmented Generation (*aRAG*). Our approach links unstructured social media content and standardized clinical assessments by retrieving relevant posts for each questionnaire item and using Large Language Models (LLMs) to complete validated psychological instruments. Our findings demonstrate two key advantages of questionnaire-guided screening: First, when completing the Beck Depression Inventory-II (BDI-II), our approach matches or outperforms state-of-the-art performance on Reddit-based benchmarks without requiring training data. Second, we show that guiding LLMs through standardized questionnaires can yield superior results compared to directly prompting them for depression screening, while also providing a more interpretable assessment by linking model outputs to clinically validated diagnostic criteria. Additionally, we show, as a proof-of-concept, how our questionnaire-based methodology can be extended to other mental conditions' screening, highlighting the promising role of LLMs as psychological assessors.[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 (MHCs) 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 (WHO, 2022).

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 different widely used standardized psychological questionnaires, specifically the Beck Depression Inventory-II (BDI-II) (Beck, 1996) for depression screening, the Self-Harm Inventory (SHI) (Sansone and Sansone, 2010) for self-harm behaviour detection, the SCOFF questionnaire (Morgan et al., 1999) for eating disorders screening, and the Pathological Gambling Diagnostic Form from DSM-V (Hopwood et al., 2012). All of them are self-reported surveys where overall scores correspond to specific severity levels of the respective conditions.

Despite their clinical validity, traditional psychological questionnaires present several practical limitations. They typically require in-person administration by trained professionals, making large-scale screening logistically challenging and resource-intensive. Access barriers disproportionately affect underserved populations, creating equity concerns in mental health assessment.

Increasingly, people are turning to social net-

[1] Code available: https://github.com/Fede-stack/Adaptive-RAG-for-Psychological-Assessment

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8975–8991
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

works as a space to express their feelings and experiences and 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.

Recent research has revealed significant limitations in using closed-source LLMs for mental disorder classification. Studies by (Amin et al., 2023) and (XU et al., 2023) demonstrate that both zero-shot and few-shot approaches struggle to match state-of-the-art (SOTA) supervised methods. These limitations stem from several challenges: (1) LLMs' difficulty in directly mapping unstructured text to diagnostic categories, (2) the complex, multi-dimensional nature of mental health assessment requiring domain expertise, and (3) the semantic gap between social media language and clinical criteria. We propose mitigating these challenges through an intermediate step: rather than attempting direct diagnosis, we instruct LLMs to complete standardized psychological questionnaires, effectively decomposing the complex diagnostic task into structured, clinically-validated assessment items.

To achieve that, we propose an adaptive RAG approach (aRAG), combining retrieval and classification, to accurately predict users' responses to psychological questionnaire items by analyzing their Reddit post history. Unlike existing methods that use fixed retrieval parameters or direct classification, our approach automatically determines the optimal number of relevant posts needed for each questionnaire item, adapting to the semantic density and relevance of available content. To the best of our knowledge, we demonstrate, for the first time, how LLMs (both open- and closed-source) can serve as effective annotators of standardized psychological questionnaires by analyzing social media posts through aRAG.

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 the results of our approach 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 MHCs.

These results confirm that LLMs can serve as effective and promising psychological assessors when their predictions are guided by standardized clinical instruments, bridging the gap between language models and psychometric practice.

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 (Ríssola et al., 2019). Transformer-based models (Vaswani et al., 2017) have significantly advanced mental health analysis on social media platforms. While initial work used BERT variants for depression detection (Raj et al., 2024; Ríssola et al., 2020), specialized models like MentalBERT (Ji et al., 2022) emerged, pre-trained specifically on mental health-related content. A novel direction explored how emotion manifests in depressed individuals' (Bucur et al., 2022) and a broader spectrum of mental disorders (De Grandi et al., 2024) from social media posts, using emotional and psychological markers to provide interpretable assessment.

Recently, Large Language Models have shown increasing promise, with approaches like MentaLLaMA (Yang et al., 2024a) offering interpretable analysis, and (Varadarajan et al., 2024) combining theoretical frameworks with computational techniques for suicide risk assessment. These advances suggest potential for real-time intervention and support (Yang et al., 2024b).

Concerning the use of NLP methods to predict psychological questionnaires' responses, early approaches used neural models to predict personality traits (Elourajini and Aïmeur, 2022) and Myers-Briggs indicators (Yang et al., 2021) from user-generated social media content. BERT embeddings have been leveraged to link social media expressions with psychological assessments (Vu et al., 2020; Atari et al., 2023). More recently, (Rosenman et al., 2024) used an LLM to impersonate interviewees and complete questionnaires, using these responses as features in a Random Forest to predict new questionnaire scores.

With regard to eRisk data, recent approaches were used to predict BDI-II responses using advanced computational methods. (Pérez et al., 2023) introduced a retrieval-based framework with item-

