Semantic Retrieval of BDI Symptoms in User Writings

Notebook for the eRisk Lab at CLEF 2025

Noam Munz[1,*,†], Eliya Naomi Aharon[1,*,†], Avi Segal[1] and Kobi Gal[1,2]

[1]Ben-Gurion University of the Negev, Beer-Sheva, Israel
[2]University of Edinburgh, Edinburgh, United Kingdom

Abstract

We present our approach to Task 1 of the CLEF eRisk 2025 Lab, which focuses on identifying depression symptoms in user-generated text. The task is formulated as a sentence ranking problem, aiming to retrieve sentences relevant to each of the 21 symptoms defined in the Beck Depression Inventory-II (BDI-II). The method employs Sentence-BERT to compute semantic similarity between user text and symptom queries derived from the BDI questionnaire's multiple-choice responses. To improve coverage, queries are expanded based on retrieval results from the training set. Additionally, sentences not referring to the user are filtered out to reduce noise from third-person narratives. Our approach achieved competitive performance, with Average Precision substantially exceeding the median of all submitted systems. This demonstrates the promise of semantic retrieval and first-person filtering for identifying fine-grained depressive symptoms at scale.

Keywords

Sentence-BERT, Semantic Similarity, Text Retrieval, Mental Health NLP, Beck's Depression Inventory-II, Large Language Models

1. Introduction

The CLEF eRisk 2025 Lab Task 1 [1, 2] focuses on identifying signs of depression in user-generated text. The task involves ranking sentences based on their relevance to 21 symptoms defined by the Beck Depression Inventory-II (BDI-II) [3], a widely used clinical tool for assessing depression severity. Participants are provided with sentence-level user writings and are tasked with returning, for each symptom, a ranked list of 1000 sentences that best reflect the user's mental state regarding that symptom. Relevant sentences may indicate either the presence or absence of the symptom.

Detecting fine-grained indicators of depression from text can support early intervention and improve access to mental health care [4], particularly in digital contexts where individuals often express their emotional states [5, 6].

The task presents several challenges. First, the dataset contains 17,553,441 texts, making retrieval computationally demanding. Second, many sentences reference people other than the author, introducing ambiguity around whose mental state is being described. This adds noise and requires disambiguation between self-disclosure and commentary about others [7, 8].

Our approach utilizes Sentence-BERT [9] embeddings to retrieve relevant sentences. We investigate the impact of query expansion on retrieval effectiveness and apply filtering to focus on first-person references, aiming to reduce noise from irrelevant or third-person content. Submissions conform to the TREC format and are evaluated using standard retrieval metrics including Average Precision (AP), R-Precision (R-PREC), Precision at 10 (P@10), and nDCG, with human relevance judgments created via pooling.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
munz@post.bgu.ac.il (N. Munz); eliyaah@post.bgu.ac.il (E. N. Aharon); avisegal@gmail.com (A. Segal); kobig@bgu.ac.il (K. Gal)
0009-0001-1153-2817 (N. Munz); 0009-0009-4641-785X (E. N. Aharon); 0000-0003-0915-5730 (A. Segal); 0000-0001-7187-8572 (K. Gal)

2. Related Work

This section reviews relevant work across computational methods for semantic retrieval and psychological foundations related to depression and its assessment.

2.1. Semantic Retrieval

Semantic representation methods have evolved with the popularity of transformer-based models [10, 11], which capture contextualized word and sentence embeddings [12, 13]. These models have demonstrated superior performance in encoding semantic information compared to traditional word embeddings [14]. Their ability to capture contextual dependencies enables more effective similarity measurements and downstream tasks like retrieval and classification [15, 16].

Query expansion and ranking are established techniques in retrieval tasks. Query expansion broadens the original query with related terms to capture a wider range of relevant information [17]. Ranking methods order results based on relevance scores, often leveraging scoring metrics or learned models [18, 19].

Large language models (LLMs) have shown strong capabilities in zero-shot classification, where tasks are performed without task-specific training [20]. By leveraging pre-trained knowledge, LLMs can generalize to new tasks [21]. This makes them especially useful in domains with limited labeled data [22].

