SymptomAI @ eRisk 2026: Layered Retrieval for ADHD Symptom Sentence Ranking*

Notebook for the eRisk Lab at CLEF 2026

Syeda Samah Daniyal[1,*,†], Yusra Faisal[1,†], Fakeha Faisal[1,†], Faisal Alvi[1] and Abdul Samad[1]

[1]Dhanani School of Science and Engineering, Habib University, Karachi, Pakistan

Abstract

The eRisk Lab explores methodologies, evaluation metrics, and real-world applications for early risk detection on the Internet, particularly in health and safety-related domains. In CLEF 2026, eRisk features Task 3, which introduces sentence-level ranking for the 18 ADHD symptoms defined in ASRS–v1.1, requiring systems to retrieve clinically relevant evidence beyond keyword matching.

We implemented progressively enhanced retrieval pipelines combining medical-domain sentence embeddings, query expansion, first-person filtering, and LLM reranking. A judge-driven diagnostic cycle using GPT-4o identified failure modes on three specific symptoms after Run 3, which were corrected through a third-person penalty, targeted negation detection, rewritten query strings, an expanded reranker pool, and per-symptom prompting. An internal GPT-4o judge evaluation of Run 4 returned a mean P@10 of 0.9278 and a mean nDCG@10 of 0.9710 across all 18 symptoms; official human-judged evaluation did not confirm this level of precision, returning a P@10 of 0.300 under majority judgments, which we attribute to evaluator–reranker correlation and analyse in detail. Despite this gap, Run 4 achieves the best AP and P@10 of the four submitted runs under official judgments, and the judge-driven diagnostic cycle is confirmed as an effective method for identifying retrieval failure modes.

Keywords

CLEF eRisk, Early Risk Detection, Transformer-based Models, Confidence Scoring, ADHD

1. Introduction

Early risk detection on the Internet is concerned with identifying markers of mental health conditions as they emerge in online discourse. The eRisk Lab was among the first evaluation initiatives to address this problem, establishing protocols that account for temporal and resource constraints in addition to accuracy. Unlike conventional text classification, eRisk formulates the task as sequential decision-making under limited time, more closely approximating real-world monitoring conditions.

The 2026 eRisk edition [1, 2] introduces Task 3: Sentence Ranking for ADHD Symptoms, which extends this paradigm to symptom-level granularity. The task defines 18 symptoms drawn from the ASRS–v1.1 checklist, and systems are required to rank candidate sentences by their relevance to each symptom. Performance is evaluated using standard information retrieval metrics, MAP, nDCG, and P@10, with the objective of surfacing clinically meaningful evidence rather than lexical overlap.

Our contributions emphasise layered retrieval pipelines that combine semantic precision with linguistic salience signals. The remainder of this paper details our system design, empirical results, and planned extensions.

CLEF 2026 Working Notes, September 21–24, 2026, Jena, Germany

*Corresponding author.

†These authors contributed equally.

samah.daniyal@gmail.com (S. S. Daniyal); yf08967@st.habib.edu.pk (Y. Faisal); fakehaafaisal@gmail.com (F. Faisal); faisal.alvi@sse.habib.edu.pk (F. Alvi); abdul.samad@sse.habib.edu.pk (A. Samad)

0009-0004-4765-5271 (S. S. Daniyal); 0009-0009-5181-0349 (Y. Faisal); 0009-0000-6147-8567 (F. Faisal); 0000-0003-3827-7710 (F. Alvi); 0009-0009-5166-6412 (A. Samad)

2. Literature Review

2.1. Task 3: Sentence Ranking for ADHD Symptoms

Recent research in computational linguistics has focused on how we can detect mental health conditions like ADHD from online text. Earlier studies mostly tried to classify whole users as either ADHD or non-ADHD. But in the eRisk setting, the focus has shifted to something more detailed. Instead of just predicting risk, researchers now try to rank individual sentences based on how closely they relate to specific clinical symptoms [3], [4]. The aim is to find clear textual evidence connected to structured symptom scales.

