Archetypes and Entropy: Theory-Driven Extraction of Evidence for Suicide Risk

Vasudha Varadarajan[1], **Allison Lahnala**[2], **Adithya V Ganesan**[1], **Gourab Dey**[1]
Siddharth Mangalik[1], **Ana-Maria Bucur**[3,4], **Nikita Soni**[1], **Rajath Rao**[1], **Kevin Lanning**[5]
Isabella Vallejo[5], **Lucie Flek**[2], **H. Andrew Schwartz**[1], **Charles Welch**[2], **Ryan L Boyd**[1]
[1]Department of Computer Science, Stony Brook University
[2]Bonn-Aachen International Center for Information Technology (b-it), University of Bonn
[3]Interdisciplinary School of Doctoral Studies, University of Bucharest
[4]PRHLT Research Center, Universitat Politècnica de València
[5]Wilkes Honors College, Florida Atlantic University
{vvaradarajan, boyd}@cs.stonybrook.edu, alahnala@uni-bonn.de

Abstract

Sensitive content warning: *This paper contains sensitive content related to suicide.*

Psychological risk factors for suicide have been extensively studied for decades. However, combining explainable theory with modern data-driven language modeling approaches is non-trivial. Here, we propose and evaluate methods for identifying language patterns indicative of suicide risk by combining theory-driven suicidal *archetypes* with language model-based and *relative entropy*-based approaches. *Archetypes* are based on prototypical statements that evince risk of suicidality while *relative entropy* considers the difference between how probable the risk-familiar and risk-unfamiliar models find user language. Each approach performed well individually; combining the two strikingly improved performance, yielding our combined system submission with a BERTScore Recall of 0.906. Further, we find diagnostic language is distributed unevenly in posts, with titles containing substantial risk evidence. We conclude that a union between theory- and data-driven methods is beneficial, outperforming more modern prompt-based methods.

1 Introduction

With the advent of large language models (LLMs) (Brown et al., 2020), studies exploring their potential for estimating suicide risk from social media data have proliferated (Coppersmith et al., 2018; Matero et al., 2019; Nock et al., 2019; Coppersmith, 2022). Such studies, however, chiefly emphasize predictive accuracy over explainability and interpretability (Schafer et al., 2021), limiting both their clinical applicability and their utility in testing theories of suicide. Our team, SWELL, takes a psychological theory-informed approach to produce evidential explanations and summaries for the assigned suicide risk score of Reddit users.

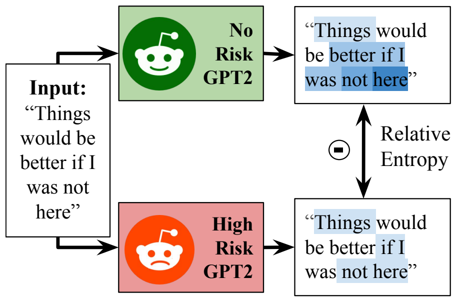

Figure 1: ***Relative entropy* method.** Two DistilGPT2 models were independently pretrained domain-adaptively (Gururangan et al., 2020) on posts from users having no suicide risk (*No Risk GPT2*) and users having high suicide risk (*High Risk GPT2*). The difference in the token entropy between the models is used as a measure of "surprisal" to extract the evidential highlights of at-risk suicide users. Highlighted spans indicate entropy values, with darker colors indicating higher entropy.

Despite substantial effort dedicated to extracting explanatory rationale for LLM answers for math, physics, and even theory of mind (Cobbe et al., 2021; Zheng et al., 2023; Saha et al., 2023), there has been limited work in building similar explanatory pipelines for mental health diagnostics. The CLPsych-2024 shared task asked teams to provide evidences and summaries for suicide risk from social media posts (Chim et al., 2024).

Our main contributions include three novel methods for suicide risk evidence extraction based on (1) theory-based *archetype* representations of suicidality including with Llama2-Chat (Touvron et al., 2023), (2) an LLM-based *relative entropy* method, and (3) a hybrid combination of *entropy* with *archetypes*. Additionally, we provide (4) a method for prompt-based explanation summaries, and (5) associations of theory-based *archetypes* with trained expert annotations. Further, we release the code associated with our submissions.[1]

[1]https://github.com/humanlab/clp24-arch-entropy

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 278–291
March 21, 2024

2 Background

Conceptualizing Suicide Risk. One of the most prominent theoretical conceptualizations of suicide is Joiners' Interpersonal Theory of Suicide (IPTS) (Van Orden et al., 2010) which is comprised of 3 factors that jointly characterize suicide risk: 1) *Acquired Capability*, a person's increased tolerance for physical pain and fear of death, which can develop over time through suicidal ideation and repeated exposure to painful or fear-inducing experiences (Smith et al., 2010); 2) *Perceived Burdensomeness*, an individual's belief that their existence or presence is a burden on others (Joiner et al., 2002); 3) *Thwarted Belongingness*, the perception or experience of not belonging to, or feeling disconnected from, meaningful social relationships despite efforts to form connections (Silva et al., 2015). Prior work suggests that suicide becomes possible when an individual experiences high levels of all 3 constructs (Joiner, 2007).

Explainable Approaches for Suicide Risk Prediction. The evolution of language modeling techniques has led to improvements in risk prediction tasks (Sawhney et al., 2022; Xu et al., 2023; V Ganesan et al., 2021; Juhng et al., 2023; V Ganesan et al., 2022), yet very little has been focused on adapting these models to be more reliable or practical for real-world applications. Heckler et al. (2022) identified *interpretability and explainability* as one of the primary challenges in supporting specialists with understanding model inferences. A number of NLP tasks such as natural language inference (Camburu et al., 2018), hate speech detection (Mathew et al., 2021), discourse relation prediction (Son et al., 2022) and commonsense reasoning (Aggarwal et al., 2021) have made long strides in building explainable models. In the vein of improving the explainability of LLMs and addressing the particular need for suicide-risk assessment models, this year's CLPsych shared task investigates evidence extraction and summarization for suicide risk from social media posts, evaluating against highlights and summaries written by experts.