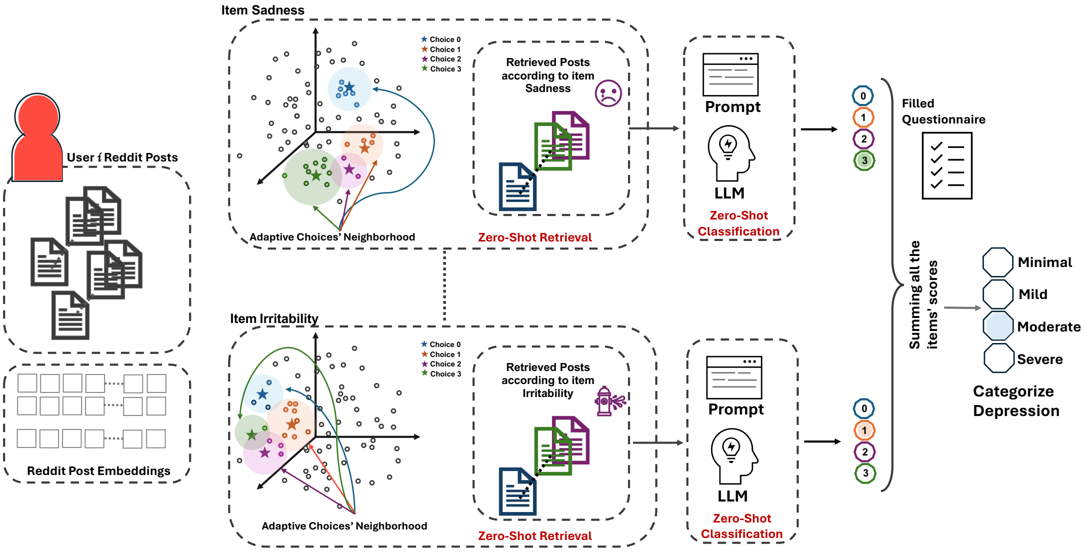


Figure 1: Pipeline of the main steps of our architecture. For each user, embeddings are created for each post and for each of the 4 choices of each item. The most relevant posts for each choice are retrieved and used as input for the LLM to generate the item score.

based classifiers for depression screening, while (Ravenda et al., 2025) proposed a probabilistic approach combining retrieval mechanisms to handle ordinal Likert scales. Both approaches innovate via retrieval-based selection of relevant social media content, departing from previous methods that used fixed post selection.

Our work differs from previous approaches by focusing on a completely unsupervised scenario, leveraging LLMs through an aRAG approach, filtering the most relevant posts and use LLMs to predict scores, establishing a semantic mappings between social media content and standardized questionnaire items. To evaluate the effectiveness of our approach 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 different MHCs, highlighting the benefits of constraining LLM predictions to individual psychological questionnaire responses.

3 Research Questions

The main **Research Questions** we address are:
(RQ1.) Is it possible to fill out a psychological questionnaire automatically based on a user's Reddit post history in an unsupervised context? How does this compare in terms of performance to SOTA models for the considered datasets?
(RQ2.) How does the model's effectiveness vary with changes in:
(RQ2a.) The LLM being used. To answer this question, we consider six LLMs: two large-scale open-source (`Qwen 2.5 70B`, `DeepSeek V3`), two lightweight open-source (`Phi-3-mini`, `Phi-3.5-mini`), and 2 closed-source (`Claude-3.5-Sonnet`, `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 certain risk levels - be extended to other MHCs without training data?
(RQ4.) What are the benefits of our *psychological-guided* approach compared to an approach that relies exclusively on prompting?

Problem Definition. While existing computational approaches typically frame mental health prediction as a direct mapping from text to diagnosis $f(\text{Text}) \rightarrow Y$ (Kim et al., 2020; Sekulić and Strube, 2019), where Text represents textual information like social media posts or interview transcriptions, this black-box formulation faces significant challenges in clinical applicability, interpretability, and generalizability across contexts (Paris et al., 2012; Friginal et al., 2017). These models, often neural, require large datasets to perform well. Our work introduces a novel paradigm that bridges computational and clinical practice by leveraging standardized psychological questionnaires as an intermediate structured representation. By reformulating the prediction task through questionnaire items, our method offers several key advantages: (1) clinical interpretability through standardized assessment criteria, (2) transparent reasoning through item-level predictions, and (3) alignment with established psychological practice. The prediction is framed as a function linking 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 $\sum_i f(\text{Text}, \text{Item}_i) \rightarrow Y$. The proposed method comprises two main steps, illustrated in Figure 1: **(1)** retrieving the most relevant posts 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.

4 Methodology

4.1 Datasets

Our analyses uses seven datasets from the 2017, 2019, 2020, and 2022 eRisk collections (Losada et al., 2017, 2019, 2020; Parapar et al., 2022), encompassing different MHCs. The depression datasets consist of two distinct tasks. The "*early detection task dataset*" from eRisk 2017 (used in Section 5.3) provides binary classification data (depression vs. control users) through user posts and comments, with users labeled based on signs of depression. The "*severity assessment task datasets*" from eRisk 2019 and 2020 (used in Section 5.1) contain user post histories along with their responses to the BDI-II questionnaire, enabling detailed depression severity measurement. The self-harm, anorexia, and pathological gambling datasets belong to the *"early detection tasks datasets"* from eRisk 2019, 2020, and 2022 (used in Section 5.2). These datasets contain posts preceding users' entry into self-harm, anorexia, and gambling communities, aiming to identify early warning signals before explicit help-seeking behavior. Table 1 summarizes the user and post distributions across conditions for the *early detection task datasets*, while Table 2 presents detailed statistics for the *severity assessment task datasets*, including the distribution of depression severity levels.

Collection	# of Users		# of Posts	
	Patient	Control	Patient	Control
eRisk 2017 - Depression	52	349	359.7	623.7
eRisk 2019 - Self-Harm	41	299	168.9	212.4
eRisk 2020 - Self-Harm	104	319	112.4	285.7
eRisk 2019 - Anorexia	73	742	241.4	745.1
eRisk 2022 - Gambling	81	1998	180.6	507.6

Table 1: Users statistics of "early detection task datasets" for eRisk 2017, 2019, 2020, and 2022 for *depression*, *self-harm*, *anorexia*, and pathological gambling.

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

Table 2: Users summary statistics of "severity assessment task datasets" for eRisk 2019 and 2020 editions.