Barrios et al. [5] found that people with ADHD often write differently. Their stories are usually shorter, less organized, and use less varied vocabulary. Using stylometric features like character n-grams, traditional models such as Support Vector Machines (SVM) achieved strong results. This shows that writing style and structure can reflect underlying cognitive differences.

Zhu et al. [6] built on this by combining traditional machine learning with Large Language Models (LLMs). Their hybrid system used LLMs together with SVM and transformer models. LLMs are good at understanding context and deeper meaning, but they can sometimes over-predict. By using ensemble methods, the system becomes more balanced and reduces false positives.

For sentence-level symptom ranking, semantic similarity methods are commonly used. Mendoza et al. [3] proposed representing each symptom as an embedding and ranking sentences using cosine similarity. Munz et al. [4] used Sentence-BERT (SBERT) to place both symptom descriptions and user sentences in the same semantic space, helping the model catch indirect or subtle symptom expressions.

Since social media text can be noisy, techniques like query expansion and first person filtering help improve accuracy [4]. Overall, the research shows that combining writing-style features, semantic embeddings, and careful filtering leads to more reliable symptom-level retrieval.

3. Approach

3.1. Task 3: ADHD Symptom Sentence Ranking

3.1.1. Dataset

The Task 3 dataset consists of two components. First, 18 clinically defined ADHD symptoms from the ASRS–v1.1, based on DSM criteria, serve as queries for sentence ranking. Second, Reddit user data in TREC format comprises approximately 4,520 user files. Each file contains structured records with <DOCNO> (sentence identifier), <TEXT> (main sentence), and <PRE>/<POST> context fields. The data is informal and noisy, characteristic of social media. The task requires 18 ranked lists, one per symptom, evaluated with MAP, nDCG@100, and Precision@10.

Figure 1 illustrates the TREC record format and the ranking pipeline input–output structure.

3.1.2. System Design

Run 1: Cosine Similarity Baseline. Using `all-MiniLM-L6-v2`, symptom descriptions and candidate sentences are encoded as embeddings and ranked by cosine similarity with no filtering or augmentation. This directly follows Mendoza et al. [3].

Run 2: Medical SBERT, Query Expansion, and Context Window. Three improvements over Run 1: (i) the embedding model is replaced with `S-PubMedBERT-MS-MARCO` for clinical-domain representations; (ii) each ASRS symptom description is expanded with clinical synonyms and informal paraphrases (e.g., "forgot to do something," "memory problems" for symptom 3); and (iii) the context window encodes PRE + TEXT + POST rather than TEXT alone. This follows Munz et al. [4].

Run 3 (*SymptomAI FP Filter LLM Rerank*): First-Person Filtering and LLM Reranking. Two additional ranking layers are applied on top of Run 2 embeddings. Self-reported expressions are boosted: sentences with strong first-person markers (*I*, *my*, *myself*) receive a $1.3\times$ score multiplier;