2.2. Depression Symptoms and the Beck Depression Inventory

Depression is a complex mental health disorder characterized by a range of emotional, cognitive, and physical symptoms. These symptoms can include sadness, loss of interest or pleasure, disturbed sleep or changes in appetite [23, 24]. Accurate identification of these symptoms is critical for diagnosis, treatment, and research [25]. Standardized tools like the Beck Depression Inventory (BDI) provide a structured way to assess the presence and severity of depressive symptoms based on self-reported data [3].

3. Methodology

Our approach consists of representing each BDI-II symptom as a set of natural language queries, computing semantic similarity scores between these queries and sentences in the dataset, and ranking sentences accordingly. To improve retrieval, we apply query expansion based on training data and post-process results to remove non-first-person statements. This section describes the steps in detail.

3.1. Problem Formulation

The dataset, denoted as $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, consists of sentences. Each sample includes the target sentence d_i along with its preceding and following sentences. In this work we use only the target sentence itself.

The 21 BDI-II symptoms are represented as a set $\mathcal{S} = \{s_1, s_2, \ldots, s_{21}\}$. Each symptom $s_i \in \mathcal{S}$ is detailed by n graded statements $\{s_{i1}, s_{i2}, \ldots, s_{in}\}$, describing increasing severity levels of the symptom.

For each symptom s_i, the goal is to produce a ranked list R_{s_i} of the top 1000 sentences from $\mathcal{D}$ that are most relevant to s_i.

The retrieval effectiveness of R_{s_i} is evaluated against human relevance judgments to measure how well the ranking aligns with actual symptom relevance.

3.2. Sentence Embedding and Similarity Scoring

We represent both user sentences $d_i \in \mathcal{D}$ and symptom graded statements $\{s_{i1}, s_{i2}, \ldots, s_{in}\}$ using Sentence-BERT, a transformer-based model that encodes sentences into fixed-size dense vectors in a

shared semantic space.

For each sentence d_i and symptom s_i, we compute the cosine similarity between the embedding of d_i and each of the embeddings corresponding to the graded statements s_{i1} through s_{in}. The final similarity score between sentence d_i and symptom s_i is defined as the maximum of these values:

$$\text{score}(d_i, s_i) = \max_{j \in \{1,2,\ldots,n\}} \cos(\text{emb}(d_i), \text{emb}(s_{ij}))$$

This results in a relevance score for each sentence–symptom pair, which we use to produce a ranked list R_{s_i} by sorting all sentences $d_i \in \mathcal{D}$ in descending order of their scores. To illustrate, we present an example for the symptom *sadness*. The graded statements for this symptom are:

- Statement 1: "I do not feel sad."
- Statement 2: "I feel sad much of the time."
- Statement 3: "I am sad all the time."
- Statement 4: "I am so sad or unhappy that I can't stand it."

Table 1
Similarity scores between user sentences and sadness symptom statements. "Stmt" refers to the corresponding symptom statement.

#	User Sentence	Stmt 1	Stmt 2	Stmt 3	Stmt 4	Max Sim
1	What do you guys do when you feel like this?	0.295	0.431	0.305	0.288	0.431
2	I'm not sad or depressed right now.	0.887	0.388	0.321	0.322	0.887
3	So you have the video?	0.043	-0.029	-0.045	0.012	0.043
4	It's hard for a lot of us around this particular holiday.	0.23	0.504	0.363	0.330	0.504
5	I feel vaguely sad all the time.	0.468	0.907	0.920	0.746	0.920

Table 1 shows the similarity between different sentences and the four symptom statements, as well as the maximum similarity per sentence. As shown, sentence 3 is clearly unrelated to sadness and receives low similarity scores across all statements. Sentences 1 and 4 are somewhat vague, they refer to difficult feelings but don't directly mention sadness, leading to moderate scores. In contrast, sentences 2 and 5 explicitly mention sadness and receive high scores, reflecting their relevance.

Notably, sentence 2 highlights how our method handles different levels of symptom expression. The user explicitly states not feeling sad, which results in a high similarity score with Statement 1 ("I do not feel sad") and low scores with the other statements. This demonstrates how comparing against all graded statements provides greater coverage.