4.2 Adaptive Zero-Shot Retrieval Strategy

In this subsection, we address the challenge of retrieving relevant social media content for psychological assessment. The number of Reddit posts per user varies widely, leading to potential issues such as exceeding the LLM token limit and reasoning degradation due to large inputs (Fraga, 2024; Li et al., 2024, 2025). To mitigate these issues, we adopt a fully unsupervised retrieval strategy based on embedding similarity, exploring 10 different dense retrieval models (see Table 3) and evaluating their effectiveness through LLM prediction accuracy. To account for variability in relevant posts 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 (see Section B). This eliminates the need to fix a priori the number of k posts to retrieve or set a threshold, making our approach *adaptive*.

Figure 1 illustrates the retrieval process. For a

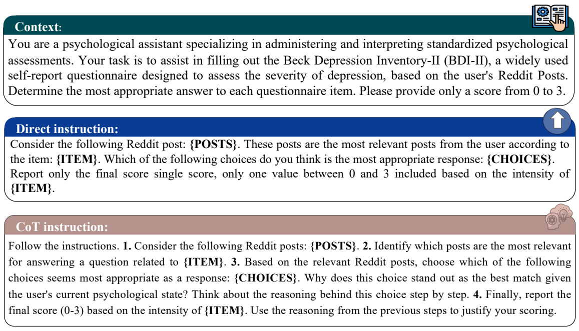
Models	MiniLM-L6	MiniLM-L12	distilBERT-v4	T5	distilBERT-tas-b	all-mpnet	GIST	sf-e5	contriever	bge-large
Cosine	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗
Emb. Dim.	384	384	768	768	768	768	768	1024	768	1024
Avg. Docs retrieved	9	15	9	15	9	20	17	14	10	13

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.

each user i, embeddings are generated for their posts (scatters) and for the four response choices (denoted by the $\star$ symbol) within an item. Let $\mathbf{P}_i = \{\mathbf{p}_{i1}, \dots, \mathbf{p}_{im}\}$ 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 four 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 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. Consider user posts and questionnaire item choices embedded in R^D, where D is the dimension of the embeddings. The optimal k^* is determined by the local density of posts around each query point, ensuring we retrieve exactly the number of posts needed to maintain semantic coherence. In other words, the space identified by the adaptive neighborhood for each item-query can be seen 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 D). We observe 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 E.2 of the Appendix.

4.3 Proposed Framework

The proposed workflow - aRAG - involves a two-step pipeline. First, we retrieve the most relevant posts for each user with respect to each questionnaire item, and then we instruct the LLM to predict the corresponding score. Our investigation combines SOTA retrieval techniques with both open and closed-source LLMs, examining how different prompting strategies affect psychological assessment accuracy. All LLMs are used with temperature set to 0 to force deterministic outputs. For predictions, we use the closed-source models `gpt-4o-mini` and `Claude-3.5-Sonnet`, alongside large-scale open-source models `Qwen 2.5 70B` and `DeepSeek V3`, as well as lightweight open-source models `Phi-3-mini` and `Phi-3.5-mini`. For prompting, we use two techniques illustrated in Figure 2. The first technique predict 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.



Context:

You are a psychological assistant specializing in administering and interpreting standardized psychological assessments. Your task is to assist in filling out the Beck Depression Inventory-II (BDI-II), a widely used self-report questionnaire designed to assess the severity of depression, based on the user's Reddit Posts. Determine the most appropriate answer to each questionnaire item. Please provide only a score from 0 to 3.

Direct instruction:

Consider the following Reddit post: **{POSTS}**. These posts are the most relevant posts from the user according to the item: **{ITEM}**. Which of the following choices do you think is the most appropriate response: **{CHOICES}**. Report only the final score single score, only one value between 0 and 3 included based on the intensity of **{ITEM}**.

CoT instruction:

Follow the instructions. **1.** Consider the following Reddit posts: **{POSTS}**. **2.** Identify which posts are the most relevant for answering a question related to **{ITEM}**. **3.** Based on the relevant Reddit posts, choose which of the following choices seems most appropriate as a response: **{CHOICES}**. Why does this choice stand out as the best match given the user's current psychological state? Think about the reasoning behind this choice step by step. **4.** Finally, report the final score (0-3) based on the intensity of **{ITEM}**. Use the reasoning from the previous steps to justify your scoring.

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

5 Results

5.1 Predicting Psychological Questionnaire Scores

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

At *the level of the questionnaire*, we examine the Hit Rate of the Depression Category (DCHR) which measures the accuracy in estimating depres-