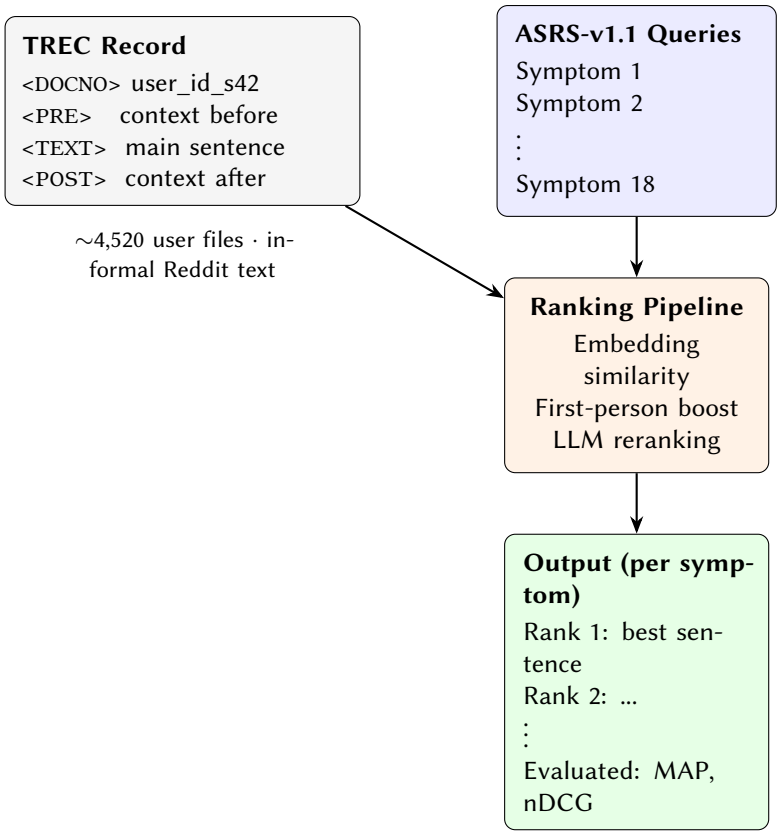

Figure 1: Task 3 dataset structure. TREC-formatted Reddit posts and 18 ASRS-v1.1 symptom queries are combined in the ranking pipeline to produce 18 ordered candidate lists.

those with weaker markers (*we*, *our*) receive $1.1\times$. The top 50 candidates per symptom are then passed to GPT-4o-mini for clinical reranking. Final scores combine both signals:

$$\text{score} = 0.6 \times s_{\text{LLM}} + 0.4 \times (s_{\text{embed}} + s_{\text{fp}}) \tag{1}$$

This hybrid approach follows Zhu et al. [6], using LLM contextual reasoning to refine precision at the top of the ranking where clinical relevance is most critical, while preserving the broad recall of embedding-based retrieval for the full candidate set.

Run 4 (*SymptomAI Query ExpNeg Filter LLM*): Judge-Driven Targeted Fixes. After submitting Run 3, GPT-4o was used as an independent judge to evaluate the top-10 results per symptom, labelling each sentence as RELEVANT, PARTIAL, or NOT_RELEVANT. The judge identified three recurring failure modes concentrated on symptoms 12, 16, and 17. Run 4 was built entirely from the corrective actions the judge's output motivated. Six changes were made.

(i) Third-person penalty. Run 3 used three pronoun tiers. Run 4 adds a fourth: sentences with third-person pronouns (*he*, *she*, *they*, *him*, *her*, *them*) receive a $0.7\times$ multiplier. The judge repeatedly flagged sentences such as "my friend always loses things" as NOT_RELEVANT because they describe a third party rather than the author's own experience. The four-tier scheme is summarised in Table 1.

(ii) Negation detection. A negation check was added for the three symptoms the judge identified as systematically broken. Sentences describing the *absence* of a symptom receive a $0.3\times$ penalty: for symptom 12 (leaving one's seat), sentences like "I sat through the entire lecture" are penalised; for symptom 16 (finishing others' sentences), "people always cut me off" is penalised because the author is being interrupted, not interrupting; for symptom 17 (difficulty waiting), "I always wait my turn" is penalised as it describes patience. These pre-LLM penalties eliminate obvious false positives without additional API calls.

(iii) Iterative query expansion (v2). The embedding query strings for six symptoms were rewritten through iterative judge-driven review: symptom 6 (hyperactivity), 10 (losing things), 12, 13 (restlessness),

Table 1
Four-tier pronoun scoring in Run 4. The third-person penalty is new relative to Run 3.

Pronouns present	Multiplier	Rationale
I, I'm, my, me, myself	$1.3\times$	Strong self-report → boost
we, our, us	$1.1\times$	Group experience → small boost
None of the above	$1.0\times$	Neutral
he, she, they, his, her, him, them	$0.7\times$	About someone else → penalise

15 (talking too much), and 17. For each, the judge's NOT_RELEVANT feedback was traced to systematic false positive clusters, and the query was edited to pull the semantic search away from those clusters. For example, symptom 6's query originally contained "active" and "drive", which attracted gym and motivational content; these were replaced with ADHD-specific phrases such as "can't slow down my brain" and "running on overdrive". The 12 remaining symptoms were unchanged, as the judge reported no recurring failures on them.

(iv) Expanded reranker pool. The number of candidates passed to GPT-4o-mini was doubled from 50 to 100 per symptom, and raised to 200 for the three weak symptoms (12, 16, 17). This ensures that correct sentences ranked between positions 50 and 100 by embedding similarity are visible to the reranker.