3 Data & Tasks

Dataset. The CLPsych-2024 shared task uses the UMD Suicidality v2 dataset (Shing et al., 2018; Zirikly et al., 2019), which contains history of posts from all subreddits for a set of users who posted on `r/SuicideWatch` (SW), a support forum for people struggling with suicidal thoughts. For each SW poster, the dataset includes a suicide risk label and a list of posts from the user. Expert annotators further assessed the posts to highlight text spans that provide evidence supporting the risk label, and wrote cohesive summaries of all evidence selected for each user.

	Risk level	AC: I/S	AC: EoE	AC: DtH	AC: HTPP	AC: ERB	AC: FSHM	PB	TB
Risk level	1.0‡	0.24	0.13	0.12	0.16	0.14	0.13	0.19	0.11
AC: I/S	0.24‡	1.0	0.35	0.40	0.28	0.33	0.53	0.40	0.35
AC: EoE	0.13	0.35	1.0	0.70	0.71	0.57	0.58	0.40	0.37
AC: DtH	0.12	0.40	0.70	1.0	0.76	0.68	0.58	0.37	0.34
AC: HTPP	0.16†	0.28	0.71	0.76	1.0	0.66	0.59	0.35	0.37
AC: ERB	0.14†	0.33	0.57	0.68	0.66	1.0	0.57	0.39	0.39
AC: FSHM	0.13	0.53	0.58	0.58	0.59	0.57	1.0	0.35	0.34
PB	0.19‡	0.40	0.40	0.37	0.35	0.39	0.35	1.0	0.64
TB	0.11	0.35	0.37	0.34	0.37	0.39	0.34	0.64	1.0

Figure 2: A descriptive Spearman correlation matrix between expert-labeled risk level from UMD Suicidality Dataset and maximum user-level *archetype* scores. Archetypes include Perceived Burdensomness (PB), Thwarted Belongingness (TB), and Acquired Capability (AC) with subtypes Ideation/Simulation (I/S), Experiences of Endurance (EoE), Desensitization to Harm (DtH), High Tolerance for Physical Pain (HTPP), Engagement in Risky Behaviors (ERB), and Familiarity with Self-Harm Methods (FSHM). Statistically significant correlations between the archetypes and risk levels are marked in the first column ($p < 0.05^{\dagger}$ and $p < 0.005^{\ddagger}$). Archetypes correlated with each other in theory-consistent ways and, additionally, were meaningfully related to expert evaluations.

The SW posts came from 195 users whom experts labeled as having no risk, and 671 users whom experts labeled into 3 suicide risk categories (*any risk*): low, moderate, and high risk. There were 1,241 posts on SW and 59,933 posts on non-SuicideWatch (NSW) subreddits. 125 users from the expert-annotated set (162 SW posts; 19,894 NSW posts) were held out as the test set. From the 195 control (*no risk*) users, the dataset also included 19,631 NSW and 195 SW posts.

Shared Tasks. The first task was to automatically extract highlights from the SW posts that explain the assigned suicide risk level of the user. The second task was to generate cohesive summaries that aggregate the evidence supporting the users' assigned suicide risk levels. These summaries were evaluated by their consistency with human expert summaries based on the same users' SW posts.

4 Methods

4.1 Evidence Extraction

We designed three general approaches described below, and experimented with variants and compositions of each. Except for the prompt-based approach, we extract highlights at the sentence level, preprocessed with the NLTK sentence tokenizer (Bird and Loper, 2004).

4.1.1 Suicidality *Archetypes*.

Several extant theories of suicide have been proposed that explain various psychological states and trajectories of suicide. For the purposes of this paper, we focused specifically on constructs from Joiner's IPTS (Joiner, 2007). Psychologists on our team formalized prototypical statements reflecting patterns of cognition relevant to suicide risk (e.g., "I've intentionally exposed myself to pain to build up my resistance"). Prototype sentences were embedded using `RoBERTa-large` (Liu et al., 2019); all embeddings were then averaged, separately by factor, to create a representative *archetype* of each construct (see Table A3). As an example, for the Ideation/Simulation archetype, the average embedding of the three statements forms the representation of the archetype.

Inspired by Garten et al. (2018), we calculate the cosine similarity between the sentence embeddings of a post and each archetype. We compute Spearman's correlation between the expert risk assessments and archetypes of Joiner's IPTS, with the users' maximum scores for each archetype to reflect the maximum risk evidence. The correlation matrix is shown in Figure 2. We find statistically significant correlations between expert-labeled risk levels and Perceived Burdensomeness, AC: Engagement in Risky Behaviors, AC: High Tolerance for Physical Pain, and AC: Ideation/Simulation, the latter having the strongest, most significant relationship with $r = 0.238$ and $p < 0.001$.

For our *archetypes*-only method ("Archetypes" in Table 1), Principal Component Analysis (PCA) (Tipping and Bishop, 1999) was applied to all 8 archetype similarity scores, reducing them from 8 to 2 dimensions. After z-scoring the sum of component scores, we highlighted spans that were either in the top-ranking 25% of each post or ≥ 1.5 standard deviations from the mean sum of components.

4.1.2 Relative Entropy.

This method is based on Lahnala et al. (2021)'s approach for studying the language of mental health professionals and peer supporters in online support forums. The entropy (used to calculate perplexity) for a token in an LM is a signal of "surprisal" of that token given the context and domain (Jurafsky and Martin, 2023, Ch. 3). Figure 1 depicts this method adapted for this work, in which, the token "here" would be particularly unexpected in this context from a no-risk user.