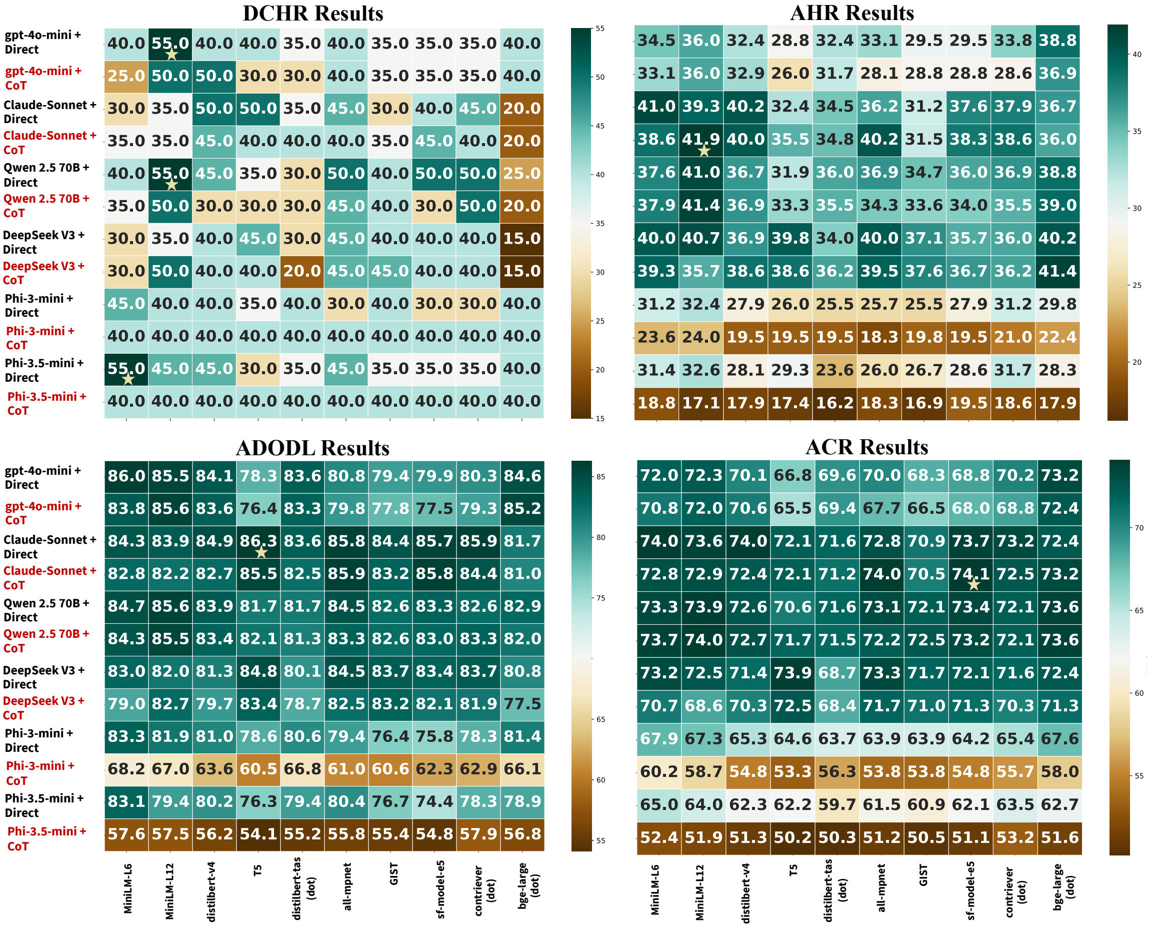


Figure 3: Results associated with the use of different combinations of LLMs, prompting strategies, and retrieval models across various metrics for the eRisk 2019 collection. The best combinations for each metric are highlighted with ⋆. The greener the color, the better the score. For readability we report only the first decimal.

sion 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 general 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.

We systematically evaluate different combinations of LLMs, dense retrievers, and prompting strategies for the eRisk 2019 dataset, using the best combinations for the eRisk 2020 collection. To address **(RQ.2)**, Figure 3 presents heatmap visualizations comparing performance across all combinations for the four evaluation metrics on the eRisk 2019 dataset.

On average, we observe that lightweight models perform significantly better with direct prompting compared to CoT approach, while this difference is less pronounced in larger architectures. Overall, we notice that closed-source LLMs and large-scale open-source models often outperform lightweight ones. This pattern can be attributed to the limited model capacity of lightweight architectures, their reduced ability to effectively manage multi-step reasoning chains during CoT prompting, and their lesser capability to capture subtle linguistic nuances crucial for mental health assessment. Specifically, the worst results are obtained when using lightweight open-source models combined with CoT prompting techniques.

After identifying the best performing combinations on the eRisk 2019 dataset from Figure 3, we use these to the 2020 collection. Table 4 shows these results alongside baseline benchmarks. For comparison, we consider the best performing models from previous work for each metric for the

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%
	`Qwen 2.5 70B` + **Direct**	MiniLM-L12	**55.00**%	85.56%	40.95%	73.81%
	`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	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%
	`Qwen 2.5 70B` + **Direct**	MiniLM-L12	41.43%	83.49%	38.78%	72.74%
	`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.

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 D of the Appendix we further discuss all the models used as benchmarks. Regarding the 2019 edition, the proposed approaches outperform the benchmarks across all considered metrics, except for AHR, where only one combination manages to outperform that edition's best model. On the 2020 eRisk collection, we achieve: (1) superior item-level metrics; our approach outperforms existing benchmarks on granular metrics (AHR and ACR); (2) SOTA depression category accuracy (DCHR); we obtain the best result with a voting-regressor ensemble based on the rounded average scores of the top 3 closed-source approaches with the highest DCHR scores in 2019 evaluations (as reported in Table 4). These results are particularly notable as we maintain high performance without requiring training data, unlike all previous approaches reported that rely on the 2019 dataset for supervision. Interestingly, as shown in Table 4, our approach demonstrates consistent performance across both eRisk editions despite their different category distributions (see Table 2). Specifically, our method maintains stable performance on three key metrics (ADODL, AHR, and ACR), with only DCHR showing variation between editions. This negative result is 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** (Beck, 1996; Warmenhoven et al., 2012). 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. In Section E.1 of the Appendix, we further justify our aRAG approach's effectiveness by comparing it against non-

RAG baselines, where we test closed-source LLMs (`gpt-4o-mini` and `Claude-3.5-Sonnet`) using direct input of all posts within their context window.

