# Overview of the CLPsych 2024 Shared Task: Leveraging Large Language Models to Identify Evidence of Suicidality Risk in Online Posts

**Jenny Chim**[1*], **Adam Tsakalidis**[1,2*], **Dimitris Gkoumas**[1], **Dana Atzil-Slonim**[3], **Yaakov Ophir**[4,5], **Ayah Zirikly**[6], **Philip Resnik**[7], **Maria Liakata**[1,2]
[1]Queen Mary University of London (UK), [2]The Alan Turing Institute (UK),
[3]Bar Ilan University (Israel), [4]Ariel University (Israel), [5]University of Cambridge (UK),
[6]Johns Hopkins University (US), [7]University of Maryland (US)
{c.chim; a.tsakalidis; m.liakata}@qmul.ac.uk

## Abstract

We present the overview of the CLPsych 2024 Shared Task, focusing on leveraging open source Large Language Models (LLMs) for identifying textual evidence that supports the suicidal risk level of individuals on Reddit. In particular, given a Reddit user, their predetermined suicide risk level ('Low', 'Moderate' or 'High') and all of their posts in the *r/SuicideWatch* subreddit, we frame the task of identifying relevant pieces of text in their posts supporting their suicidal classification in two ways: (a) on the basis of evidence highlighting (extracting sub-phrases of the posts) and (b) on the basis of generating a summary of such evidence. We annotate a sample of 125 users and introduce evaluation metrics based on (a) BERTScore and (b) natural language inference for the two sub-tasks, respectively. Finally, we provide an overview of the system submissions and summarise the key findings.

## 1 Introduction

Recent statistics on mental health related problems during and after the COVID-19 pandemic are striking. In the US, almost 50% of adults aged 18-44 reported a mental illness in 2023,[1] whereas similar rates of the EU population had experienced emotional or psychosocial problems between June 2022-23.[2] Partially due to the limited accessibility of support services, individuals often seek support in online social media by sharing their thoughts and concerns and engaging in discussions with their peers. Research at the intersection of natural language processing (NLP) and mental health has focused on exploiting such user generated content in order to automatically detect vulnerable users (Coppersmith et al., 2015; Shing et al., 2018; Zirikly et al., 2019) or monitor their well-being over time (Tsakalidis et al., 2022b; Tseriotou et al., 2023). However, in real-world scenarios, detection is only part of the need: downstream evaluation and intervention would be facilitated by an understanding of *why* a user's text led them to be flagged (Ophir et al., 2022).

Large language models (LLMs) (Brown et al., 2020; Sanh et al., 2022; Chowdhery et al., 2022; Touvron et al., 2023) are currently dominating the field of NLP. Work at the intersection of NLP and mental health has leveraged such models for classification (Amin et al., 2023), data augmentation (Liyanage et al., 2023) or reasoning (Xu et al., 2023), among others. Recent research explores the language understanding and mental health reasoning capabilities of LLMs using instruction fine-tuning and Chain-of-Thought prompting (CoT) (Yang et al., 2023; Xu et al., 2023). Instead of direct phrase extraction, LLMs are instructed to provide step-by-step reasoning, leveraging inherent knowledge to generate human-like language (Xu et al., 2023). Such approaches pose the risk of incorrect predictions and flawed reasoning, especially in complex conversations (Li et al., 2023).

This year's CLPsych Shared Task focused on *leveraging open source LLMs for the purpose of finding evidence in online posts that supports the level of suicidal risk of their author*. In particular, we define two sub-tasks (thereafter 'tasks') on the basis of (a) highlighting and (b) summarising such supporting evidence. Working with the UMD Reddit Suicidality dataset (Shing et al., 2018; Zirikly et al., 2019), we present the process of defining the task (Section 3), selecting and annotating a subset of 125 Reddit users (Section 4), introducing our evaluation metrics (Section 5) and summarising the approaches and the best-performing system of each team (Section 7).

In this overview paper we make the following

[*]Denotes equal contribution.

[1]https://www.apa.org/news/press/releases/2023/11/psychological-impacts-collective-trauma

[2]https://europa.eu/eurobarometer/surveys/detail/3032



*Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 177–190

contributions:

- we introduce two novel tasks on identifying evidence that supports the suicidal risk level of a particular user;
- we describe the annotation process;
- we provide an overview of the approaches followed by the participating teams, our evaluation approach and an overview of the results.

## 2 Related Work

### 2.1 NLP and Mental Health

Related work during the last decade has been primarily focusing on classifying documents (Sawhney et al., 2022a) or users, with the latter being performed at a static (Coppersmith et al., 2015; Shing et al., 2018; Zirikly et al., 2019; Sawhney et al., 2022b) (e.g., suicide level of an individual) or a longitudinal basis (Tsakalidis et al., 2022b,a; Hills et al., 2023). Recent work has started paying attention to more fine-grained analysis with respect to mental health as well as explaining model predictions. The 2023 eRisk Task 1 focused on ranking of sentences based on their relevance to depressive symptoms (Parapar et al., 2023). Nguyen et al. (2022) proposed a spectrum of BERT-based methods for depression detection that are constrained by the presence of PHQ-9 symptoms for improved generalizability and interpretability of the models. Nemesure et al. (2021) used SHAP values (Lundberg and Lee, 2017) to explain predictions for generalized anxiety and major depressive disorder prediction models. Zirikly and Dredze (2022) used the PHQ-9 questions as auxiliary tasks to provide explanations for a depression detection model using LIME (Ribeiro et al., 2016) and measured performance on a manually annotated sample of highlighted text spans. Garg (2024) also annotated a dataset with highlighted text spans over several 'wellness' dimensions. In this year's Shared Task we also highlight text spans of online posts, which serve as evidence for an online user's suicide risk level, and we further accompany this with a summarisation of such evidence found at the user level. The task then sets out to explore to what extent such text spans and summaries can be obtained by leveraging open source LLMs.