3.3. Query Expansion

To improve recall and capture a broader range of symptom expressions, we apply query expansion using phrases derived from previous years' datasets. For each symptom s_i, we compute similarity scores between all sentences from the 2023 and 2024 datasets and the original n BDI-II symptom graded statements $\{s_{i1}, s_{i2}, \ldots, s_{in}\}$ as described previously. This results in a similarity score for each sentence with respect to each symptom.

We then iterate over each symptom and select the top k sentences with the highest similarity scores as additional query representations for that symptom (the choice of k is discussed in Section 4). These selected sentences are treated as pseudo-relevance feedback and appended to the original query set:

$$\{s_{i1}, s_{i2}, \ldots, s_{in}\} \longrightarrow \{s_{i1}, \ldots, s_{n+k}\}$$

This approach aims to build an exhaustive query set for each symptom by covering diverse phrasings and ways users may express the symptom. Examples of original and expanded queries for two symptoms are shown in Table 2.

Table 2
Query Expansion Examples for *Sadness* and *Loss of Energy*

Symptom	Original Statements	Expanded Statements
Sadness	I do not feel sad.	Every time I get really sad.
	I feel sad much of the time.	Why do I feel sad?
	I am sad all the time.	Sometimes I'm just sad.
	I am so sad or unhappy that I can't stand it.	I just feel sad and cold inside all the time.
Loss of Energy	I have as much energy as ever.	I have so much energy now its immense.
	I have less energy than I used to have.	But I don't have the energy for anything at the moment.
	I don't have enough energy to do very much.	I often don't have the energy.
	I don't have enough energy to do anything.	I've never had much energy.

Similarity between a test sentence d_i and a symptom s_i is then computed as the maximum cosine similarity across this expanded set of $n + k$ phrases.

3.4. First Person Filtering

First-person statements offer the most direct insight into a user's mental health, as they capture self-reported experiences related to depressive symptoms [7, 8]. Since the competition task required including only sentences that provide information about the writer, we applied first-person filtering to improve the quality of the final ranking. This helps reduce noise from sentences referring to others or general situations. We employed three approaches to identify first-person language.

1. **Pronoun Filter**: a simple keyword-based filter checked for the presence of first-person pronouns including: *I*, *me*, *my*, *we*, *ourselves*, *mine*, *our*, *ours*, *I'm*, *I've*.
2. **SpaCy Filter**: spaCy[1], an open-source natural language processing library, was used to identify whether the grammatical subject of the sentence was in the first person based on syntactic dependencies and morphological features.
3. **LLM-Based Filter**: we employed a large language model (LLM) in a zero-shot classification setting to identify first-person narratives without task-specific training [26]. Specifically, we used Claude Sonnet 3.7 [27], a top-ranked model on the Hugging Face Chatbot Arena Leaderboard[2], to analyze whether texts reflected the writer's personal experience by focusing on self-references and symptom connection. Details of prompt evaluation and refinement, including the prompt text, are provided in Section 4.2.

After filtering, we produced new ranked lists R_s^{FP} containing only sentences identified as first-person narratives.

4. Experiments

4.1. Datasets

Our experiments utilize datasets from three different years: 2023, 2024, and 2025. The 2023 and 2024 datasets include labeled sentences, where each sentence is annotated with a binary indication of relevance to a symptom. For both years, we report the number of sentences for the full datasets as well as for annotated subsets based on majority vote and full annotator consensus (Table 3). The 2025 dataset used for the current task is unlabeled and contains only raw user sentences.

All datasets follow the TREC format, where each sample includes a document ID, the target sentence, as well as the preceding and following sentences (though only the target sentence is used in this work).

[1] https://spacy.io/
[2] https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard

Table 3
Number of sentences in each dataset by year and annotation type

Dataset	Number of Sentences
2023	4,264,693
2023 Majority Vote	21,581
2023 Consensus	21,580
2024	15,542,200
2024 Majority Vote	14,823
2024 Consensus	14,823
2025	17,553,441

4.2. Hyperparameter Tuning

We tuned two key hyperparameters to optimize retrieval performance.