(v) Per-symptom reranker prompt. The Run 3 prompt used general instructions. Run 4 provides four tailored examples per symptom (scores 1.0, 0.6, 0.2, 0.0) so the LLM's judgement is grounded in what relevance actually means for that specific ADHD behaviour. Coping strategies ("I keep a planner so I don't forget") are now explicitly scored at 0.2, and negation at 0.0, with examples. A token-limit increase and regex fallback were added to handle JSON truncation errors that occasionally crashed Run 3.

(vi) LLM-as-judge evaluation. Run 4 integrates a two-step evaluation loop using GPT-4o, a stronger model than the GPT-4o-mini used for reranking, deliberately chosen to ensure evaluator independence. Step 13 evaluates the top-10 results per symptom, reporting P@10 and nDCG@10. Step 14 extends this to the top-100, reporting P@100 and nDCG@100. This fast feedback loop was what made all the other changes in Run 4 possible: without per-sentence, per-symptom feedback, failure modes could only be identified after official submission.

4. Results and Discussion

4.1. Task 3: ADHD Symptom Sentence Ranking

LLM-as-judge evaluation results. Official eRisk evaluation scores are pending the submission deadline. The internal GPT-4o judge evaluation of Run 4 returned a mean P@10 of 0.9278 and a mean nDCG@10 of 0.9710 across all 18 symptoms. Table 2 shows per-symptom P@10, nDCG@10, and relevant-in-top-10 counts from the judge report. Fifteen of 18 symptoms achieved P@10 = 1.0. The three weak symptoms targeted in Run 4, symptoms 12, 16, and 17, show partial recovery: Symptom 12 reached P@10 = 0.7 with nDCG@10 = 0.9951.

Score distribution in Run 4 submissions. Table 3 shows the top-10 submission scores for Symptom 1 from the Run 4 output file. Scores compress near the upper end: the top five entries range from 0.9737 to 0.7371, with a sharper drop from position 5 to position 6. This is expected from the combined scoring formula (Equation 1): the 0.6 LLM weight pulls top-ranked candidates close together when the reranker assigns uniformly high relevance scores. It reflects genuine multi-sentence relevance, not retrieval error.

What the judge evaluation exposed. Table 4 summarises the two distinct error types the GPT-4o judge identified, neither of which aggregate metrics would have surfaced. Semantic inversion occurs when a sentence describing the *absence* of a symptom scores highly because the relevant vocabulary

Table 2
Task 3 Run 4 internal judge evaluation results per symptom. P@10 and nDCG@10 reported by GPT-4o judge. Symptoms 12, 16, 17 were the targeted weak cases from Run 3.

Symptom	ASRS-v1.1 Description (abbrev.)	P@10	nDCG@10	Relevant/10
1	Finishing final details of a project	1.0000	0.8781	10/10
2	Organising tasks requiring sequencing	1.0000	1.0000	10/10
3	Remembering appointments / obligations	1.0000	1.0000	10/10
4	Avoiding tasks requiring sustained effort	1.0000	0.9547	10/10
5	Fidgeting when seated for long periods	1.0000	0.8957	10/10
6	Feeling overly active / compelled to move	1.0000	1.0000	10/10
7	Making careless mistakes	1.0000	1.0000	10/10
8	Difficulty with sustained attention	1.0000	0.9530	10/10
9	Difficulty concentrating on conversation	1.0000	0.9555	10/10
10	Misplacing things	0.9000	0.9000	9/10
11	Distracted by external stimuli	1.0000	1.0000	10/10
12	Leaving seat in situations requiring stay	0.7000	0.9951	7/10
13	Feeling restless	1.0000	1.0000	10/10
14	Difficulty unwinding / relaxing	1.0000	1.0000	10/10
15	Talking excessively	1.0000	1.0000	10/10
16	Finishing others' sentences	1.0000	1.0000	10/10
17	Difficulty waiting one's turn	1.0000	1.0000	10/10
18	Interrupting others	1.0000	1.0000	10/10
Mean		**0.9278**	**0.9710**	

Table 3
Run 4 top-10 submission scores for Symptom 1 (finishing final details of a project). Score compression at the top reflects LLM-assigned uniform high relevance.