5.2 Beyond Depression: Identifying Signs of different MHCs through questionnaire

In this section, as a proof-of-concept, we extend the use of questionnaires to different mental health conditions, such as self-harm, anorexia, and pathological gambling. 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 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 posts using SHI questions as queries, followed by LLM-generated responses. For SHI, since each item has binary yes/no responses, we use only the questions as retrieval queries. 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. Figure 4 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 approach achieves superior precision in the 2020 edition. These results are particularly significant considering that our approach do not need any training data and requires no training data, yet achieves competitive performance compared to fully supervised models that leverage extensive domain-specific training.

We further applied the same methodology to the early detection of anorexia and pathological gambling, using the SCOFF questionnaire and the Pathological Gambling Diagnostic Form respectively. In both cases, we used the same pipeline as in the SHI scenario. As reported in Figure 4, our method achieves the highest F1 score for anorexia and gambling tasks, outperforming the top-scoring systems of the respective editions (Mohammadi et al., 2019; Ragheb et al., 2019; Mármol-Romero et al., 2022; Fabregat et al., 2022) in this specific metric. These findings reinforce the generalizability of our adaptive RAG framework, demonstrating that standardized psychological questionnaires can effectively guide LLMs to detect a wide range of MHCs in a fully unsupervised setting. 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.

Even though in some cases the results do not outperform the benchmark models, this may be attributed not only to the lack of training data in our approach, but also to the fact that our questionnaire-guided approach occasionally fails to find sufficient evidence to support a diagnosis based on the psychometric tool used. As a result, our approach tends to be more conservative in assigning final diagnoses, sometimes refraining from matching the diagnostic criteria defined by the questionnaires unless the retrieved content provides clear and consistent indicators.

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. To prevent trivial classification by the LLM, we removed the word depression and related terms from posts in which users explicitly self-diagnosed.

We tested the same aRAG approach that proved effective for previous tasks, using the same combination of methods but guided by different types of

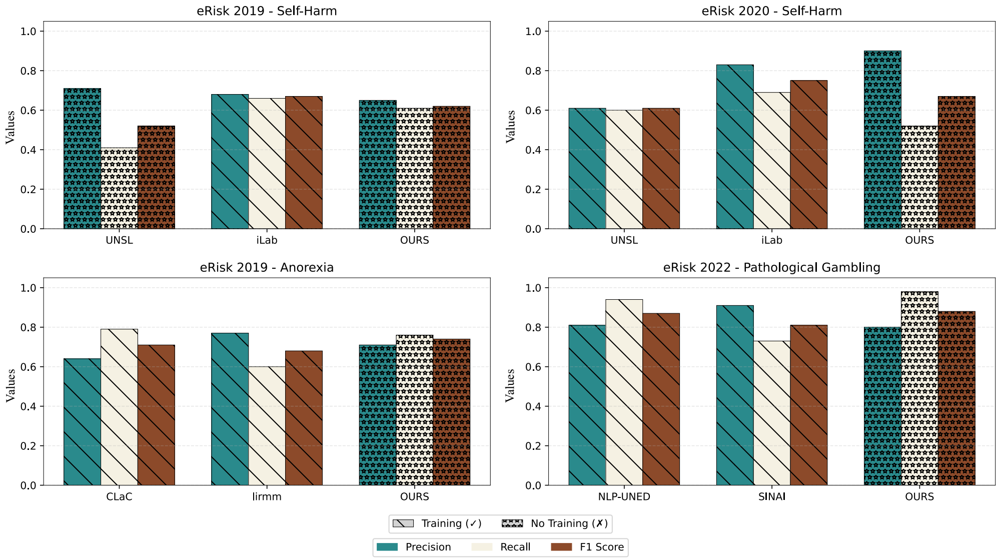


Figure 4: Performance of our approach compared to the best approach in each of the two eRisk editions for the detection of self-harm, Anorexia, and Pathological Gambling. Models with ✓ use training data while models with ✗ are fully unsupervised. Precision, Recall, and F1 metrics are reported.

	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 (subscale)	14	0.36	0.64	0.46
	BDI-II	20	0.42	**0.76**	0.54
	Quest. Ensemble		0.53	0.69	**0.60**

Table 6: 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.

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. While for BDI-II we used choice options as queries, for PHQ-9 and DASS-42 we only used the questionnaire items as queries since only the questions contain textual content.

Each of the three questionnaires includes an established optimal cut-off score, denoted as τ in Table 6, for identifying clinically significant depression, specifically at the moderate severity threshold. The table also reports the results obtained 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

We introduce a novel aRAG approach that leverages standardized psychological questionnaires to guide LLMs in mental health screening, requiring no training data. We demonstrate the advantage of our approach in supporting mental health screening tasks. The results show the advantages of our approach in: automatically completing questionnaires **(RQ1.)**, proving effective not only for depression screening but also extending successfully to other conditions like self-harm detection, anorexia, and pathological gambling, as well as potentially other MHCs **(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.)**, by providing a structured interpretation of the user's psychological state and enabling an estimation of the severity level through clinically questionnaire scores.

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 MHCs.

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. Additionally, such cut-offs may not be fully consistent in the context of social media analysis and may require further adaptation.

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.

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.

Mohammad Atari, Ali Omrani, and Morteza Dehghani. 2023. Contextualized construct representation: leveraging psychometric scales to advance theory-driven text analysis.

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. 1996. Manual for the beck depression inventory-ii. *(No Title)*.

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.

Ana-Maria Bucur, Adrian Cosma, and Liviu P Dinu. 2022. Life is not always depressing: Exploring the happy moments of people diagnosed with depression. *arXiv preprint arXiv:2204.13569*.

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)*.