### 2.2 LLMs for evidence extraction

The use of Large Language Models (LLMs) in evidence extraction is an ongoing area of research and discussion. LLMs have shown promise in retrieving supporting evidence for generated responses and in self-detecting hallucinations within them (Huo et al., 2023). In the context of medical evidence, domain-agnostic LLMs, like GPT-3, have been found to be potentially precise at zero- and few-shot information extraction from clinical unstructured texts (Agrawal et al., 2022), yet prone to inconsistent generated summaries, raising concerns about potential harm due to misinformation (Tang et al., 2023). In NLP for mental health, existing work has predominantly explored the capabilities of LLMs to predict critical mental states (e.g., stress and depression) or high-risk actions (suicide) by forcing LLaMA-2 or GPT3 to act as an expert in a zero- or few-shot setting (Lamichhane, 2023; Amin et al., 2023; Yang et al., 2023). Other work has systematically explored the mental health reasoning capabilities of various LLMs in an instruction fine-tuning setting, employing CoT prompting to elucidate the reasoning behind their predictions (Yang et al., 2023; Xu et al., 2023). However, these approaches do not directly extract precise phrases from the text. Instead, they instruct LLMs to provide step-by-step reasoning or explanations for their output, leveraging inherent knowledge and paraphrasing the text to generate human-like natural language based on embedded knowledge (Xu et al., 2023). This could result not only in incorrect predictions but also in flawed reasoning processes, particularly in more complex conversation contexts (Li et al., 2023).

### 2.3 LLMs for Summarisation

LLMs have demonstrated promising summarisation performance across document types including news articles (Goyal et al., 2022; Zhang et al., 2023b) and instructional texts (Maynez et al., 2023), and have shown significant improvements in challenging areas such as meeting transcripts (Laskar et al., 2023) and long narratives (Chang et al., 2024). While most use simple prompts (e.g. "*Summarize the following article:*"), prior work on news (Wang et al., 2023) and social media (Song et al., 2024) suggest that multi-step prompting strategies with prompt design informed by domain expertise can steer models to produce improved information-rich summaries. Nonetheless, how to effectively leverage the generative capabilities of LLMs while ensuring outputs are grounded in supporting evidence and consistent with expert knowledge remains an ongoing research problem,

especially in high stake applications such as mental health.

## 3 Task Definition and Instructions

We define two tasks aimed at leveraging LLMs in order to find evidence within text that has been shared by particular online social media users supporting their pre-assigned Suicide Risk Level ('Low', 'Moderate' or 'High'). The distinction between the two tasks is based on the way that this evidence is expected to be provided.

**Task A** For our first task, participants were asked to provide the evidence supporting the pre-defined Suicide Risk Level of a user by *highlighting* relevant phrases within the text posted by the user. Each user could have multiple posts in the dataset; Task A was defined at the post (document) level – i.e., highlighting relevant phrases within each post made by a particular user.

**Task B** Our second task required *generating a summary* of evidence supporting a user's assigned risk level, across multiple posts made by the user. As opposed to Task A, Task B was performed at the user level – i.e., generating a single summary per user. Summaries were limited to 300 tokens.

No ground truth data were provided to the teams, except for a single example of a user with a pre-assigned Suicide Risk Level for whom we shared the expected highlights (Task A) and summary (Task B), as annotated by our experts (see Section 4.2). Compared to the latest CLPsych Shared Tasks, where the expected outputs were a class label either at the user level (Zirikly et al., 2019) or on a longitudinal basis (Tsakalidis et al., 2022a), this year's edition was considerably more open-ended. We therefore provided a list of 'aspects to consider' to the teams, which were compiled on the basis of our internal annotation instructions. These aspects, which were based on literature on suicidal risk (see Section 4.2), included, but were not limited to, the following:

- **Emotions**: How does the individual feel? From feeling sad to experiencing unbearable psychological pain, the self-disclosed emotions of the user could play an important role in the risk level assigned to the individual.
- **Cognitions**: What are the individual's thoughts and perceptions about suicide? For example, what is the level and frequency of suicidal thoughts? Does the individual intend to self-harm/suicide? Does the individual have a plan about it?
- **Behaviour and Motivation**: What are the individual's acts or behavior related to suicide? For example, do they have access to means and a concrete plan? What is the user's ability to handle difficult/stressful situations ('behaviour')? What is the motivation behind their wish to be dead?
- **Interpersonal and social support**: Does the individual have social support/stable relationships? How does the individual feel towards significant others?
- **Mental health-related issues**: Consider psychiatric diagnoses associated with suicide such as schizophrenia, bipolar/anxiety/eating disorder, previous self-harm/suicidal attempts and others.
- **Context/additional risk factors**: For example, socioeconomic and demographic factors, exposure to suicide behaviour by others, chronic medical condition, ...

Each team was allowed to provide (up to) three submissions for each task. Additional submissions were also allowed in order to facilitate ablation and further analysis by the teams, but were not included in our official results presented in this overview paper. Upon receiving the submissions, we returned the results based on our evaluation metrics (see Section 5) on a test set of 125 users (see Section 4).