Domain adaptation (Gururangan et al., 2020) of LLMs on low-risk or no-risk data leads to higher entropy for tokens signaling high-risk in the high-risk data. However, as out-of-domain expressions can also have high entropy, we calibrate the entropy by domain-adaptive pre-training of two LLMs; one with lower-risk data and one with higher-risk data. We hypothesize that higher differences from subtracting token entropies of higher-risk models from lower-risk models are signals of risk-associated language.

To calculate the relative entropy, we subtract the entropy of the token-level predictions of one model from the other. For a model, H, trained on high-risk data, we can subtract the entropy this model assigns to high-risk data from the entropy assigned by a model, L, trained on low-risk data. To obtain the entropy difference for a sequence of tokens, S, in a given sentence, we calculate the maximum[2] of token entropy differences within the sentence:

$$E_{L,H} = \max_{s \in S}\{log(p_L(s)) - log(p_H(s))\} \quad (1)$$

We applied domain-adaptive pretraining to DistilGPT2 (Sanh et al., 2019) for each of the risk categories: none (a), low (b), moderate (c), high (d) and any (b,c,d), and calculate the entropy differences between sentences for each language model pair. In our system, four pairs of models were considered: no-low, no-moderate, no-high and no-any. We applied PCA to reduce the dimensionality of these four elements to a single relative entropy score and qualitatively examined the scores to determine a threshold for selecting sentences as spans, resulting in the top 30% of sentences.

[2]We conducted a qualitative analysis and found that the maximum performed better than the mean or median.

	Highlighted Evidence				Summarized Evidence	
	Recall ↑	Precision ↑	W.Recall ↑	H.Mean ↑	Mean Consist. ↑	Max Contra. ↓
Random (25%)	0.887	0.894	0.790	0.891	0.969	0.094
Archetypes (25%)	0.897	**0.914**	**0.816**	0.905	**0.973**	**0.080**
ArchPrompts-Llama2c	0.884	**0.914**	0.741	**0.910**	0.972	0.082
Entropy-DistilGPT2 (30%)	0.901	0.884	0.621	0.892	0.967	0.094
Entropy-DGPT2 x Archs (30%)	**0.906**	0.897	0.648	0.901	0.970	0.092

Table 1: **Scores based on shared task's annotations.** The first row indicates a baseline which is a random selection of 25% of sentences from each post. Our submissions to the shared task were Archetypes, ArchPrompts-Llama2c (LLama2-chat 13b prompted to extract sentences evidential of the 3 major archetypes), and Entropy-DistilGPT2 x Archs (combining our Entropy based approach with Archetype scores). The scores on the right compare the gold evidence summaries with the evidence summaries generated by Llama2c with highlighted evidence spans from each method as the inputs.

We also applied similar techniques to another LM, HaRT (Human-aware Recurrent Transformer) (Soni et al., 2022) which is a user-level LM that models message-level context along with author-specific context, helping capture the surprisal of language specific to the author. The dataset for domain-adaptive pretraining included a limited number of historical posts from other subreddits for each user in the SW test set. We encode users' NSW and SW posts in a chronological order by concatenating them with a separator token. Two models were trained for none and any risk levels, and we followed the same entropy calculations. In §5, we discuss a comparison of this user-level variant of the entropy method against a combination of archetypes and entropy (see Table 3).

4.1.3 Prompt-based evidence highlights.

Our submission based on Llama2-Chat used Joiner's constructs in a few-shot setting to extract highlights from the posts. We created instructions that included a definition of each construct alongside five prototypical examples of highlights extracted from the posts for the respective construct. We then prompted a self-hosted instance of Llama2-Chat (13B) with these instructions to generate a list of highlights that correspond with each construct for each post. The full prompts are in Appendix B.

4.2 Evidence Summarization

For each system detailed in §4.1, we prompt Llama2-chat (13B) with detailed instructions to summarize the highlighted evidence of the user explaining the assigned risk level. The instruction was framed to incorporate different factors of language (Emotional State, Cognitive Processes, Behavior and Motivation, Interpersonal Relationships and Social Support, Mental Health Issues, and Other Risk Factors) while summarizing the highlights with the objective of explaining the risk category. For the prompt and more details about the method, please see Appendix B. Llama2-Chat was provided up to 10 highlights in order to avoid running into problems caused by long context (Liu et al., 2023) and the highlights were uniformly sampled from all posts for each user.

	Rec	Prec	*M/p*
Archetypes	**0.892**	**0.899**	**3.75**
ArchPrompt-Llama2c	0.789	0.797	4.39
Entropy-DGPT2	0.867	0.861	5.41
Entropy-DGPT2 x Archs	0.881	0.865	6.40

Table 2: Recall, Precision on the set of internal expert annotations and mean spans extracted per post (*M/p*). The *M/p* for the internal expert annotation was 6.35.

4.3 Internal Annotations

To support our experimental evaluations, we collected our own set of annotations of evidence from experts, based on Joiner's IPTS (§2). We selected 50 posts from 50 unique users that were not part of the heldout test set for the shared task. These were annotated by two clinical experts following the guidelines outlined in the Appendix (Table A6). We used the annotations to internally validate our systems and select the best models (see Table 2).

5 Results

We discuss the result of our experiments with the methods described in §4. We report the results of our official submissions in Appendix A.[3]

[3]We intended our first official submission to be Archetypes-based and the second to be based on Entropy x Archetypes. Instead, due to a couple of interesting bugs, we re-did experiments to validate our findings and report them in §5 and Table 1. The initial submissions are described in Appendix A.

Archetypes capture relevant highlights. Table 1 shows that theory-driven approaches such as Archetypes (1) outperform random chance; and (2) interestingly, we observe that small yet strong encoder language models (RoBERTa-large) generalize archetypal utterances of suicidal risk with few examples better than large generative models (Llama2c). Further, performance on internal expert annotations in Table 2 validates the generalizability of Archetypes from internal to shared task annotator pools. Archetypes also have an average of 3.75 highlights per post, which when coupled with the overall performance, indicates highly informative spans are selected as evidence.