Rank	Sentence ID	System	Score
1	KUZnTb_44_5	SymptomAIQueryExpNegFilterLLM	0.9737
2	BS1vJ3_347_5	SymptomAIQueryExpNegFilterLLM	0.9736
3	9FCyPy_88_0	SymptomAIQueryExpNegFilterLLM	0.9735
4	DCNR2N_113_4	SymptomAIQueryExpNegFilterLLM	0.9703
5	Khg94C_23_6	SymptomAIQueryExpNegFilterLLM	0.7371
6	Khg94C_23_5	SymptomAIQueryExpNegFilterLLM	0.7353
7	39QQmI_741_4	SymptomAIQueryExpNegFilterLLM	0.7352
8	s56wxB_50_3	SymptomAIQueryExpNegFilterLLM	0.7337
9	1tTXdP_430_3	SymptomAIQueryExpNegFilterLLM	0.7329
10	VgbCMs_129_0	SymptomAIQueryExpNegFilterLLM	0.7329

is present; "I sat through the entire lecture" contains sitting-related language that cosine similarity cannot distinguish from active symptom expression. Referential mismatch occurs when third-party descriptions pass all linguistic filters; "my friend always loses things" is correctly excluded by the pronoun tier but would score highly without it. Neither error type is recoverable by reranking alone, because both appear in the top-50 candidate pool that the LLM reranker sees. Pre-LLM corrections address these upstream, which is architecturally more efficient than asking the reranker to overcome retrieval failures.

What remains unresolved. Official MAP and nDCG@100 scores from the eRisk organisers are the primary outstanding validation. The internal judge evaluation is directionally informative but uses the same model family (GPT-4o) as the reranker (GPT-4o-mini), which introduces evaluator–reranker correlation that could inflate precision estimates. The stylometric Run 5 extension, incorporating character n-grams and vocabulary diversity, remains unsubmitted; without it, the pipeline has no

Table 4
Failure modes identified by the GPT-4o judge in Run 3 outputs, with the corrections applied in Run 4.

Failure Type	Example Sentence	Why It Passed Run 3	Run 4 Correction
Semantic inversion (negation)	"I sat through the entire lecture" (Symptom 12)	Relevant vocabulary present; cosine similarity cannot detect negated symptom	$0.3\times$ negation penalty applied pre-LLM
Referential mismatch (third-person)	"My friend always loses things" (Symptom 10)	First-person filter only; third party was not penalised	$0.7\times$ third-person multiplier applied pre-LLM
Symptom-adjacent vocabulary	"active," "drive" attracting gym content (Symptom 6)	Query string contained these terms; embedding similarity pulled off-topic results	Iterative query rewrite; ADHD-specific phrases substituted

signal for the structural writing patterns Barrios et al. found to be diagnostically meaningful for ADHD independently of semantic content.

Official evaluation preliminary results. Table 5 consolidates SymptomAI's four runs under the two relevance-judgment settings used in the official eRisk 2026 evaluation [1]: majority-vote, where a sentence is considered relevant if at least two of three expert assessors agreed, and unanimity, where all three assessors must agree.

Table 5
SymptomAI official scores for all four runs under majority-vote and unanimity relevance judgments. Scores sourced from the eRisk 2026 preliminary results report [1].

		Majority-vote				Unanimity			
Run	**Method (abbrev.)**	**AP**	**R-PREC**	**P@10**	**NDCG**	**AP**	**R-PREC**	**P@10**	**NDCG**
Run 1	Cosine baseline	0.052	0.098	0.194	0.267	0.048	0.072	0.133	0.234
Run 2	Medical SBERT	0.012	0.041	0.056	0.094	0.005	0.015	0.028	0.066
Run 3	FP filter + LLM rerank	0.037	0.056	0.189	0.155	0.030	0.040	0.128	0.126
Run 4	Judge-driven fixes	0.067	0.096	0.300	0.215	0.070	0.093	0.206	0.192