## 4 Data and Annotation

### 4.1 Data

We use a subset of the The University of Maryland (UMD) Reddit Suicidality Dataset, Version 2 (Shing et al., 2018; Zirikly et al., 2019) for both tasks. The dataset contains posts made by a larger number of Reddit users.[3] The data was previously annotated at the user level with respect to level of suicide risk ('No', 'Low', 'Moderate' or 'Severe' risk labels), where the main difference between Moderate and Severe is that the latter indicates imminent or crisis-level risk. This annotation was performed in two ways (by (a) crowdsourcing and

[3] https://www.reddit.com/

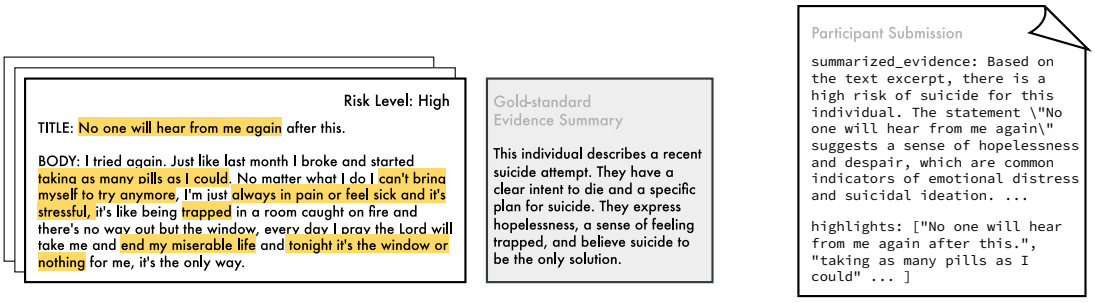


Figure 1: Example posts, gold evidence spans and summary, and corresponding submission data. Texts have been paraphrased for privacy. Participants are provided with posts and user-level risk labels, then asked to predict supporting evidence spans (Task A) and synthesise summaries (Task B). Each user can have multiple posts.

(b) experts, where the experts annotated a subset of the users) and it involved an annotator reading all of the posts that a user had made in the *r/SuicideWatch* subreddit in order to make a labelling decision for that particular user.

The inter-annotator agreement was higher amongst the expert annotators (Shing et al., 2018); we therefore ignored the crowdsourced annotations for this Shared Task and focused strictly on the 245 users annotated by the experts. Since our task involves finding evidence about the suicide risk level of a particular user, we only kept the 'Low', 'Moderate' and 'Severe' classes (209 users) and ignored the 'No risk' category. Also, since the original annotation was performed on the basis of the *r/SuicideWatch* posts only, we further focused explicitly on those 332 posts made by the 209 users. Lastly, we selected 125/209 users (162/332 posts) to be annotated by our annotators (see Section 4.2) and serve as our ground truth during the evaluation stage of the Shared Task. This final selection was based on (a) filtering out any users whose posts were very short, (b) ignoring users with more than 3 posts in *r/SuicideWatch* to accommodate faster annotation (i.e., prioritising more users instead of more posts in our evaluation data) and (c) prioritising the inclusion of 'Severe', followed by 'Moderate' risk users. In the end, 93 users had only one post, 27 users had two posts and five users had three posts. Table 1 shows the overall numbers of users and posts in *r/SuicideWatch* that were selected for annotation purposes (and therefore, our gold standard during evaluation), as described next.

| | No | Low | Moderate | Severe | **Total** |
|---|---|---|---|---|---|
| Original (users) | 36 | 50 | 115 | 44 | 245 |
| Annotated (users) | – | 13 | 74 | 38 | 125 |
| Original (posts) | 45 | 77 | 162 | 93 | 377 |
| Annotated (posts) | – | 17 | 91 | 54 | 162 |

Table 1: Summary of the data that was annotated in this Shared Task and used as our ground truth, compared to the original UMD Reddit Suicidality Dataset.

### 4.2 Annotation

The annotators were two graduate students (fluent English speakers) in a clinical psychology training program at Bar-Ilan University. Their task was to read the posts of each user on *r/SuicideWatch*, and highlight text spans as evidence supporting the suicide risk level previously assigned by experts in Shing et al. (2018). Next, they were asked to synthesize the textual evidence and related clinical observations in a short summary.

Annotators were provided with detailed guidelines and expert annotated examples. The guidelines for the annotations were based on the clinical literature about suicidal risk (Posner et al., 2011; Turecki et al., 2019; Rogers et al., 2023) and their main aspects are provided as a list in Section 4.1. We conducted two rounds of training supervised by a senior clinical psychology expert. In each round, annotators labelled posts independently. We manually checked the agreement on these posts, then addressed areas of disagreement and clarified task guidelines in training meetings. Next, the team was asked to refine their existing annotations and work on new ones. We repeated this process until satisfactory agreement levels were obtained upon manual inspection, where the most important key phrases were captured by both annotators, and the summarised evidences were mutually consistent. Out of the 125 posts, 13 were labeled twice by both annotators. The final pairwise relaxed F1 (Hripcsak and Rothschild, 2005; Deléger et al., 2012) over evidence spans from these doubly annotated instances is .96.

## 5 Evaluation Metrics

### 5.1 Task A

The main metric we consider is the recall of evidence spans. For a given user, given predicted evidence spans $H$ and gold evidence spans $E$, we average the maximum recall-oriented BERTScore (Zhang et al., 2020):

$$\text{Recall} = \frac{1}{|E|} \sum_{e \in E} \max_{h \in H} R_{\text{BERT}}(e, h)$$

To provide a more holistic view of evidence identification performance, we compute precision by averaging the maximum precision-oriented BERTScore for each predicted evidence span $h \in H$ against each gold evidence span $e \in E$:

$$\text{Precision} = \frac{1}{|H|} \sum_{h \in H} \max_{e \in E} P_{\text{BERT}}(e, h)$$

We also report a weighted version of recall, which is sensitive to predicted evidence lengths relative to gold evidence lengths. For a given user with gold evidence spans of cumulative token count $n_{\text{gold}}$ and predicted spans with cumulative token count $n_{\text{pred}}$, if the predicted evidence spans are longer than the gold-standard ones, we apply weight $w$ to the user-level recall:

$$w = \begin{cases} \frac{n_{\text{gold}}}{n_{\text{pred}}} & \text{if } n_{\text{pred}} > n_{\text{gold}} \\ 1 & \text{otherwise} \end{cases}$$

Finally, we report F1, the harmonic mean between precision and unweighted recall, $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.