Entropy combined with Archetypes further improves Recall. We find that entropy-based methods have a high recall owing to better coverage of highlights signaling suicide risk, however, this comes at a small cost of precision, as seen in Table 1. Since *Archetypes* reflect theory-driven signals and *Entropy* captures data-driven signals of suicide risk, we combine Archetypes with Entropy by multiplying the scores and selecting the top-scoring 30% of sentences. This produced the best recall and improved the precision of the entropy-based method by a significant margin in the case of shared task annotations. In the case of internal expert annotations, Archetypes fared better, likely due in part to our internal annotation schema being consistent with Joiner's IPTS theory. For summaries and extracted spans for each system for a paraphrased example, see Appendix A2.

	Recall	Precision	W.Recall
Post-structure experiments			
Random 25%	0.887	**0.894**	0.790
Title only	0.862	**0.894**	**0.840**
25% body	**0.884**	0.892	0.699
Title + 25% body	0.883	0.893	0.788
***Entropy-variant experiments* (top 40%)**			
DGPT2 x Archs	**0.915**	**0.892**	**0.542**
HaRT	0.912	0.887	0.525

Table 3: Recall, precision, and weighted recall for the post-structure experiments and entropy-variant experiments for Task A.

We further compare the performance boost afforded by using Entropy-DGPT2 x Archetypes against better modeling of user-level context using Entropy-HaRT, selecting top 40% of sentences from both the methods as the suicide risk evidence. Table 3 (bottom) shows a comparison of the two entropy-variant experiments, and we find that the combination of Entropy-DGPT2 x Archetypes is better across all three performance metrics.

Title of a post is highly informative. In Table 3 (top), we explore the post structure of SuicideWatch posts to understand the effects of the title and body. We experiment with three conditions; using only the title, the first 25% of the body and the title and 25% body together. Our results from using the title and the first 25% of the posts show that they outperform a random sampling of 25% of posts. Interestingly, when using only the title, we get the highest weighted recall across all methods, supporting that titles are highly informative (Matero et al., 2019) and potentially pointing to signals of suicide being presented upfront in SW posts on platforms with a similar post structure.

Llama2-Chat is consistent with Summarization. Summaries generated by Llama2-chat (13B) had high consistency and low contradiction scores across all submissions. This may have resulted from (1) the model's ability to precisely identify the suicide risk from appropriate psychological dimensions inferred from the span(s), and (2) a prompt carefully crafted to consider the important psychological dimensions to provide the summary.

While these summarized explanations are more convincing for high-risk users, we also find that the model is extremely sensitive to the inputs. For example, posts with very few spans from low-risk users were still surmised to exhibit a "heightened risk of suicide."

6 Conclusion

We combined theory-driven archetypes with data-driven language models to extract evidence from users' social media posts that support the assigned suicide risk levels. We found that scores derived for Joiner's constructs show a significant correlation with assigned suicide risk. Combining the relative entropy scores with Joiner construct scores improves upon relative entropy alone, which is demonstrated in the experimental results on both the shared task test set as well as our set of internal annotations. These rigorous data-driven methods grounded in theory also outperformed extensive prompting of instruction-following LLMs. Still, archetypes alone yield the highest precision in both evaluations. This demonstrates the importance of theoretically derived constructs in language modeling approaches to build explainable approaches for mental health diagnosis.

Limitations

While often not characterized within the context of IPTS, research has identified numerous other, more specific factors and pathways to suicidality, such as an omnibus need for "escape" from aversive self-awareness (Baumeister, 1990) and substance dependence (Pompili et al., 2012). A comprehensive review of suicide risk factors is beyond the scope of this paper, however, several such risk factors played a role in our approach to understanding and capturing suicide risk. We limited our methods to the most prominent factors as described in §2.

For the scope of our work, we limited our study of prompt-based methods for both evidential highlighting as well as evidence summarization to a single modern large language model – Llama2-chat. While modern LLMs lack social (Choi et al., 2023; Ziems et al., 2023; Varadarajan et al., 2023; Lahnala et al., 2022) and personal understanding (Havaldar et al., 2023; V Ganesan et al., 2023) from language, experiments using the same prompting structure with other socially and human aware LMs (Dey et al., 2024) could potentially produce results that outperform the methods described in this paper.

The studied data is limited to the English-speaking Reddit and may contain other data-specific biases (Chancellor et al., 2019) such as sampling bias towards certain groups. Furthermore, the subjectivity of interpretation of suicidality across individuals (Keilp et al., 2012) and the possibility for annotator biases (Hovy and Spruit, 2016) could implicate limitations in model training and evaluation approaches.

Ethics Statement

While the essence of our work is to aid in the detection of at-risk users, it is imperative that any interventions be well-thought, failing which may lead to counter-productive outcomes, such as users moving to fringe platforms, which would make it harder to provide assistance (Kumar et al., 2015). Care should be taken so as not to create stigma, and interventions must be carefully planned by consulting relevant stakeholders such as clinicians, designers, and researchers (Chancellor et al., 2016), to maintain social media as a safe space for individuals looking to express themselves (Chancellor et al., 2019).

We do not seek to make any diagnostic claims with our work; rather, we aim to help prioritize individuals in need of immediate help. Our approach should hence not act as a standalone method in risk assessment (De Choudhury et al., 2016). It is critical to avoid misuse of algorithmic inferences by bad actors (Chancellor et al., 2019), as in the case of Samaritan's radar (Hsin et al., 2016), by only selectively sharing the evaluations made by our study (De Choudhury et al., 2016). It is also vital to incorporate accessible interpretations (Chancellor et al., 2019). While we highlight the role of NLP as part of forming a human-in-the-loop framework, it is further essential that clinicians are not overburdened (Chancellor et al., 2019).