Run progression. Run 4 achieves the best AP and P@10 across all runs under both settings, confirming that the judge-driven corrections, third-person penalty, negation detection, query rewriting, and expanded candidate pool, produced measurable improvement over Run 3. However, Run 1, the MiniLM cosine baseline with no filtering, achieves the highest NDCG under both settings (majority 0.267; unanimity 0.234), outperforming Run 4 (majority 0.215; unanimity 0.192). This divergence exposes a precision–depth tradeoff: Run 4 surfaces more relevant sentences near the top of the ranking (P@10 gain of 0.106 over Run 1 under majority), but the score compression documented in Table 3 reduces discriminability across the full ranking depth, degrading NDCG. Run 2, the medical SBERT configuration using `S-PubMedBERT-MS-MARCO`, is the weakest run across every metric (majority AP 0.012, P@10 0.056), despite being designed as an improvement over Run 1. The likely cause is domain mismatch: `S-PubMedBERT` is trained on formal clinical text and encodes the colloquial, fragmented language of Reddit posts less effectively than the general-domain `all-MiniLM-L6-v2`. Run 3, which applies first-person filtering and LLM reranking on top of Run 2 embeddings, partially recovers from this failure (AP 0.037, P@10 0.189), demonstrating that LLM reranking can compensate for a weak candidate pool

but cannot substitute for retrieval-stage representational quality.

Gap between internal and official evaluation. The internal GPT-4o judge returned a mean P@10 of 0.9278 for Run 4; the official P@10 is 0.300 under majority judgments, a factor of 3.1 difference. This confirms the evaluator–reranker correlation risk identified in Section 4. The reranker (GPT-4o-mini) and the judge (GPT-4o) belong to the same model family and share representational biases. Sentences that GPT-4o-mini ranked highly were systematically more likely to satisfy GPT-4o's relevance criteria, independently of agreement with human expert assessors. This circularity inflated internal precision estimates substantially. Two conclusions follow. First, the judge-driven diagnostic cycle remains a valid method for identifying failure modes across runs: the AP improvement from Run 3 to Run 4 ($+0.030$ majority) is confirmed by official scores. Second, internal LLM-as-judge evaluation cannot serve as a proxy for official metrics when the reranker and evaluator share a model family. Future evaluation loops should use either a structurally different model (e.g. an open-weight model distinct from the API-based reranker) or human spot-checks on a stratified sample of retrieved sentences.

Symptom-level difficulty in the relevance pool. Analysis of the organiser-released relevance judgments reveals substantial variation in symptom retrievability. Symptom 4 (avoiding tasks requiring sustained effort) has the highest pool density, 31.5% of assessed sentences judged relevant under majority, and the highest annotator agreement (76% of majority-relevant sentences retained under unanimity), making it the most tractable retrieval target. Symptom 16 (finishing others' sentences) has the lowest density (4.8%) and the lowest annotator agreement (36% retained), reflecting genuine clinical ambiguity: this behaviour overlaps with normal conversational patterns and is rarely resolved by sentence-level evidence alone. Table 6 summarises the five hardest and five easiest symptoms by pool density. The three symptoms targeted by the judge-driven cycle in Run 4 (12, 16, 17) are among the five lowest-density symptoms in the pool, confirming that the internal judge correctly identified the hardest retrieval targets. However, low pool density also means that precision improvements on these symptoms contribute disproportionately little to aggregate AP, limiting the measurable gain from Run 4's targeted corrections.

Table 6
Five hardest and five easiest ADHD symptoms by relevance pool density (relevant/assessed, majority judgments) and annotator agreement rate (unanimity relevant / majority relevant). Symptoms targeted by Run 4 judge-driven corrections are marked †.