### 5.2 Task B

Following prior work in general domain (Maynez et al., 2020) and mental health summarisation (Song et al., 2024), we leverage predictions from a natural language inference (NLI) model (Laurer et al., 2024) for summary evaluation.[4] We consider consistency to be the absence of contradiction. For each sentence in a submitted summary $s \in S$, we use the NLI model to compute its mean probability of contradicting each sentence in the corresponding gold-standard evidence summary $g \in G$, taking the gold sentence as premise and the submitted sentence as hypothesis:

$$\text{CS} = \frac{1}{|S| \cdot |G|} \sum_{s \in S} \sum_{g \in G} (1 - \text{NLI}(\text{Contradict}|g, s))$$

To complement consistency, we also evaluate summaries by their contradiction to expert summaries. We expect there to be some natural contradictory information in most summaries, since summarised evidence can include both risk factors and protective factors. We compute the contradiction score by averaging the *maximum* contradiction probability of a predicted sentence against gold evidence summary sentences:

$$\text{CT} = \frac{1}{|S|} \sum_{s \in S} \max_{g \in G} \text{NLI}(\text{Contradict}|g, s).$$

## 6 Participating Process & Teams

### 6.1 Registration Process

The registration process included (a) a team member initialising the process by filling an online form as the team representative, (b) reading and signing a data sharing agreement and (c) receiving instructions on how to download the data in a password protected zip folder. Each team member would also sign up for their team upon completing an individual registration form. The (b) data sharing agreement (among others) prohibited transferring any part of the data to third party providers in order to use their LLMs.

### 6.2 Participating Teams

Overall 23 teams (75 members) registered for the task. Members of four teams mentioned that they had participated in a previous CLPsych Shared Task, whereas members of three teams stated that they had previous experience with the UMD Suicidality Dataset. 15 out of 23 teams submitted their outputs for either of the two tasks – a percentage of 65% compared to 60% for Shared Task 2022 (Tsakalidis et al., 2022a) – and 13 teams submitted a paper at the end of the Shared Task (see Table 2).

## 7 Results

### 7.1 Overview

**Task A** The results are summarised in Table 3. The highest evidence recall comes from systems employing different approaches, including relying on smaller expert models for sentence-level predictions (SophiaADS), CoT prompting (UoS NLP), and prompting then post-processing (UniBuc Archeology). To improve precision and reduce incorrect outputs (e.g., hallucinations and unintended text normalisation where LLM corrects

[4] https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

| Team Name | #Members | #Submissions Task A | Task B | Paper submitted |
|---|---|---|---|---|
| CSIRO | 6 | 3 | – | ✓ |
| DONUTS Colaboratory | 6 | 2 | 2 | |
| INF@UoS | 2 | 1 | 1 | ✓ |
| ISM | 2 | 3 | 3 | ✓ |
| LAMA | 3 | 3 | 3 | ✓ |
| MHNLP | 1 | 1 | 1 | |
| SBC | 1 | 3 | 3 | ✓ |
| SCALAR-NITK | 3 | 1 | 1 | ✓ |
| SKKU-DSAIL | 5 | 3 | 3 | ✓ |
| sophiaADS | 3 | 3 | 3 | ✓ |
| SWELL | 11 | 3 | 3 | ✓ |
| UniBuc Archaeology | 3 | 3 | 3 | ✓ |
| UoS NLP | 4 | 3 | 3 | ✓ |
| UZH_CLyp | 2 | 1 | 1 | ✓ |
| Xinhai | 3 | 3 | 3 | ✓ |
| **Total (sum)** | **55** | **36** | **33** | **13/15** |

Table 2: Summary of the team information and submissions for the CLPsych Shared Task 2024.

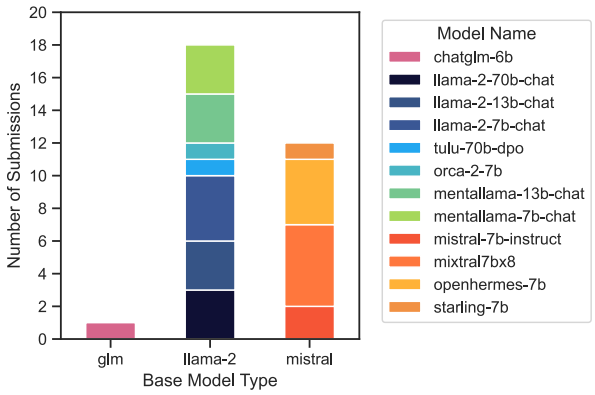


Figure 2: LLMs used in official submissions, grouped by model family and lineage.

typos in noisy user text), most teams applied post-processing procedures to align predicted spans to the original text, and some employed formal grammars to constrain model outputs (CSIRO, SBC).

**Task B** Submissions that achieved the highest consistency scores commonly incorporated domain knowledge, such as using expert models to retrieve emotionally charged text before summarising (UZH_CLyp), designing detailed instructions around the Shared Task guidelines (SBC, SWELL), and summarising evidence spans that were extracted based on psychology theory, e.g. Joiner's Interpersonal Theory of Suicide (SWELL). While there was no definitively superior LLM, top performing submissions on this task used Mistral (Jiang et al., 2023) and its derivative Openhermes[5], as well as LLaMA-2 (Touvron et al., 2023) and its mental health oriented derivative MentaLLaMA (Yang et al., 2023).