Alessandro De Grandi, Federico Ravenda, Andrea Raballo, and Fabio Crestani. 2024. The emotional spectrum of llms: Leveraging empathy and emotion-based markers for mental health support. *arXiv preprint arXiv:2412.20068*.

Francesco Denti, A Di Noia, A Mira, et al. 2023. Bayesian nonparametric estimation of heterogeneous intrinsic dimension via product partition models. In *Book of the Short Papers SEAS IN 2023*, pages 316–321. Pearson.

Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. 2022. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12(1):20005.

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.

Hermenegildo Fabregat, Andres Duque, Lourdes Araujo, and Juan Martinez-Romo. 2022. Uned-nlp at erisk 2022: Analyzing gambling disorders in social media using approximate nearest neighbors. In *CLEF (Working Notes)*, pages 894–904.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. 2017. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140.

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.

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.

Siran Li, Linus Stenzel, Carsten Eickhoff, and Seyed Ali Bahrainian. 2025. Enhancing retrieval-augmented generation: A study of best practices. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6705–6717, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Alba María Mármol-Romero, Salud María Jiménez Zafra, Flor Miriam Plaza del Arco, M Dolores Molina-González, María Teresa Martín Valdivia, and Arturo Montejo-Ráez. 2022. Sinai at erisk@ clef 2022: Approaching early detection of gambling and eating disorders with natural language processing. In *CLEF (Working Notes)*, pages 961–971.

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)*.

Elham Mohammadi, Hessam Amini, and Leila Kosseim. 2019. Quick and (maybe not so) easy detection of anorexia in social media posts. In *CLEF (Working Notes)*.

John F Morgan, Fiona Reid, and J Hubert Lacey. 1999. The scoff questionnaire: assessment of a new screening tool for eating disorders. *Bmj*, 319(7223):1467–1468.

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.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2022. erisk 2022: pathological gambling, depression, and eating disorder challenges. In *European Conference on Information Retrieval*, pages 436–442. Springer.

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*.

Waleed Ragheb, Jérôme Azé, Sandra Bringay, and Maximilien Servajean. 2019. Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media. In *CLEF (Working Notes)*.

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 *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pages 1014–1018.

Esteban A Ríssola, Seyed Ali Bahrainian, and Fabio Crestani. 2019. Anticipating depression based on online social media behaviour. In *Flexible Query Answering Systems: 13th International Conference, FQAS 2019, Amantea, Italy, July 2–5, 2019, Proceedings 13*, pages 278–290. Springer.

Esteban A Ríssola, Seyed Ali Bahrainian, and Fabio Crestani. 2020. A dataset for research on depression in social media. In *Proceedings of the 28th ACM conference on user modeling, adaptation and personalization*, pages 338–342.

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.

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.

Vasudha Varadarajan, Allison Lahnala, Adithya V Ganesan, Gourab Dey, Siddharth Mangalik, Ana-Maria Bucur, Nikita Soni, Rajath Rao, Kevin Lanning, Isabella Vallejo, et al. 2024. Archetypes and entropy: Theory-driven extraction of evidence for suicide risk. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 278–291.

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.

Franca Warmenhoven, Eric van Rijswijk, Yvonne Engels, Cornelis Kan, Judith Prins, Chris van Weel, and Kris Vissers. 2012. The beck depression inventory (bdi-ii) and a single screening question as screening tools for depressive disorder in dutch advanced cancer patients. *Supportive care in cancer*, 20:319–324.

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.

WHO. 2022. Who special initiative for mental health (2019-2023): Universal health coverage for mental health.

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 ABIDE

Adaptive Binomial Intrinsic Dimension Estimator (ABIDE) (Noia et al., 2024) is an approach to estimate the Intrinsic Dimension (ID) of the data. The idea of ID is to quantify the complexity of high-dimensional datasets. In essence, it represents the minimum number of variables needed to describe the underlying structure of data without significant loss of information (Denti et al., 2022, 2023). In fact, in many real-world scenarios, especially for what concerns text, there are often hidden relationships and dependencies among these features. This means that the data might actually lie on a lower-dimensional manifold embedded within the high-dimensional space.

However, estimating ID can be difficult, especially in real-world scenarios, where datasets often have complex structures and ID can vary depending on the scale at which the data are observed.

ABIDE addresses the scale-dependency challenge through a novel adaptive approach. At its core, ABIDE uses the concept of k^*, which represents the optimal neighborhood size for each data point. k^* is not fixed across all the observations, but varies for each point, allowing the method to adapt to local data characteristics. The algorithm works iteratively. It starts with an initial ID estimate, using the 2NN method (Facco et al., 2017). Then, for each data point, it determines the largest neighborhood (k^*) where the data density can be considered approximately constant. Using these k^* values, ABIDE recalculates the ID estimate. This process is repeated, refining at each iteration both the ID estimate and the k^* values for each point. The iteration continues until convergence is reached.

C Dense Retrieval Models

We use a pool of different dense retrieval models: *msmarco-MiniLM-L-6-v3*[2], *msmarco-MiniLM-L-12-v3*[3], *msmarco-distilbert-base-v4*[4], *sentence-*