Issues with summarization methods also suggest that today's open-source LLMs are still not at the stage to run post-hoc explanations for suicide risk associated with the text. These models need to be fine-tuned and could be guardrailed using RLHF (Ouyang et al., 2022).

Acknowledgements

We would like to thank Veerle Eijsbroek and Katarina Kjell for their efforts in creating high-quality and IPTS-based annotations of the SuicideWatch posts. These expert annotations were invaluable for allowing us to conduct internal evaluations. We also thank the American Association of Suicidology for their assistance in making the dataset available for conducting this study.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Roy F. Baumeister. 1990. Suicide as escape from self. *Psychological Review*, 97(1):90–113. Place: US Publisher: American Psychological Association.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. 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, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 79–88, Atlanta, GA, USA.

Stevie Chancellor, Zhiyuan Lin, Erica L. Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW '16, pages 1171–1184, San Francisco, CA, USA.

Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the clpsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. "Association for Computational Linguistics".

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Glen Coppersmith. 2022. Digital life data in the clinical whitespace. *Current Directions in Psychological Science*, 31(1):34–40.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2098–2110, San Jose, CA, USA.

Gourab Dey, Adithya V Ganesan, Yash Kumar Lal, Manal Shah, Shreyashee Sinha, Matthew Matero, Salvatore Giorgi, Vivek Kulkarni, and H. Andrew Schwartz. 2024. SOCIALITE-LLAMA: An instruction-tuned model for social scientific tasks. In *18th Conference of the European Chapter of the Association for Computational Linguistics*, St. Julian's, Malta. Association for Computational Linguistics.

Justin Garten, Joe Hoover, Kate M. Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior Research Methods*, 50(1):344–361.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Wesllei Felipe Heckler, Juliano Varella de Carvalho, and Jorge Luis Victória Barbosa. 2022. Machine learning for suicidal ideation identification: A systematic literature review. *Computers in Human Behavior*, 128:107095.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, pages 591–598, Berlin, Germany.

Honor Hsin, John Torous, and Laura Roberts. 2016. An adjuvant role for mobile health in psychiatry. *JAMA Psychiatry*, 73(2).

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

Thomas E. Joiner, Jeremy W. Pettit, Rheeda L. Walker, Zachary R. Voelz, Jacqueline Cruz, M. David Rudd, and David Lester. 2002. Perceived burdensomeness and suicidality: Two studies on the suicide notes of those attempting and those completing suicide. *Journal of Social and Clinical Psychology*, 21(5):531–545. Publisher: Guilford Publications Inc.

Swanie Juhng, Matthew Matero, Vasudha Varadarajan, Johannes Eichstaedt, Adithya V Ganesan, and H. Andrew Schwartz. 2023. Discourse-level representations can improve prediction of degree of anxiety. In

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1500–1511, Toronto, Canada. Association for Computational Linguistics.

Dan Jurafsky and James H Martin. 2023. *Speech and Language Processing (3rd ed. draft)*.

John G Keilp, Michael F Grunebaum, Marianne Gorlyn, Simone LeBlanc, Ainsley K Burke, Hanga Galfalvy, Maria A Oquendo, and J John Mann. 2012. Suicidal ideation and the subjective aspects of depression. *Journal of Affective Disorders*, 140(1).

Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, pages 85–94, Guzelyurt, TRNC, Cypru.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158, Abu Dhabi, United Arab Emirates. 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.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Matthew K Nock, Franchesca Ramirez, and Osiris Rankin. 2019. Advancing our understanding of the who, when, and why of suicide risk. *JAMA psychiatry*, 76(1):11–12.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Maurizio Pompili, Gianluca Serafini, Marco Innamorati, Massimo Biondi, Alberto Siracusano, Massimo Di Giannantonio, Giancarlo Giupponi, Mario Amore, David Lester, Paolo Girardi, and Anne Maria Möller-Leimkühler. 2012. Substance abuse and suicide risk among adolescents. *European Archives of Psychiatry and Clinical Neuroscience*, 262(6):469–485.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. 2023. Can language models teach weaker agents? teacher explanations improve students via theory of mind. *arXiv preprint arXiv:2306.09299*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMCˆ2 Workshop*.

Ramit Sawhney, Shivam Agarwal, Atula Tejaswi Neerkaje, Nikolaos Aletras, Preslav Nakov, and Lucie Flek. 2022. 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.

Katherine M. Schafer, Grace Kennedy, Austin Gallyer, and Philip Resnik. 2021. A direct comparison of theory-driven and machine learning prediction of suicide: A meta-analysis. *PLOS ONE*, 16(4):e0249833. Publisher: Public Library of Science.

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.

Caroline Silva, Jessica D. Ribeiro, and Thomas E. Joiner. 2015. Mental disorders and thwarted belongingness, perceived burdensomeness, and acquired capability for suicide. *Psychiatry Research*, 226(1):316–327.

Phillip N. Smith, Kelly C. Cukrowicz, Erin K. Poindexter, Valerie Hobson, and Lee M. Cohen. 2010. The acquired capability for suicide: A comparison of suicide attempters, suicide ideators, and non-suicidal controls. *Depression and Anxiety*, 27(9):871–877.

Youngseo Son, Vasudha Varadarajan, and H. Andrew Schwartz. 2022. Discourse relation embeddings:

Representing the relations between discourse segments in social media. In *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*, pages 45–55, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Andrew Schwartz. 2022. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 622–636, Dublin, Ireland. Association for Computational Linguistics.

Michael E Tipping and Christopher M Bishop. 1999. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482.

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*.

Adithya V Ganesan, Yash Kumar Lal, August Nilsson, and H. Schwartz. 2023. Systematic evaluation of GPT-3 for zero-shot personality estimation. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 390–400, Toronto, Canada. Association for Computational Linguistics.