**LLM Characteristics and Resources.** As per the data use agreement, participants were forbidden from using Cloud APIs, relying on private and self-hosted instances. Figure 2 outlines the employed models. All submissions used instruction-tuned LLMs. The majority of submissions used models that are 7B or smaller (52%), the rest includes 13B and 8x7B mixture-of-expert models (35%) and 70B models (13%). Models were typically deployed with quantization, in some cases using libraries such as llama.cpp to run on consumer hardware.[6]

[5] https://huggingface.co/teknium/OpenHermes-2-Mistral-7B

[6] https://github.com/ggerganov/llama.cpp

### 7.2 Individual Team Submissions

**UoS_NLP** Singh et al. (2024) explored prompting strategies with Mixtral7bx8 (Jiang et al., 2024), a LLM with the same high-level architecture as Mistral (Jiang et al., 2023) but utilising mixture-of-experts layers, and Tulu-2-DPO-70B (Ivison et al., 2023), LLaMA-2 further instruction finetuned using direct preference optimisation (Rafailov et al., 2023). Their best performing evidence extraction approach involved few-shot CoT prompting Tulu, choosing exemplars by embedding posts with social media fine-tuned RoBERTA (Barbieri et al., 2020) then applying k-means clustering and manually selecting representative examples. For evidence summarisation, their best approach involved zero-shot instruction prompting Mixtral with additional meta-information, i.e. inferred emotion, inferred sentiment, and suicide risk label.

**SCALAR-NITK** Koushik et al. (2024) used attention weights from hierarchical attention networks (Yang et al., 2016) to extract evidence spans. For evidence summarisation, they zero-shot prompted LLaMA-2-7B-chat, providing the content of the user's post(s) concatenated with their extracted evidence spans as input.

**LAMA** Alhamed et al. (2024) used LLaMA-7B-chat with instruction prompting. For evidence extraction, they zero-shot prompted the LLM and combined the outputs with keywords extracted using a suicide lexicon (Alhamed et al., 2022) as well as manually curated depression-related keywords. Evidence summaries were separately obtained by first prompting to provide explanations of the individual's suicide risk level then synthesising them.

| Team | Task A: Evidence Extraction | | | | | Task B: Evidence Summarisation | | |
|---|---|---|---|---|---|---|---|---|
| | Hybrid | Recall | Precision | Weighted Recall | Harmonic Mean | Hybrid | Mean Consistency | Max Contradiction |
| SophiaADS | ✓ | **.944** | .906 | .489 | .924 | ✓ | .944 | .175 |
| UoS NLP | ✓ | .943 | .916 | .527 | **.929** | | .966 | .107 |
| UniBuc Archaeology | | .939 | .890 | .390 | .914 | | .973 | .081 |
| ISM | ✓ | .935 | .911 | .564 | .923 | ✓ | .961 | .125 |
| SKKU-DSAIL | ✓ | .922 | .912 | .549 | .917 | | .970 | .096 |
| CSIRO | | .919 | **.917** | .701 | .917 | | – | – |
| SWELL | ✓ | .915 | .892 | .542 | .903 | ✓ | .973 | .081 |
| UZH_CLyp | | .910 | .916 | .742 | .913 | ✓ | **.979** | .064 |
| MHNLP | ✓ | .910 | .888 | .197 | .909 | | .873 | .204 |
| SBC | | .907 | .912 | .738 | .909 | | .976 | .079 |
| Xinhai | | .887 | .906 | .617 | .911 | | .958 | .126 |
| SCALAR-NITK | ✓ | .886 | .893 | **.784** | .889 | | .901 | .233 |
| DONUTS Colaboratory | | .872 | .900 | .626 | .907 | | .942 | .159 |
| INF@UoS | ✓ | .850 | .893 | .630 | .896 | ✓ | .934 | .165 |
| LAMA | ✓ | .577 | .899 | .513 | .888 | | .964 | **.060** |

Table 3: Evaluation scores for Task A, by selecting the top-performing submission of each team on the basis of Recall. The associated Task B evaluation scores are shown on the right. 'Hybrid' denotes that the shown submission incorporated non-LLM techniques, including using inputs derived via non-LLM methods, and excluding standard post-processing. For details and methods explored in other submissions, please refer to Section 7.2.

| Team | Mean Consistency | Max Contradiction |
|---|---|---|
| UZH_CLyp | **.979** | .064 |
| UoS NLP | .977 | .079 |
| SBC | .977 | .083 |
| SKKU-DSAIL | .973 | .086 |
| SWELL | .973 | .081 |
| UniBuc Archaeology | .973 | .081 |
| LAMA | .964 | **.060** |
| ISM | .961 | .125 |
| Xinhai | .959 | .121 |
| SophiaADS | .944 | .175 |
| DONUTS Colaboratory | .942 | .159 |
| INF@UoS | .934 | .165 |
| SCALAR-NITK | .901 | .233 |
| MHNLP | .873 | .204 |

Table 4: Evaluation scores for Task B, by selecting the top-performing submission of each team on the basis of Mean Consistency.

**Xinhai** Zhu et al. (2024) used instruction prompting on a version of the open-source ChatGLM-3-6B (Du et al., 2022) model adapted to healthcare data. They revised their prompt using GPT-4. For evidence span extraction, they ensured LLM predictions obtained from instruction prompting were text spans directly present in the input texts using regular expressions and aligning phrases by their semantic similarity.