Query Expansion Size k: We tested values of k ranging from 10 to 100 in increments of 10. For each k, we evaluated retrieval quality using the merged 2023-2024 consensus labeled dataset (Section 4.4). Performance improved up to $k = 30$ and then slightly declined, so we selected $k = 30$ for the final expansions.

LLM Prompt Refinement: We applied the similarity scoring and query expansion to the 2023 and 2024 datasets to obtain the top 100 sentences per symptom, creating a pool of 2,100 sentences. After removing duplicates, we randomly sampled 200 sentences for evaluation. Using the prompted LLM, each sentence was labeled for first-person language. Two annotators assessed labeling accuracy. The prompt wording was iteratively refined to improve the LLM's accuracy until improvements plateaued.

Prompt Used for Annotation (Final Version)

Analyze the following text to determine if it provides information about the writer's personal experience with the specified symptom.
`<symptom>{symptom}</symptom>`
`<text>{text}</text>`
Consider the text informative (YES) if it reveals anything about the writer's personal relationship with the symptom – whether they have it, had it, are recovering from it, don't have it, etc.
Consider the text non-informative (NO) if the symptom is only mentioned in relation to other people or discussed generally without personal connection to the writer.
Pay special attention to first-person language and direct self-references that connect the writer to the symptom.
Return only "YES" or "NO" based on your analysis.

4.3. Configurations

We evaluated five retrieval configurations to assess the impact of query expansion and first-person filtering strategies:

- **sbert**: baseline model using Sentence-BERT with original BDI-II symptom queries.
- **sbert-w-expansion**: adds top-30 high-scoring training sentences to each symptom's query set for expanded semantic coverage.
- **sbert-w-expansion-w-naive-fp**: applies the **Pronoun Filter**, which detects first-person language using keyword matching.
- **sbert-w-expansion-w-spacy-fp**: applies the **spaCy Filter**, which identifies first-person grammatical subjects via syntactic parsing.
- **sbert-w-expansion-w-naive-fp-w-claude**: applies the **LLM-Based Filter**, which verifies first-person relevance through LLM-based classification.

4.4. Evaluation

To evaluate our retrieval configurations, we used a labeled test set created by combining the 2023 and 2024 consensus datasets. These datasets contain high-confidence binary annotations indicating sentence relevance to each BDI-II symptom. The merged evaluation set includes 36,403 annotated sentences.

Following common practice [28], we generated a ranked list of 100 sentences per symptom for each configuration and evaluated it against the labeled set. The evaluation metrics were:

- **Precision@k** for $k \in \{10, 30, 50\}$: the proportion of relevant sentences in the top-k positions of each ranking.
- **Average Precision (AP)**: the average of precision scores at each rank where a relevant sentence appears.
- **R-Precision (R-PREC)**: precision at R, where R is the total number of relevant sentences for a given symptom.
- **nDCG** (normalized Discounted Cumulative Gain): a rank-aware metric that rewards placing relevant items higher in the list.

All metrics were computed per symptom and then averaged over the 21 BDI-II symptoms.

4.5. Implementation

All experiments were conducted on a machine with an NVIDIA RTX 6000 GPU using the `sentence-transformers/nli-roberta-base-v2` model via the SentenceTransformers library. The spaCy filter used spaCy's *en_core_web_sm* pipeline.

5. Results

We now describe the final competition results. For each of the 21 symptoms defined by the BDI-II questionnaire, participating teams were required to submit a ranked list of up to 1,000 relevant sentences. Each team was permitted to submit up to five system configurations for evaluation. In total, 17 teams took part in the eRisk 2025 Task 1 competition, resulting in 67 submitted runs.

The evaluation process involved three expert assessors who independently judged the relevance of sentences for each symptom. Relevance was determined using two complementary criteria: under majority voting, a sentence was deemed relevant if at least two assessors agreed; under unanimity voting, relevance required consensus among all three assessors. System performance was assessed using standard ranking metrics, including Average Precision, R-Precision, Precision@10, and NDCG.

We also report the top-performing runs submitted by other teams for each evaluation setting. Specifically, we include two configurations from Team INESC-ID: one that achieved the highest scores in Average Precision, R-Precision, and Precision@10, and another that achieved the best NDCG. In addition, we report the mean and median scores across all submitted runs, following common practice in prior work [29].

