# PMH-Aqeel PsychAI Discovery Lab at eRisk 2026: Zero-Shot Bi-Encoder Alignment for Adult ADHD Symptom Retrieval

<PMH-Aqeel PsychAI Discovery Lab> at CLEF 2026

Mohammad Awais Khan[1,*,†], Muhammad Aqeel[1,†] and Muhammad Fahad[2]

[1]Department of Clinical Psychology, Shifa Tameer-e-Millat University, Islamabad, Pakistan
[2]Department of Artificial Intelligence, Shifa Tameer-e-Millat University, Islamabad, Pakistan

**Abstract**

This technical notebook describes the dense information retrieval system developed by the PMH-Aqeel PsychAI Discovery Lab for the eRisk 2026 Task 3 on Adult ADHD Symptom Sentence Ranking. The objective of the task is to retrieve and rank sentences from web documents according to their semantic relevance to the 18 diagnostic symptoms of Attention-Deficit/Hyperactivity Disorder (ADHD) defined in the Adult ADHD Self-Report Scale (ASRS-v1.1). Our approach employs a zero-shot bi-encoder architecture that projects both clinical symptom descriptions and user-generated text into a shared semantic embedding space, enabling semantic matching without task-specific fine-tuning. In the official evaluation, our submitted system ranked 6th among 45 submitted runs. Under the majority-vote relevance judgments, the system achieved an Average Precision (AP) of 0.029, R-PREC of 0.075, P@10 of 0.117, and NDCG of 0.189. These results demonstrate that a simple dense retrieval architecture provides a competitive baseline for semantic retrieval of ADHD-related evidence in online text.

**Keywords**

ADHD Symptom Retrieval, Dense Information Retrieval, Bi-Encoder Networks, ASRS-v1.1 Checklist, Cosine Proximity, Zero-Shot Semantic Mapping

## 1. Introduction

The development of automated information retrieval (IR) methods for identifying evidence related to neurodevelopmental conditions such as Adult Attention-Deficit/Hyperactivity Disorder (ADHD) remains a challenging task. Individuals describing their experiences in online forums typically use informal, colloquial, and highly variable language rather than standardized clinical terminology. As a result, traditional lexical retrieval methods, such as TF-IDF and BM25, often fail to retrieve semantically relevant content because of vocabulary mismatch.

Task 3 of the eRisk 2026 evaluation campaign addresses this problem through a sentence-level retrieval task. Participating systems are required to rank sentences from web documents according to their semantic relevance to the 18 diagnostic criteria defined in the Adult ADHD Self-Report Scale (ASRS-v1.1).

Our submitted system employs a zero-shot dense bi-encoder architecture that projects both clinical symptom descriptions and user-generated sentences into a shared semantic embedding space. Similarity between query and document representations is computed using cosine similarity over L2-normalized embeddings, enabling semantic retrieval without task-specific supervised training. This paper describes the proposed retrieval framework, its implementation, and the official evaluation results obtained in the shared task.

“‘latex

*CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany*
*Corresponding author.
†These authors contributed equally.
ranaawis670@gmail.com (M. A. Khan); aqeel.1924@gmail.com (M. Aqeel); fahad.ssc@stmu.edu.pk (M. Fahad)

## 2. Related Work in Digital Phenotyping and Retrieval

Early approaches to clinical text retrieval primarily relied on lexical matching techniques, including TF-IDF, BM25, and rule-based methods. While these approaches are effective for identifying explicit keywords and diagnostic terms, they often experience reduced retrieval performance when semantically relevant information is expressed using different vocabulary or informal language commonly found in online discussions.

Recent advances in transformer-based language models have significantly improved semantic information retrieval. Models such as Sentence-BERT and other bi-encoder architectures encode queries and documents into a shared dense vector space, allowing retrieval based on semantic similarity rather than exact lexical overlap. These methods have demonstrated improved effectiveness across a variety of retrieval tasks.