Sym.	ASRS-v1.1 Description (abbrev.)	Pool	Density	Agreement
Five hardest (lowest density)				
16†	Finishing others' sentences	231	0.048	0.36
12†	Leaving seat when required to stay	294	0.051	0.53
6	Feeling overly active / compelled to move	331	0.066	0.36
7	Making careless mistakes	337	0.068	0.48
17†	Difficulty waiting one's turn	298	0.070	0.52
Five easiest (highest density)				
4	Avoiding tasks requiring sustained effort	321	0.315	0.76
13	Feeling restless	262	0.275	0.83
14	Difficulty unwinding / relaxing	304	0.227	0.75
2	Organising tasks requiring sequencing	306	0.193	0.75
8	Difficulty with sustained attention	303	0.185	0.38

A secondary pattern in Table 6 is that annotator agreement does not track directly with pool density. Symptom 8 (difficulty with sustained attention) has the fifth-highest density (18.5%) but the second-lowest annotator agreement (38%), suggesting that sustained attention failures are frequently expressed in social media text but that assessors disagree substantially on which sentences constitute clinically informative evidence. This symptom-level annotator disagreement is a task-level challenge that retrieval systems cannot resolve: the supervision signal itself is uncertain for a non-trivial subset of symptoms.

Competitive positioning. Across 45 submitted runs from 12 teams, SymptomAI Run 4 ranks approximately 21st by AP (0.067) and 14th by P@10 (0.300) under majority judgments. The leading runs, NeuroRankUB (AP 0.248, P@10 0.528) and erisk-cedri EnsembleV3 (AP 0.190, P@10 0.400), both use ensemble retrieval combining multiple dense retrievers with reciprocal rank fusion prior to any reranking stage. SymptomAI's architecture used a single embedding model per run, which constrains recall at the candidate generation stage. LLM reranking can improve precision within a candidate pool but cannot recover sentences the base retriever excluded. A retrieval-stage ensemble with model diversity across general and clinical embedding spaces would address this recall bottleneck directly and is the most consequential architectural change for a future iteration.

5. Conclusion

This paper presented SymptomAI's approach to eRisk 2026 Task 3. A layered retrieval pipeline progresses from cosine similarity through domain-adapted embeddings, query expansion, and LLM reranking, with first-person linguistic signals boosting self-reported symptom expressions. Run 4 extends this with a judge-driven diagnostic cycle: a GPT-4o evaluator applied after Run 3 identified failure modes on three specific symptoms, which were corrected through a third-person penalty, targeted negation detection, rewritten query strings, an expanded reranker pool, and per-symptom prompting. This evaluation loop demonstrated that embedding-level retrieval has systematic blind spots, opposite-meaning sentences and third-party descriptions, that pre-LLM filters address more efficiently than additional reranking passes.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for grammar and spelling checking.

References

[1] 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 the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, 21–24 September, 2026, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2026.

[2] 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 - 17th International Conference of the CLEF Association, CLEF 2026, Jena, Germany, September 21–24, 2026, Proceedings, volume To be published of *Lecture Notes in Computer Science*, Springer, 2026.

[3] L. Mendoza, J. Suarez, E. Puertas, J. Martinez, J. Serrano, Cotecmar-utb at erisk 2025: Semantic-centroid symptom ranking and early depression detection using adaptive decision rule, in: Notebook Papers of CLEF 2025 eRisk Lab, CEUR Workshop Proceedings, 2025. URL: https://ceur-ws.org/Vol-4038/paper_126.pdf.

[4] N. Munz, E. Aharon, A. Segal, K. Gal, Semantic retrieval of bdi symptoms in user writings, in: Notebook Papers of CLEF 2025 eRisk Lab, CEUR Workshop Proceedings, 2025. URL: https://ceur-ws.org/Vol-4038/paper_129.pdf.

[5] J. Barrios, E. Poznyak, J. L. Samson, H. Rafi, S. Gabay, F. Cafiero, M. Debbané, Detecting adhd through natural language processing and stylometric analysis of adolescent narratives, Frontiers in Child and Adolescent Psychiatry 4 (2025). URL: https://www.frontiersin.org/journals/child-and-adolescent-psychiatry/articles/10.3389/frcha.2025.1519753/full. doi:10.3389/frcha.2025.1519753.

[6] Y. Zhu, Y. Guo, N. Marchuck, Y. Wang, A. Sarker, Leveraging large language models and traditional machine learning ensembles for adhd detection from narrative transcripts, 2025. URL: https://arxiv.org/abs/2505.21324. doi:10.48550/arXiv.2505.21324. arXiv:2505.21324.