Adithya V Ganesan, Matthew Matero, Aravind Reddy Ravula, Huy Vu, and H. Andrew Schwartz. 2021. Empirical evaluation of pre-trained transformers for human-level NLP: The role of sample size and dimensionality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4515–4532, Online. Association for Computational Linguistics.

Adithya V Ganesan, Vasudha Varadarajan, Juhi Mittal, Shashanka Subrahmanya, Matthew Matero, Nikita Soni, Sharath Chandra Guntuku, Johannes Eichstaedt, and H. Andrew Schwartz. 2022. WWBP-SQT-lite: Multi-level models and difference embeddings for moments of change identification in mental health forums. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 251–258, Seattle, USA. Association for Computational Linguistics.

Kimberly A. Van Orden, Tracy K. Witte, Kelly C. Cukrowicz, Scott Braithwaite, Edward A. Selby, and Thomas E. Joiner. 2010. The interpersonal theory of suicide. *Psychological review*, 117(2):575–600.

Vasudha Varadarajan, Swanie Juhng, Syeda Mahwish, Xiaoran Liu, Jonah Luby, Christian Luhmann, and H. Andrew Schwartz. 2023. Transfer and active learning for dissonance detection: Addressing the rare-class challenge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11923–11936, Toronto, Canada. Association for Computational Linguistics.

Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2023. Mental-llm: Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385*.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. 2023. Large language models for scientific synthesis, inference and explanation. *arXiv preprint arXiv:2310.07984*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, pages 1–53.

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*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

Appendices

A Official Submissions to CLPsych 2024 Shared Task.

The results of our official submissions to both Shared Task A and B are shown in Table A1.

A.1 Task A: Evidence Extraction.

A.1.1 SWELL-1: First 25% Title + Body

For our first submission, we picked the first 25% of sentences from the concatenated title and body for each post as evidence of the assigned suicide risk level. As discussed in §5, the title of a post is highly informative, and with its inclusion, this system scored the highest weighted recall (0.808) among all the official submissions for Task A.

A.1.2 SWELL-2: Top 40% Archetypes + Entropy

In this approach, each sentence from the posts was scored by taking the product of the maximum archetype score and the relative entropy score (as described in §4.1). Spans were selected by using the top-scoring 40% of sentences. For the official submission, the training data was comprised of mostly crowd-annotated posts, while the validation set was comprised entirely of expert-annotated

	Highlighted Evidence				Summarized Evidence	
	Recall ↑	Precision ↑	W.Recall ↑	H.Mean ↑	Mean Consist. ↑	Max Contra. ↓
1. First 25%	0.881	0.895	**0.808**	0.888	0.972	0.080
2. 40% Entropy-DGPT2 x Archs	**0.915**	0.892	0.542	0.903	0.973	0.081
3. ArchPrompts-Llama2c	0.884	**0.914**	0.741	**0.910**	0.972	0.082

Table A1: **Official Submissions**: Recall, precision and weighted recall for our official submissions to CLPsych Shared Task 2024 Task A. The recall of entropy-based systems is much higher than Table 1 due to picking top 40% of the sentences rather than top 30%, which is reflected in the decreased weighted recall.

posts. This likely contributed to our model overfitting on the crowd-annotated domain. We mixed the distribution of crowd and expert-annotated data across the train and validation sets in our experiments after the official submission, which are the results we present in the main paper (with top 30%). Since the Shared Task emphasizes the importance of recall, our official submission included the top 40% highest-scoring sentences. While this method succeeds in improving recall, the weighted recall is much lower than SWELL-1's. This shows that the evidence selected by SWELL-2 captures most expert-annotated evidence of suicide risk but also many potentially irrelevant sentences. We find from other experiments in §5 that when we select top 30% instead of 40%, the recall is reduced too, indicating that a 40% cutoff yields some of the most evidential highlights, albeit with a higher false positive rate.

A.1.3 SWELL-3: ArchPrompts-LLaMA2c

The third submission included the highlights extracted using prompting, as explained in §4.1.3 and is same as the one reported in Table 1. The prompts were based on Joiner's IPTS theory as well (§4.1.3), instructing the model to find highlights relevant to each archetype. The exact instruction can be found in Table A4. This theory-based method yields precise spans that evince suicide risk much better than SWELL-1 and SWELL-2, while not trading off performance on weighted recall.

A.2 Task B: Evidence Summarization

The highlights from the Task A for SWELL-1, SWELL-2 and SWELL-3 were further prompted to be summarized, as explained in §4.2. The exact instruction can be found in Table A5. All the three methods yielded similar results, generating summaries highly consistent with the expert summaries.

B Prompts for Evidence Extraction and Summarizing

The instruction prompts we developed are shown in Table A4 for the evidence extraction task (Task A) and Table A5 for the summarization task (Task B).

C Paraphrased examples

The following are sample summaries generated based on the example post and highlighted evidence selected by each system shown in Table A2.

C.1 SWELL-1 (First 25% Title + Body)

Generated summary: “There is a high risk of suicide for this individual. The first excerpt reveals that they spent two hours trying to "man up" and hang themselves, and even wrote a note to their family. This suggests a high level of emotional distress and a possible loss of hope. The second excerpt highlights the individual’s sobering reality of nearly killing themselves today, which further substantiates their emotional state. Additionally, the use of the phrase "man up" implies a sense of shame or stigma associated with their emotions, which may hinder their ability to seek help. The individual’s cognitive processes may be impaired due to emotional distress, as evidenced by the length of time they spend attempting to take their own life. Their behavior and motivation have also changed, as they have lost interest in activities they once enjoyed. The lack of social support and mental health issues, such as depression or anxiety, may also contribute to their suicide risk. Overall, the text excerpts suggest a high risk of suicide for this individual.”