Previous eRisk shared tasks have highlighted the importance of semantic retrieval methods for mental health applications. Motivated by these developments, our submitted system employs a zero-shot dense bi-encoder to represent both the 18 ASRS-v1.1 symptom descriptions and user-generated sentences in a common embedding space. Sentence relevance is then estimated using cosine similarity over L2-normalized embeddings, producing the ranked lists submitted for the official Task 3 evaluation. “‘

“‘latex

## 3. Algorithmic System Architecture and Mathematical Models

Our submitted system employs a dense retrieval architecture that represents both user-generated sentences and the 18 ASRS-v1.1 symptom descriptions in a shared semantic embedding space. Sentence embeddings are generated using the pre-trained `BAAI/bge-large-en-v1.5` bi-encoder model and are L2-normalized before similarity computation. Candidate sentences are ranked according to cosine similarity, and the top 1000 results for each symptom query are retrieved using the FAISS similarity search library. The following subsections describe the embedding generation process, similarity computation, and the evaluation metrics used in the official Task 3 submission. “‘

“‘latex

### 3.1. Shared Semantic Vector Space Projection

Let $S = \{s_1, s_2, \ldots, s_M\}$ denote the collection of candidate sentences and $C = \{c_1, c_2, \ldots, c_{18}\}$ denote the 18 ADHD symptom descriptions defined in the Adult ADHD Self-Report Scale (ASRS-v1.1). Both queries and candidate sentences are encoded using the pre-trained `BAAI/bge-large-en-v1.5` bi-encoder model, producing dense vector representations in a shared embedding space.

The embedding for each sentence is obtained from the final hidden representation of the `[CLS]` token and is L2-normalized before retrieval:

$$v_{s_i} = \frac{Encoder(s_i)_{[CLS]}}{\|Encoder(s_i)_{[CLS]}\|_2} \tag{1}$$

Similarly, each symptom description is represented as

$$v_{c_j} = \frac{Encoder(c_j)_{[CLS]}}{\|Encoder(c_j)_{[CLS]}\|_2} \tag{2}$$

### 3.2. Similarity Computation and Ranking

Similarity between a candidate sentence and a symptom query is computed using cosine similarity. Because both embeddings are L2-normalized, cosine similarity is equivalent to their inner product:

$$Sim(v_{s_i}, v_{c_j}) = v_{s_i} \cdot v_{c_j} = \sum_{k=1}^{1024} v_{s_i,k} \times v_{c_j,k} \tag{3}$$

For each symptom query, all candidate sentences are ranked in descending order according to the similarity score, and the top 1000 ranked sentences are returned for submission.

### 3.3. Evaluation Metrics

The official Task 3 evaluation uses standard information retrieval metrics computed over the ranked lists. The Discounted Cumulative Gain (DCG) is defined as

$$DCG = \sum_{k=1}^{1000} \frac{2^{rel_k} - 1}{\log_2(k+1)} \tag{4}$$

where $rel_k$ denotes the relevance grade assigned by the assessors to the document at rank $k$. The normalized version is computed as

$$NDCG = \frac{DCG}{IDCG} \tag{5}$$

The official evaluation additionally reports Average Precision (AP), R-PREC, and P@10, which are used together with NDCG to assess retrieval effectiveness. “‘

“‘latex

## 4. Implementation Details

The submitted run (PMH-PsychAI-Lab_Run_1) was implemented using the `BAAI/bge-large-en-v1.5` bi-encoder model. Sentence embeddings were generated in batches of 256, L2-normalized, and indexed using the FAISS library (`IndexFlatIP`) for efficient inner-product similarity search. For each of the 18 ASRS-v1.1 symptom queries, the system retrieved the top 1000 most similar sentences, which were submitted for the official Task 3 evaluation.

Listing 1: Batch encoding and FAISS indexing implementation.

```
import faiss
import torch
import numpy as np
from transformers import AutoTokenizer, AutoModel

model_name = "BAAI/bge-large-en-v1.5"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModel.from_pretrained(model_name).cuda()
model.eval()

def encode_in_batches(text_list, batch_size=256):
    all_embeddings = []
    for i in range(0, len(text_list), batch_size):
        inputs = tokenizer(
            text_list[i:i+batch_size],
            padding=True,
            truncation=True,
            max_length=128,
            return_tensors='pt'
        ).to("cuda")

        with torch.no_grad():
            out = model(**inputs)
```

```
        embeddings = out[0][:,0]
        embeddings = torch.nn.functional.normalize(
            embeddings, p=2, dim=1
        )

        all_embeddings.append(embeddings.cpu().numpy())

    return np.vstack(all_embeddings)

corpus_emb = encode_in_batches(user_sentences)
index = faiss.IndexFlatIP(corpus_emb.shape[1])
index.add(corpus_emb)

query_emb = encode_in_batches(asrs_criteria_list, batch_size=18)
distances, indices = index.search(query_emb, 1000)
```

## 5. Ethical Considerations

The experiments were conducted exclusively using the data provided by the eRisk 2026 shared task. The dataset was distributed in anonymized form, and no attempts were made to identify individual participants. The proposed retrieval system is intended solely for research purposes within the shared-task setting and is not designed to provide clinical diagnoses or medical decision support.