**SophiaADS** Tanaka and Fukazawa (2024) proposed a hybrid solution comprising task-specific models, handcrafted rules, and MentaLLaMA-chat-7b. For evidence extraction, they first picked sentences corresponding to high probabilities of suicide risk, as predicted by a bert-base-uncased classifier (Devlin et al., 2019). The latter was fine-tuned on a binary sentence-level suicide ideation dataset heuristically developed from the Shared Task data. In cases of insufficient evidence, they added the most negative sentences as predicted by a Tweet sentiment classifier (Barbieri et al., 2020), and supplemented with predictions from instruction-prompting MentaLLaMA as necessary. To summarise evidence, the team combined LLM summaries with rules that produce descriptions of risk level, posting behaviour, and several mental health related risk factors.

**ISM** Tran and Matsui (2024) leveraged Mixtral-8-7B-Instruct (Jiang et al., 2024) in two distinct stages: a) knowledge self-extraction and b) output refinement. During the knowledge self-extraction phase, participants provided users' posts along with the associated risk levels, prompting the model to address the task. The resulting output comprises a set of generated highlights, summaries, and identifications of suicide risks. Next, they selected the most optimal generated outputs aligned with the risk level to enrich the model's knowledge in stage 2, creating an enhanced and knowledge-rich representation (i.e., concatenation of the best knowledge responses). In the final step, the model underwent an iterative refinement process, continuously prompting for adjustments to the newly generated summaries and highlights until no further changes were observed.

**CSIRO** Chen et al. (2024) introduced instructive prompting for a range of psychological and socioeconomic factors to extract evidence aligned with users' suicidal risk from LLaMA-2-70b-chat

in a zero-shot setting. They investigated prompt engineering approaches across three different variations: a) A naive approach, instructing the model to extract phrases as evidence supporting suicide risk, thereby evaluating the inherent knowledge of the model. b) They enhanced the input content with a collection of psychological and socioeconomic factors, namely factor-orientied instruction. c) Finally, they reformulated the risk levels provided by annotators into a set of selected risk factors. The model was then guided by rules to choose the most appropriate prompt based on a user's risk level.

**SBC** Blanco-Cuaresma (2024) investigated open-source LLMs – OpenHermes, Orca2, Starling 7B alpha – in a one-shot setting. They employed the same crafted prompts, consisting in prefixed psychological and social factors provided by the organizers, to extract evidence from users' posts or to summarize evidence associated with their risk level. When extracting evidence, they utilized Backus-Naur Form (BNF), which is a metasyntactic notation for context-free grammars. This approach ensured that the order of words in the generated output matched the order of those in the users' posts.

**INF@UoS** Preiss and Chen (2024) proposed a two-stage pipeline to address span extraction and summarization related to suicidal risk levels. In the first stage, they fine-tuned a suicide risk classifier, i.e., MentalRoBERTa (Ji et al., 2022). Additionally, they employed Linguistic Inquiry and Word Count (LIWC-22) to extract informative features from the language, including desire for connection, certainty, negative tones, emotions, negative emotions, sadness emotions, mental health behavior, persuasiveness, and feelings. The additional extracted information was integrated with users' posts to train the classifier. Subsequently, SHAP was utilized to identify crucial phrases from the input content that contributed to the classifier's decision. In stage two, they prompted Mistral-7B-Instruct to generate summaries across five diverse factors—emotion, cognition, social support, mental health issues, and conceptual risk—using the extracted phrases from stage one.

**SKKU-DSAIL** For Task A, Jeon et al. (2024) prompted MentaLLaMA by assigning it a 'psychiatrist' identity and further providing it with (a) an example (partially highlighted) post, (b) a list of suicide-related words present in the post (Lee et al., 2022), (c) the post under consideration and (d) the suicide risk level of its author. For Task B, they used a similar setting, followed by two methods ('extract-then-generate' (Zhang et al., 2023a), integrating the highlighted phrases from Task A, and SOLAR (Kim et al., 2023)) for tackling hallucinations and inconsistencies in the generated summaries.

**UZH_CLyp** Uluslu et al. (2024) provided Mistral-7B-Instruct with the post and the author's label, asking it to extract the highlights for Task A as a suicide prevention therapist expert. For Task B, the levels of three emotions were calculated at the post-level and the top-5 saddest posts were included in the prompt (alongside the post, the user's risk level and the emotions) in order to generate the summary. In their ablation analysis, the authors showed that selecting the top-5 saddest posts had a large (positive) impact on model performance.

**SWELL** For Task A, Varadarajan et al. (2024) followed three approaches: (a) they constructed 'suicidality archetypes' on the basis of Joiner's IPTS (Joiner, 2007) in order to calculate their similarity against the sentence embeddings of a given post and extract the spans with the highest similarity; (b) they fine-tuned separate LMs using data from users with different suicidal risk levels and calculated the difference in entropy between these models for each sentence in a given post (Lahnala et al., 2021); (c) they prompted LLaMA-2 to extract sentences signalling any of the three main Joiner's IPTS constructs. For Task B they prompted LLaMA-2 in a few-shot setting, providing it with highlights and asking it to generate a summary by considering the six aspects present in Section 3.

**UniBuc Archaeology** Sandu et al. (2024) experimented with 'traditional' NLP approaches and LLMs: (a) for Task A, they used SHAP (Lundberg and Lee, 2017) on the outputs of a logistic regression trained to split 'No' vs 'Low/Moderate/High' risk users on the basis of tfidf ngrams and performed Task B as an extractive summarisation task; (b) they prompted OpenHermes 2.5 based on Mistral for extracting highlights and summarising the evidence.

### 7.3 Performance by Risk Level

Figure 3 summarises performance on test users at each risk level aggregated over all submissions. For the complete table of performance per team, see Table 3. While mean evidence recall values are

relatively similar, for precision and metrics assessing summary consistency the lower the risk level the lower the average performance. This suggests that linguistic cues for lower to moderate risk can be subtler compared to those of higher risk levels, and it may be more challenging to describe protective factors and the *absence* of risk factors. Future approaches should aim to more fully capture the nuances within the spectrum of suicide risk factors.