In both majority (Table 4) and unanimity (Table 5) voting evaluations , our **sbert** configuration achieved the highest Average Precision, R-Precision, and NDCG among our tested methods. This suggests heuristic query expansions and filters may add noise that lowers overall ranking quality. However, combining query expansion with first-person filtering improved Precision@10, indicating that first-person filtering may help prioritize personal disclosures at the top. Among these filters, the LLM-based approach performed best, likely due to its enhanced semantic understanding of context.

In the unanimity voting evaluation (Table 5), all our configurations scored lower across metrics compared to our majority voting results, reflecting the stricter relevance criterion. The relative benefit of first-person filtering on Precision@10 was slightly higher under this stricter setting, though the **sbert** configuration still remained strongest on overall ranking metrics.

Compared to other teams, our approach consistently performed above the overall mean and median across all reported metrics, demonstrating competitiveness in this domain.

Table 4
Majority Voting Results. Best results are in bold.

Team	Configuration	AP	R-PREC	P@10	NDCG
INESC-ID	unanimity	**0.354**	**0.433**	**0.876**	0.575
INESC-ID	max	0.350	0.407	0.648	**0.653**
All Teams	Mean	0.126	0.185	0.406	0.300
All Teams	Median	0.077	0.168	0.400	0.290
BGU Data-Science Lab	sbert	0.240	0.324	0.743	0.516
	sbert-w-expansion	0.197	0.281	0.652	0.444
	sbert-w-expansion-w-spacy-fp	0.220	0.287	0.767	0.463
	sbert-w-expansion-w-naive-fp	0.227	0.296	0.767	0.475
	sbert-w-expansion-w-naive-fp-w-claude	0.232	0.305	0.767	0.483

Table 5
Unanimity Voting Results. Best results are in bold.

Team	Configuration	AP	R-PREC	P@10	NDCG
INESC-ID	unanimity	**0.269**	**0.383**	**0.509**	0.561
INESC-ID	max	0.223	0.308	0.386	**0.582**
All Teams	Mean	0.083	0.141	0.224	0.265
All Teams	Median	0.052	0.117	0.205	0.270
BGU Data-Science Lab	sbert	0.171	0.272	0.419	0.489
	sbert-w-expansion	0.119	0.223	0.381	0.389
	sbert-w-expansion-w-spacy-fp	0.135	0.237	0.462	0.412
	sbert-w-expansion-w-naive-fp	0.138	0.240	0.448	0.420
	sbert-w-expansion-w-naive-fp-w-claude	0.143	0.244	0.443	0.429

6. Conclusions and Future Work

We presented a retrieval approach using Sentence-BERT embeddings combined with query expansion and first-person filtering to identify BDI-II symptoms in user text. While query expansion and filtering aimed to improve retrieval, the baseline model without expansions performed best on most ranking metrics. This suggests that adding heuristic expansions and filters may introduce noise and reduced overall ranking quality. However, filtering to emphasize self-references helped increase the number of relevant results at top ranks.

Our study was limited to five configurations, which constrains detailed understanding of each component's contribution. This is because our internal evaluations were performed on a smaller labeled dataset, where some symptoms were underrepresented. The official competition results, which are more robust, were also available only for the five submitted configurations.

Future work should focus on several key areas. A thorough ablation study is needed to isolate the effects of query expansion and different filtering methods, addressing the limitations of our current evaluations. In addition, query expansion could be improved by curating higher-quality phrases through qualitative analysis and incorporating more diverse data sources to better capture varied symptom expressions. Considering sentence context rather than treating sentences in isolation may help better reflect user intent and improve consistency. Training symptom-specific classifiers on labeled data that integrate first-person detection directly into the model could further enhance precision beyond semantic similarity. Finally, exploring the use of large language models to generate candidate queries or symptom expressions, despite the higher computational cost, is another promising direction.

7. Acknowledgments

This work was funded in part by the Israeli Science Foundation grant no. 1302/21.

Declaration on Generative AI

The authors used generative AI tools to assist with grammar refinement and phrasing corrections throughout the writing process.