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

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

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

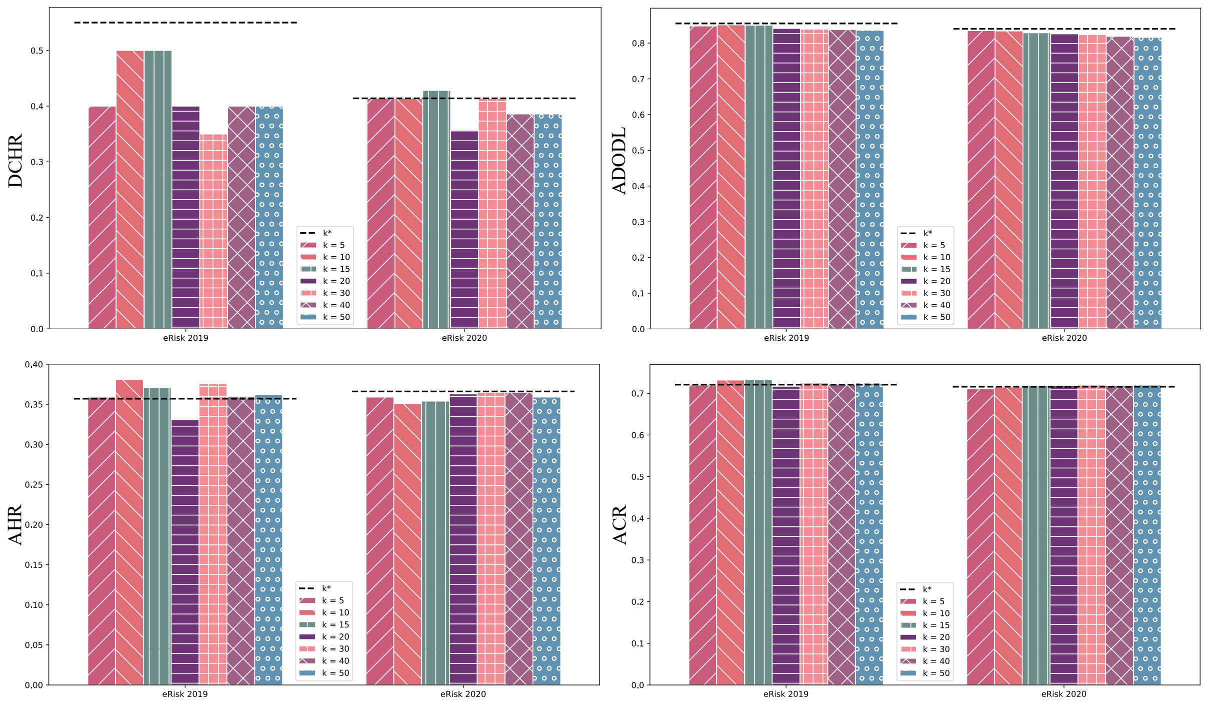


Figure 5: Comparison of the performance of the gpt-4o-mini + Direct Prompt combination with MiniLM-L12-v3 retrieval model on the eRisk 2019 and 2020 datasets, varying by k value used to select the number of documents to retrieve. k^* represent our adaptive method, and $k = 15$ the average number of k^*.

T5-base[5], *msmarco-distilbert-base-tas-b*[6], *all-mpnet-base-v2*[7], *GIST*[8], *sf-model-e5*[9], *contriever-msmarco*[10], *bge-large*[11]. 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.

D 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,

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

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

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

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

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

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

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

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.

E Ablation Study

E.1 Justification of the RAG approach

To further justify our retrieval-based approach, aka aRAG, we compare the performance of the most effective DCHR configurations using two closed-source LLMs against their performance when directly prompted to answer BDI-II items without filtering relevant posts. Specifically, we compare Claude 3.5 Sonnet + Direct prompt combined with T5 and gpt-4o-mini + Direct prompt combined with MiniLM-L12-v3. Given that some users have a significantly high number of posts, we input all posts that fit within each LLM's context window based on timestamp order. We focus on the questionnaire-level metrics, DCHR and ADODL, which allow us to assess the accuracy of both approaches in determining depression severity. As shown in Figure 6, the aRAG approach consistently outperforms its no-retrieval counterpart across both collections and LLMs. Notably, we observe particularly wide performance gaps when using gpt-4o-mini on the eRisk 2019 collection (55.0% vs 35.0% for DCHR and 85.5% vs 82.0% for ADODL) and Claude on the 2020 edition (47.1% vs 31.4% for DCHR and 83.9% vs 80.9% for ADODL).

E.2 The Impact of the Adaptive k^* Dense Retrieval Approach

Figure 5 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 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`	🟥	†	🟥	🟥
`Qwen 2.5 70B`	🟥	🟥	🟥	🟥
`DeepSeek V3`	🟥	†	🟥	🟥
`Phi-3-mini`	🟥	‡	‡	‡
`Phi-3.5-mini`	🟥	‡	‡	‡

Table 7: 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.

E.3 The Impact of Different Prompting Strategies

In Table 7, 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, $\alpha = 0.05$). Lightweight open-source models (`Phi-3-mini` and `Phi-3.5-mini`) perform well only in direct prompt contexts, while they perform poorly when the prompting technique is CoT, tending to overestimate questionnaire scores (difference is statistically significant for ADODL, AHR, and ACR metrics).