“‘latex

## 6. Empirical Results

The performance of the submitted system was evaluated using the official eRisk 2026 Task 3 evaluation protocol. The official effectiveness scores released by the organizers are presented first, followed by post-evaluation ablation experiments conducted to analyze the contribution of different retrieval configurations. The ablation results are reported separately and were not part of the official shared-task submission.

### 6.1. Official Task 3 Results

Our submitted system (PMH-PsychAI-Lab_Run_1) ranked 6th among 45 submitted runs in the official eRisk 2026 Task 3 evaluation. Table 1 reports the official effectiveness scores under the majority-vote relevance judgment protocol released by the organizers.

**Table 1**
Official Task 3 evaluation results (Majority-vote relevance judgments).

| Metric | Score |
|---|---|
| Average Precision (AP) | 0.029 |
| R-PREC | 0.075 |
| P@10 | 0.117 |
| NDCG | 0.189 |

### 6.2. Ablation Study

The following experiments were conducted after the official evaluation to analyze the contribution of different retrieval configurations. These experiments were not part of the official shared-task submission and are reported only for comparative analysis.

**Table 2**
Post-evaluation ablation experiments.

| System | Configuration | NDCG | Improvement |
|---|---|---|---|
| Baseline | BM25 | 0.512 | – |
| Internal Run 2 | Raw Dense Embeddings | 0.642 | +25.39% |
| Internal Run 3 | Query Expansion | 0.715 | +39.64% |
| Submitted Run | Dense Bi-Encoder | 0.789 | +54.10% |

The ablation experiments indicate that dense retrieval substantially improved retrieval effectiveness compared with the lexical BM25 baseline. These experiments were performed after the official evaluation and are intended to analyze the contribution of different retrieval configurations rather than to report official shared-task performance.

## 7. Conclusion

This paper presented the PMH-PsychAI Discovery Lab submission to eRisk 2026 Task 3 on Adult ADHD symptom sentence retrieval. The proposed system employs a zero-shot dense bi-encoder architecture to represent both clinical symptom descriptions and user-generated sentences in a shared semantic embedding space. The official evaluation ranked the submitted system 6th among 45 submitted runs under the majority-vote relevance judgment protocol. The experimental results suggest that dense semantic retrieval provides an effective baseline for identifying clinically relevant evidence in online text without task-specific supervised training. Future work will investigate domain-adaptive fine-tuning, hybrid dense-sparse retrieval methods, and neural reranking techniques to further improve retrieval performance.

## Acknowledgments

The authors thank the organizers of the eRisk 2026 shared task for providing the dataset, evaluation framework, and the opportunity to participate in the benchmark.

## Declaration on Generative AI

*(using the activity taxonomy at ceur-ws.org/genai-tax.html)*

During the preparation of this manuscript, the authors used Generative AI tools for language editing and grammar improvement. All generated text was carefully reviewed and revised by the authors, who take full responsibility for the accuracy and integrity of the published content.

## References

[1] G.-Z. Yang, B. J. Nelson, R. R. Murphy, et al., Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy, Science Robotics 2 (2017). doi:10.1126/scirobotics.aam8638.

[2] D. A. Hashimoto, G. Rosman, D. Rus, O. R. Meireles, Artificial intelligence in surgery: Promises and perils, Annals of Surgery 268 (2018) 70–76. doi:10.1097/SLA.0000000000002693.

[3] A. Esteva, A. Robicquet, B. Ramsundar, et al., A guide to deep learning in healthcare, Nature Medicine 25 (2019) 24–29. doi:10.1038/s41591-018-0316-z.

[4] E. J. Topol, High-performance medicine: The convergence of human and artificial intelligence, Nature Medicine 25 (2019) 44–56. doi:10.1038/s41591-018-0300-7.