References

[1] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings, Part II, volume To be published of *Lecture Notes in Computer Science*, Springer, 2025.

[2] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 9-12 September, 2025, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[3] A. T. Beck, R. A. Steer, G. Brown, Beck depression inventory–ii, Psychological assessment (1996).

[4] K. Mao, W. Yuqi, C. Jie, A systematic review on automated clinical depression diagnosis. npj mental health research, 2 (1), 20, 2023.

[5] Y. Ophir, R. Tikochinski, C. S. Asterhan, I. Sisso, R. Reichart, Deep neural networks detect suicide risk from textual facebook posts, Scientific reports 10 (2020) 16685.

[6] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, A. Ulhaq, Depression detection from social network data using machine learning techniques, Health information science and systems 6 (2018) 1–12.

[7] T. Edwards, N. S. Holtzman, A meta-analysis of correlations between depression and first person singular pronoun use, Journal of Research in Personality 68 (2017) 63–68.

[8] X. Ren, H. A. Burkhardt, P. A. Areán, T. D. Hull, T. Cohen, Deep representations of first-person pronouns for prediction of depression symptom severity, in: AMIA Annual Symposium Proceedings, volume 2023, 2024, p. 1226.

[9] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[10] A. Ettinger, What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models, Transactions of the Association for Computational Linguistics 8 (2020) 34–48.

[11] J. Turton, D. Vinson, R. E. Smith, Deriving contextualised semantic features from bert (and other transformer model) embeddings, arXiv preprint arXiv:2012.15353 (2020).

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), 2019, pp. 4171–4186.

[14] K. Ethayarajh, How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings, arXiv preprint arXiv:1909.00512 (2019).

[15] A. Bialer, D. Izmaylov, A. Segal, O. Tsur, Y. Levi-Belz, K. Gal, Detecting suicide risk in online counseling services: A study in a low-resource language, arXiv preprint arXiv:2209.04830 (2022).

[16] A. Abolghasemi, S. Verberne, L. Azzopardi, Improving bert-based query-by-document retrieval

with multi-task optimization, in: European Conference on Information Retrieval, Springer, 2022, pp. 3–12.
[17] B. Aklouche, I. Bounhas, Y. Slimani, Query expansion based on nlp and word embeddings., in: TREC, 2018.
[18] H. Steck, C. Ekanadham, N. Kallus, Is cosine-similarity of embeddings really about similarity?, in: Companion Proceedings of the ACM Web Conference 2024, 2024, pp. 887–890.
[19] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, X. Cheng, A deep look into neural ranking models for information retrieval, Information Processing & Management 57 (2020) 102067.
[20] Y. Chae, T. Davidson, Large language models for text classification: From zero-shot learning to fine-tuning, Open Science Foundation 10 (2023).
[21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
[22] B. Ding, C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W. Xia, J. Hu, L. A. Tuan, S. Joty, Data augmentation using llms: Data perspectives, learning paradigms and challenges, in: Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 1679–1705.
[23] K. S. Kumar, S. Srivastava, S. Paswan, A. S. Dutta, et al., Depression-symptoms, causes, medications and therapies, The Pharma Innovation 1 (2012) 37.
[24] J. LeMoult, I. H. Gotlib, Depression: A cognitive perspective, Clinical psychology review 69 (2019) 51–66.
[25] L. S. Goldman, N. H. Nielsen, H. C. Champion, A. M. A. Council on Scientific Affairs, Awareness, diagnosis, and treatment of depression, Journal of general internal medicine 14 (1999) 569–580.
[26] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, 2023. URL: https://arxiv.org/abs/2205.11916. arXiv:2205.11916.
[27] Anthropic, Claude language model (version 3.7), https://www.anthropic.com/claude, 2023. Accessed on [date].
[28] V. Pavlu, J. Aslam, A practical sampling strategy for efficient retrieval evaluation, College of Computer and Information Science, Northeastern University (2007).
[29] A. Barachanou, F. Tsalakanidou, S. Papadopoulos, Rebecca at erisk 2024: search for symptoms of depression using sentence embeddings and prompt-based filtering, Working Notes of CLEF (2024) 9–12.