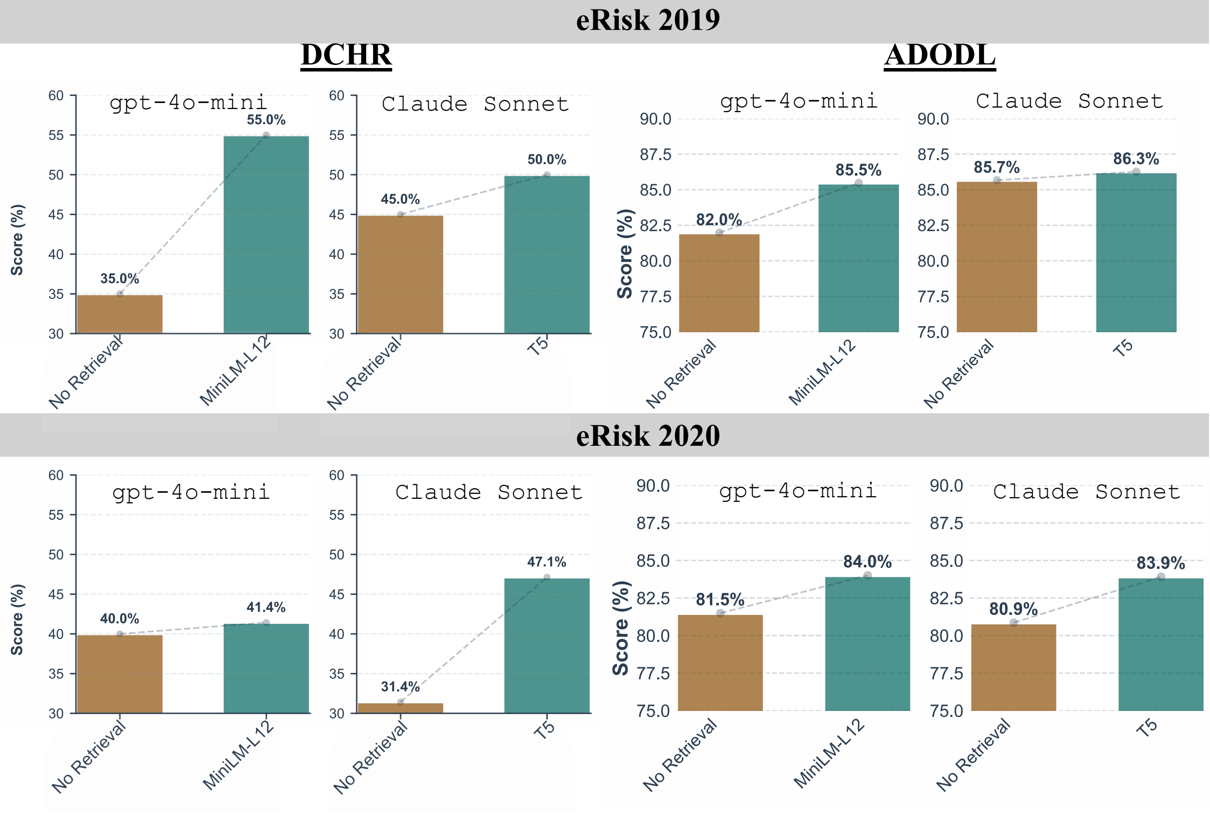


Figure 6: Performance comparison between aRAG and no-retrieval approaches for closed-source LLMs across DCHR and ADODL metrics on eRisk 2019 and 2020 datasets

For closed-source (gpt-4o-mini, Claude-3.5-Sonnet) and large-scale open-source LLMs (Qwen 2.5 70B, DeepSeek V3), the performance difference between prompting strategies varies. Specifically, gpt-4o-mini shows significant differences only under t-test across most metrics. Claude-3.5-Sonnet and DeepSeek V3 demonstrate significant differences solely in ADODL according to t-test, while Qwen 2.5 70B shows no statistically significant differences across any metric.

E.4 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 7, we examine the distribution of LLMs scores across different retrieval approaches and their rankings for each metric w.r.t. eRisk 2019 dataset collection (after removing anomalous values given by lightweight open-source models in combination with CoT prompting strategies). While no single retrieval 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 on average.

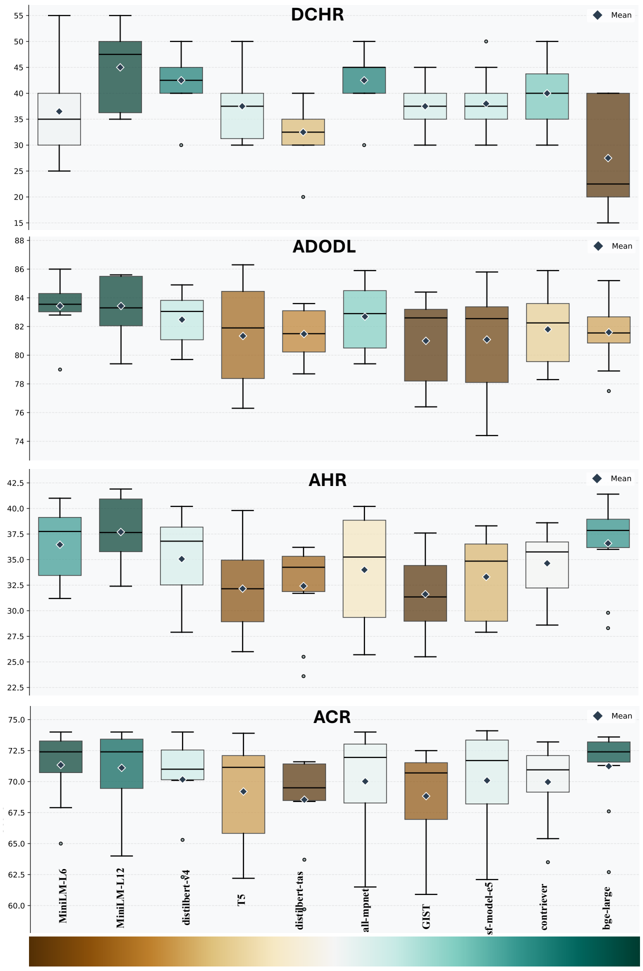


Figure 7: Distribution of adaptive RAG scores conditioned on different retrieval approaches and their rankings for each metric on the eRisk 2019 dataset. Rankings are based on mean scores (shown as diamonds). The color gradient from brown to teal indicates performance ranking, with darker teal representing better performance.