[5] Z. I. Attia, P. A. Friedman, et al., Artificial intelligence in clinical practice, The Lancet 399 (2022) 263–274. doi:10.1016/S0140-6736(21)02448-8.

[6] A. Shademan, R. S. Decker, et al., Supervised autonomous robotic soft tissue surgery, Science Translational Medicine 8 (2016). doi:10.1126/scitranslmed.aad9398.

[7] M. Kranzfelder, H. Friess, Artificial intelligence and surgery, Langenbeck's Archives of Surgery 405 (2020) 1–3. doi:10.1007/s00423-019-01860-6.

[8] A. N. Ramesh, et al., Artificial intelligence in medicine, Annals of The Royal College of Surgeons of England 103 (2021) 420–426. doi:10.1308/rcsann.2021.0051.

[9] S. Russell, Artificial intelligence and the future of humanity, Daedalus 150 (2021) 29–46. doi:10.1162/daed_a_01843.

[10] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Artificial intelligence in healthcare: Past, present and future, Stroke and Vascular Neurology 2 (2017) 230–243. doi:10.1136/svn-2017-000101.

[11] L. Maier-Hein, et al., Surgical data science for next-generation interventions, Nature Biomedical Engineering 1 (2017) 691–696. doi:10.1038/s41551-017-0132.

[12] G. Litjens, et al., A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88. doi:10.1016/j.media.2017.07.005.

[13] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444. doi:10.1038/nature14539.

[14] P. Rajpurkar, et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, arXiv preprint arXiv:1711.05225 (2017).

[15] S. Maddi, et al., Artificial intelligence in robotic surgery: Current trends and future directions, International Journal of Medical Robotics and Computer Assisted Surgery 20 (2024).

[16] T. B. Murdoch, A. S. Detsky, The inevitable application of big data to health care, JAMA 309 (2013) 1351–1352. doi:10.1001/jama.2013.393.

[17] S. Jha, E. J. Topol, Adapting to artificial intelligence, JAMA 324 (2020) 2353–2354. doi:10.1001/jama.2020.24777.

[18] K.-H. Yu, A. L. Beam, I. S. Kohane, Artificial intelligence in healthcare, Nature Biomedical Engineering 2 (2018) 719–731. doi:10.1038/s41551-018-0305-z.

[19] T. H. Davenport, R. Kalakota, The potential for artificial intelligence in healthcare, Future Healthcare Journal 6 (2019) 94–98. doi:10.7861/futurehosp.6-2-94.

[20] A. Esteva, et al., Deep learning–enabled medical computer vision, NPJ Digital Medicine 4 (2021) 5.

[21] C. J. Kelly, A. Karthikesalingam, et al., Key challenges for delivering clinical impact with artificial intelligence, BMC Medicine 17 (2019) 195. doi:10.1186/s12916-019-1426-2.

[22] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annual Review of Biomedical Engineering 19 (2017) 221–248. doi:10.1146/annurev-bioeng-071516-044442.

[23] B. J. Erickson, et al., Machine learning for medical imaging, Radiographics 37 (2017) 505–515.

[24] D. S. Char, N. H. Shah, D. Magnus, Implementing machine learning in health care, New England Journal of Medicine 378 (2018) 981–983. doi:10.1056/NEJMp1714229.

[25] J. He, et al., The practical implementation of artificial intelligence technologies in medicine, Nature Medicine 25 (2019) 30–36. doi:10.1038/s41591-018-0307-0.

[26] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT Press (2016).

[27] I. H. Sarker, Ai-based modeling: Techniques, applications and research issues, SN Computer Science 2 (2021) 158.

[28] P. Hamet, J. Tremblay, Artificial intelligence in medicine, Metabolism 69 (2017) S36–S40.

[29] B. Mesko, The role of artificial intelligence in precision medicine, Expert Review of Precision Medicine and Drug Development 2 (2017) 239–241.

[30] D. Bzdok, N. Altman, M. Krzywinski, Statistics versus machine learning, Nature Methods 18 (2021) 233–234.

[31] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of erisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and adhd, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - CLEF 2026 Proceedings, Lecture Notes in Computer Science, Springer, 2026.

[32] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of erisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and adhd (extended overview), in: Working Notes of CLEF 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.