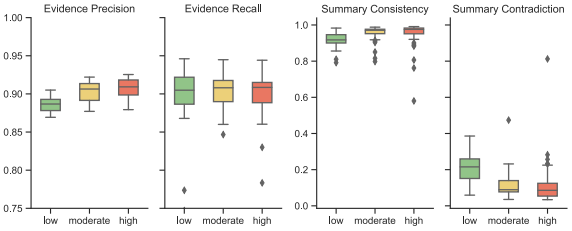


Figure 3: Mean performance by user's risk level. From left to right: evidence precision, recall, summary consistency, summary contradiction. Higher is better except for summary contradiction.

## 8 Conclusion

This work presented the overview of the CLPsych Shared Task 2024, focusing on leveraging open source LLMs to find supporting textual evidence for the suicide risk level of an online user, based on their online posts. We defined two tasks for finding such evidence – based on (a) text highlighting of relevant spans at the post-level and (b) summarising the evidence at the user-level. We generated a dataset of 125 social media users to facilitate evaluation and introduced the associated evaluation metrics for measuring system performance. Lastly, we have summarised the approaches taken by 13 teams and provided an overview of their results, their commonalities and novel aspects of their work.

## Limitations

As in the vast majority of prior work on leveraging social media for user-level mental health assessments, this year's Shared Task involved users who were classified with respect to their suicide risk level on the basis of content they generated. This implies that the annotation of their suicidality risk level, as well as this Shared Task's additional annotations (see Section 4.2), have been made on the basis of self-report. Moreover, the present tasks were conducted using social media posts made on a particular subreddit in the English language, by users who willingly self-disclosed their thoughts and feelings. Generalisation of the approaches presented in this work to other contexts (e.g., in psychotherapy sessions) remains an open question. Lastly, we have examined the presence of evidence around suicidality at the post-level; importantly longitudinal linguistic cues that might be present in the data cannot be captured by our annotations – and therefore, by the approaches outlined in this work.

## Ethics

This task explored the extent to which evidence for suicidal risk from online posts can be obtained by leveraging information inherent in open source LLMs, and how this information can be further summarised. However, the task cannot make any claims about the potential evidence providing explanations for suicidal risk and neither do the aggregate summaries constitute such explanations. The motivation behind the task was to explore the possibilities for evidence extraction provided by LLMs and the corresponding limitations. We hope that this is a first step to research that can actually make causal links between evidence and suicidality and augment models with symbolic of inference methods that can reveal reasoning processes.

The task also does not promote in any way the notion that LLMs could provide evidence for diagnosis that would not involve a human. Any such evidence would need to be reviewed by a human expert and our intuition is that better models could help augment the capacity of clinical experts by providing information that would not otherwise be available to them.

The UMD Reddit Suicidality Dataset was made available for the shared task following a determination by the University of Maryland College Park IRB that doing so was exempt from IRB review according to U.S. federal regulations. All of the data have been provided to the participants in an anonymised fashion. An application form was required to be signed by each of the teams before accessing the data, clarifying that only the listed members could have access to the dataset and the location where it would be hosted locally had to be stated. Even though we are using publicly available data from Reddit, we prohibited the use of any third-party LLMs that would require sending (part of) the data in the provider's servers, as to protect the suicide risk label of each user in the UMD Reddit Suicidality Dataset.

## Acknowledgements

This work was supported by a UKRI/EPSRC Turing AI Fellowship to Maria Liakata (grant ref EP/V030302/1) and the Alan Turing Institute (grant ref EP/N510129/1). Philip Resnik was supported by U.S. NSF award 2124270. The shared task organizers would like to express their gratitude to the anonymous users of Reddit whose data feature in this year's shared task dataset; to the clinical experts from Bar-Ilan University who annotated the data for both tasks; to the American Association of Suicidology; to all team members for their participation; and to EACL for its support for CLPsych.

## References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Falwah Alhamed, Julia Ive, and Lucia Specia. 2022. Predicting moments of mood changes overtime from imbalanced social media data. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 239–244, Seattle, USA. Association for Computational Linguistics.

Falwah Alhamed, Julia Ive, and Lucia Specia. 2024. Using large language models (llms) to extract evidence from pre-annotated social media data. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Mostafa M Amin, Erik Cambria, and Björn W Schuller. 2023. Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt. *arXiv preprint arXiv:2303.03186*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Sergi Blanco-Cuaresma. 2024. Psychological assessments with large language models: A privacy-focused and cost-effective approach. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

Jiyu Chen, Vincent Nguyen, Xiang Dai, Diego Molla-Aliod, Cecile Paris, and Sarvnaz Karimi. 2024. Exploring instructive prompts for large language models in the extraction of evidence for supporting assigned suicidal risk levels. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39.

Louise Deléger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnár, Laura Stoutenborough, Michal Kouril, Keith A. Marsolo, and Imre Solti. 2012. Building gold standard corpora for medical natural language processing tasks. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2012:144–53.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Muskan Garg. 2024. Wellxplain: Wellness concept extraction and classification in reddit posts for mental health analysis. *Knowledge-Based Systems*, 284:111228.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Anthony Hills, Adam Tsakalidis, and Maria Liakata. 2023. Time-aware predictions of moments of change in longitudinal user posts on social media. In *International Workshop on Advanced Analytics and Learning on Temporal Data*, pages 293–305. Springer.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Siqing Huo, Negar Arabzadeh, and Charles LA Clarke. 2023. Retrieving supporting evidence for llms generated answers. *arXiv preprint arXiv:2306.13781*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2.