C.2 SWELL-2 (Archetypes + Entropy)

Generated summary: “There is a high risk of suicide for this individual. They have expressed a desire to end their life, with a history of suicidal thoughts and a lack of motivation to engage in activities they once enjoyed. Their emotional state is characterized by sadness, hopelessness, and a sense of emptiness. They have also reported poor grades, social isolation, and a lack of close relationships, which can further exacerbate their risk of suicide. Additionally, their cognitive processes are impaired, as evidenced by their negative thought patterns and decision-making abilities. Overall, this individual is at a high risk of suicide and should be given immediate attention and support.”

C.3 SWELL-3 (ArchPrompts-LLaMA2c)

Generated summary: “There is a high risk of suicide for this individual. They express feelings of emotional distress, hopelessness, and a lack of motivation to engage in activities they once enjoyed. They have a negative self-image, feeling like they have "piss poor grades" and are not amounting to anything. They have also considered suicide, spending two hours trying to "man up" and hang themselves. Additionally, they have no real friends and have never been close to a girl, which can further exacerbate their feelings of isolation and low self-esteem. These factors, combined with their mental health issues and lack of social support, contribute to a high risk of suicide.”

SWELL1: First 25% Title + Body
I almost killed myself today. I wrote a note for my family and tried to man up and hang myself. I couldn't do it. For the past year my life has felt strange and just hasn't been worth living. Some days I'm great, and others I'm thinking about how to kill myself. My life just feels...empty. My grades suck, don't have any friends, never had a girlfriend. My life is meaningless, so what's even the point?
SWELL2: Entropy-DGPT2 x Archetype (Top 40%)
I almost killed myself today.`score=1.18` I wrote a note for my family and tried to man up and hang myself.`score=0.76` I couldn't do it.`score=0.49` For the past year my life has felt strange and just hasn't been worth living.`score=0.69` Some days I'm great, and others I'm thinking about how to kill myself.`score=1.28` My life just feels...empty.`score=1.50` My grades suck, don't have any friends, never had a girlfriend.`score=1.56` My life is meaningless, so what's even the point?`score=1.42`
SWELL3: Prompt-LLaMA2c
I almost killed myself today. I wrote a note for my family and tried to man up and hang myself. I couldn't do it. For the past year my life has felt strange and just hasn't been worth living. Some days I'm great, and others I'm thinking about how to kill myself. My life just feels...empty. My grades suck, don't have any friends, never had a girlfriend. My life is meaningless, so what's even the point?

Table A2: A paraphrased example showing the highlighted evidence from each system.

Construct	Prototype Sentence	Relevant Work
Acquired Capability: Ideation/Simulation	I think about putting a rope around my neck. I want to put a gun in my mouth and pull the trigger. I plan on taking a bunch of pills and just fall asleep forever.	Smith et al. (2010)
Acquired Capability: Experiences of Endurance	I've been through so much pain in my life that I feel like nothing can hurt me anymore. Overcoming those challenges made me realize I can endure a lot more than I thought. I've become numb to the pain and it takes a lot to bother me now.	
Acquired Capability: Desensitization to Harm	I've seen and experienced so much violence that it doesn't faze me anymore. I can handle situations that used to terrify me, it's like I'm immune to the fear. I've become desensitized to the pain.	
Acquired Capability: High Tolerance for Physical Pain	I can tolerate physical pain much better now; it's almost like I've trained myself. I've intentionally exposed myself to pain to build up my resistance. I've been through so many injuries that pain doesn't bother me as much.	
Acquired Capability: Engagement in Risky Behaviors	I've taken risks that others wouldn't even consider; it's like I'm not afraid of consequences anymore. Engaging in extreme activities has made me less afraid of potential harm. I've deliberately put myself in dangerous situations, and it doesn't scare me like it used to.	
Acquired Capability: Familiarity with Self-Harm Methods	I've researched and learned about different ways to harm myself. Knowing about self-harm methods doesn't bother me. I'm not disturbed by the idea of self-harm anymore.	
Perceived Burdensomeness	Everyone would be better off without me. The world would be a better place without me. I add nothing to the world. Things would be better if I was not here.	Joiner et al. (2002)
Thwarted Belongingness	I am alone. I don't fit in anywhere. Everyone hates me.	Silva et al. (2015)

Table A3: The suicide-related constructs were composed into "archetypes" for the Shared Task. The primary archetypes of interest reflected Joiner's Interpersonal Theory of Suicide (acquired capability, perceived burdensomeness, and thwarted belongingness). Multiple rounds of evaluation and refinement were performed to ensure coherence and consistency with hypothesized aspects of suicidality.

Constructs	Instruction Prompt
Acquired Capability	Instruction: The Acquired Capability refers to a person's increased tolerance for physical pain and fear of death, which develops over time through repeated exposure to painful or fear-inducing experiences. People who experience high levels of Acquired Capability may have exposure to self-harm, previous suicide attempts, or experiences in environments where pain and death are commonplace.Read the user written input text provided below and highlight the text spans for Acquired Capability, if present. If it is not present, print 'Not Present'. Input: Running on empty I've gotten to the point in my life where I don't know what makes me happy anymore. I've been in this long term relationship, I've met new people, work is not bad. I have good things but they don't make happy anymore. This all happened a couple months back and its gotten worse and worse. I ended up cutting myself again. Lying to people and telling them I'm seeing help when I'm not. I've gotten to the point where I'll cut just because my life has nothing better to keep me busy. So I hurt myself. I feel empty except for the anxiety that one day, I will end it. I've traveled to try and spark new adventures and memories. Nothing has been able to remove this feeling in my stomach. This feeling of having nothing. If my life has no meaning. Of I can't give it meaning. It should end soon. Text Spans: Acquired Capability: 'I ended up cutting myself again', 'I've gotten to the point where I'll cut just because my life has nothing better to keep me busy', 'So I hurt myself', 'I feel empty except for the anxiety that one day, I will end it', 'If my life has no meaning. Of I can't give it meaning. It should end soon'
Perceived Burdensomeness	Instruction: Perceived Burdensomeness specifically refers to an individual's belief that their existence or presence is a burden on others. People who experience high levels of perceived burdensomeness believe that their family, friends, or society in general would be better off without them.Read the user written input text provided below and highlight the text spans for Perceived Burdensomeness, if present. If it is not present, print 'Not Present'. Input: Is life really worth living? Im not suicidal like at all... But i do question if life is even worth living. I know people die of cancer and this is extremely rude of me to say since i have a good crack at a happy life. But sorry not to sound depressing but i dont see how anyone can be happy in such a selfish world. This world just flat out sucks. So why am i here? Why are you here? Shouldn't there be a goal in the end? Besides getting married and having kids and being happy with a normal life? Text Spans: Perceived Burdensomeness: 'So why am i here? Why are you here? Shouldn't there be a goal in the end?'
Thwarted Belongingness	Instruction: Thwarted Belongingness refers to the perception or experience of not belonging or feeling disconnected from meaningful social relationships despite their efforts to form connections. People who experience thwarted belongingness feel isolated or excluded from social groups.Read the user written input text provided below and highlight the text spans for Thwarted Belongingness, if present. If it is not present, print 'Not Present'. Input: Running on empty I've gotten to the point in my life where I don't know what makes me happy anymore. I've been in this long term relationship, I've met new people, work is not bad. I have good things but they don't make happy anymore. This all happened a couple months back and its gotten worse and worse. I ended up cutting myself again. Lying to people and telling them I'm seeing help when I'm not. I've gotten to the point where I'll cut just because my life has nothing better to keep me busy. So I hurt myself. I feel empty except for the anxiety that one day, I will end it. I've traveled to try and spark new adventures and memories. Nothing has been able to remove this feeling in my stomach. This feeling of having nothing. If my life has no meaning. Of I can't give it meaning. It should end soon. Text Spans: Thwarted Belongingness: 'Lying to people and telling them I'm seeing help when I'm not'

Table A4: Prompt for Evidence Extraction

Task	Instruction Prompt
Summarization	Instruction: You are required to analyze and summarize text excerpts from an individual, who has been assigned a Moderate risk of suicide. In your summary, compile all the evidence that substantiates this level of risk. While preparing your summary, take into account the following aspects that can be deduced from their text: 1. Emotional State: Look for signs of emotional distress, sadness, or hopelessness. 2. Cognitive Processes: Pay attention to their thought patterns, decision-making abilities, and problem-solving skills. 3. Behavior and Motivation: Observe any changes in their behavior or lack of motivation to engage in activities they once enjoyed. 4. Interpersonal Relationships and Social Support: Assess the quality of their relationships and the level of social support they receive. 5. Mental Health Issues: Identify any signs of mental health disorders or issues, such as depression or anxiety. 6. Other Risk Factors: Consider any other potential risk factors that may contribute to their suicide risk. Your task is to provide a comprehensive summary that encapsulates all these factors, based on the evidence found in the individual's text. Write your summary in a single paragraph without numbering. Spans: 1. It would be so easy, and relieve me of this pain forever.You don't know what it's like going through life despising yourself, and everything you are.Everything in this world is against me even thought it isn't.I'm going to end it.Edit: In case anyone was wondering, I didn't go through with it. 2. I'm too apathetic to even kill myself anymore, I just don't know what to do at all. 3. I don't want to live anymore. 4. It would be so easy, and simple, and I could end all of this pain. 5. No one cares about me. Summary: Based on the text excerpts provided, the individual is at a Moderate risk of suicide. They express a strong desire to end their life, stating that it would be ""easy"" and ""relieve [them] of this pain forever"" (1). They also mention feeling ""despis[ed]"" and that ""everything in this world is against [them]"" (1), indicating a sense of hopelessness and isolation. Additionally, they mention being ""apathetic"" and not knowing what to do (2), which suggests a lack of motivation and problem-solving skills. The individual also expresses a belief that no one cares about them (5), further contributing to their feelings of isolation and despair. These factors, combined with the individual's explicit statements about wanting to end their life, substantiate a Moderate risk of suicide.</s>"

Table A5: Prompt for Summarization of selected spans

Joiner's Constructs	Definition	Facets/Symptoms
Thwarted belongingness	Thwarted belongingness refers to the perception or experience of not belonging or feeling disconnected from meaningful social relationships despite their efforts to form connections. People who experience thwarted belongingness feel isolated or excluded from social groups.	Loss of Social Support Isolation/Loneliness Perceived/Actual Rejection
Perceived burdensomeness	Perceived burdensomeness specifically refers to an individual's belief that their existence or presence is a burden on others. People who experience high levels of perceived burdensomeness believe that their family, friends, or society in general would be better off without them.	Belief in Being a Burden Failure to Contribute Perceived Lack of Worth
Acquired Capability	The Acquired Capability refers to a person's increased tolerance for physical pain and fear of death, which develops over time through repeated exposure to painful or fear-inducing experiences. People who experience high levels of Acquired Capability may have exposure to self-harm, previous suicide attempts, or experiences in environments where pain and death are commonplace.	Simulation Experiences of Endurance Desensitization to Harm High Tolerance for Physical Pain Engagement in Risky Behaviors Familiarity with Self-Harm Methods
Protective Factors	Protective Factor can be any factor that indicates an improvement in the person's mental health– for example, an expression of resilience, gratefulness, seeking therapy etc. It can be something that hints at the opposite of Joiner's constructs: good social support and belonging, feel worthy and grateful for life, feeling pain and being careful about their own life.	

Table A6: Annotation Schema for Internal Experts