Hyolim Jeon, Dongje Yoo, Daeun Lee, Sejung Son, Seungbae Kim, and Jinyoung Han. 2024. A dual-prompting for interpretable mental health language models. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

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, Marseille, France. European Language Resources Association.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

T. Joiner. 2007. *Why people die by suicide*. Harvard University Press, Cambridge, Mass.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

L Koushik, M Vishruth, and M Anand Kumar. 2024. Detecting suicide risk patterns using hierarchical attention networks with large language models. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Allison Lahnala, Yuntian Zhao, Charles Welch, Jonathan K. Kummerfeld, Lawrence C An, Kenneth Resnicow, Rada Mihalcea, and Verónica Pérez-Rosas. 2021. Exploring self-identified counseling expertise in online support forums. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4467–4480, Online. Association for Computational Linguistics.

Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Daeun Lee, Migyeong Kang, Minji Kim, and Jinyoung Han. 2022. Detecting suicidality with a contextual graph neural network. In *Proceedings of the eighth workshop on computational linguistics and clinical psychology*, pages 116–125.

Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. 2023. Compressing context to enhance inference efficiency of large language models. *arXiv preprint arXiv:2310.06201*.

Chandreen Liyanage, Muskan Garg, Vijay Mago, and Sunghwan Sohn. 2023. Augmenting Reddit posts to determine wellness dimensions impacting mental health. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 306–312, Toronto, Canada. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. 2023. Benchmarking large language model capabilities for conditional generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9194–9213, Toronto, Canada. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Matthew D Nemesure, Michael V Heinz, Raphael Huang, and Nicholas C Jacobson. 2021. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific reports*, 11(1):1980.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8446–8459, Dublin, Ireland. Association for Computational Linguistics.

Yaakov Ophir, Refael Tikochinski, Anat Brunstein Klomek, and Roi Reichart. 2022. The hitchhiker's guide to computational linguistics in suicide prevention. *Clinical Psychological Science*, 10(2):212–235.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2023. Overview of erisk 2023: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 294–315. Springer.

Kelly Posner, Gregory K Brown, Barbara Stanley, David A Brent, Kseniya V Yershova, Maria A Oquendo, Glenn W Currier, Glenn A Melvin, Laurence Greenhill, Sa Shen, et al. 2011. The columbia–suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American journal of psychiatry*, 168(12):1266–1277.

Judita Preiss and Zenan Chen. 2024. Incorporating word count information into depression risk summary generation: Inf@uos clpsych 2024 submission. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Megan L Rogers, Min Eun Jeon, Sifan Zheng, Jenelle A Richards, Thomas E Joiner, and Igor Galynker. 2023. Two sides of the same coin? empirical examination of two proposed characterizations of acute suicidal crises: Suicide crisis syndrome and acute suicidal affective disturbance. *Journal of psychiatric research*, 162:123–131.

Anastasia Sandu, Teodor Mihailescu, and Sergiu Nisioi. 2024. Cheap ways of extracting clinical markers from texts. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Ramit Sawhney, Shivam Agarwal, Atula Tejaswi Neerkaje, Nikolaos Aletras, Preslav Nakov, and Lucie Flek. 2022a. Towards suicide ideation detection through online conversational context. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1716–1727.

Ramit Sawhney, Atula Tejaswi Neerkaje, and Manas Gaur. 2022b. A risk-averse mechanism for suicidality assessment on social media. *Association for Computational Linguistics 2022 (ACL 2022)*.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Loitongbam Gyanendro Singh, Junyu Mao, Rudra Mutalik, and Stuart E Middleton. 2024. Extracting and summarizing evidence of suicidal ideation in social media contents using large language models. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Jiayu Song, Jenny Chim, Adam Tsakalidis, Julia Ive, Dana Atzil-Slonim, and Maria Liakata. 2024. Clinically meaningful timeline summarisation in social media for mental health monitoring.

Rika Tanaka and Yusuke Fukazawa. 2024. Suicide risk assessment and summarization using bert and mentallama. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. medrxiv.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vu Tran and Tomoko Matsui. 2024. Team ism at clpsych 2024: Extracting evidence of suicide risk from reddit posts with knowledge self-generation and output refinement using a large language model. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022a. 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, Seattle, USA. Association for Computational Linguistics.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022b. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660.

Talia Tseriotou, Adam Tsakalidis, Peter Foster, Terence Lyons, and Maria Liakata. 2023. Sequential path signature networks for personalised longitudinal language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5016–5031.

G Turecki, DA Brent, D Gunnell, RC O'Connor, MA Oquendo, J Pirkis, and BH Stanley. 2019. Suicide and suicide risk. nature reviews disease primers, 5 (1), 1-22.

Ahmet Yavuz Uluslu, Andrianos Michail, and Simon Clematide. 2024. Utilizing large language models to identify evidence of suicidality risk through analysis of emotionally charged posts. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Vasudha Varadarajan, Allison Lahnala, Adithya V Ganesan, Gourab Dey, Siddharth Mangalik, Ana-Maria Bucur, Nikita Soni, Rajath Rao, Kevin Lanning, Isabella Vallejo, Lucie Flek, H. Andrew Schwartz, Charles Welch, and Ryan Boyd. 2024. Archetypes and entropy: Theory-driven extraction of evidence for suicide risk. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Xuhai Xu, Bingshen 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. *arXiv preprint arXiv:2307.14385*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. Extractive summarization via ChatGPT for faithful summary generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*,

pages 3270–3278, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023b. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Jingwei Zhu, Ancheng Xu, Minghuan Tan, and Min Yang. 2024. Xinhai@clpsych 2024 shared task: Prompting healthcare-oriented llms for evidence highlighting in posts with suicide risk. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Ayah Zirikly and Mark Dredze. 2022. Explaining models of mental health via clinically grounded auxiliary tasks. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 30–39.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.