NeuroRankUB at eRisk 2026: Embedding-Based Retrieval and Re-ranking for ADHD Symptom Search

eRisk at CLEF 2026

Teodora Berca[1,*], Daria Horga[1,*], Raluca Negoiță-Crețu[1,*] and Antonia-Maria Popeangă[1,*]

[1]*Faculty of Mathematics and Computer Science, University of Bucharest, Bucharest, Romania*

Abstract

Attention-Deficit/Hyperactivity Disorder (ADHD) is a neurodevelopmental condition whose symptoms may surface in online written discourse. Natural Language Processing (NLP) can help identify ADHD-related signals in text, supporting early risk assessment and mental health research. In these working notes, we describe our participation in the ADHD symptom sentence retrieval task at the eRisk shared lab, where systems were required to retrieve relevant sentences for the 18 symptoms of the Adult ADHD Self-Report Scale (ASRS-v1.1) from a corpus of over 4.1 million sentences.

As no labeled training data were provided in this first edition of the task, we explored retrieval pipelines combining sentence embeddings and cosine similarity for candidate selection, followed by re-ranking with cross-encoders and Large Language Models (LLMs). Our submitted NeuroRankUB systems achieved strong performance, with the best run (*final ADHD ranking*) ranking first under both majority-vote and unanimity relevance judgements, while three of our submissions occupied the top three positions under unanimity evaluation.

Keywords

ADHD, eRisk, information retrieval, sentence retrieval, sentence embeddings, cross-encoder re-ranking, large language models, natural language processing

1. Introduction

Attention-Deficit/Hyperactivity Disorder (ADHD) represents a significant and growing global health concern. Between 1990 and 2021, the global prevalence of ADHD increased by approximately 18.7%, while associated disability-adjusted life years (DALYs) increased by 18.6%, with the highest burden observed among children and adolescents [1]. As ADHD frequently persists into adulthood and is associated with substantial socio-economic impact, there is increasing interest in computational approaches that can support ADHD-related research and mental health assessment.

The increasing availability of publicly shared textual content on online platforms has created new opportunities for Natural Language Processing (NLP) methods to analyze behavioral and mental health indicators through language. Large-scale textual corpora derived from online interactions may provide valuable evidence for identifying symptom-related expressions, enabling the study of mental health signals without relying exclusively on manually collected datasets.

This motivation is reflected in Task 3 of the eRisk Lab at CLEF 2026, which focuses on ADHD symptom sentence retrieval [2, 3]. Given a large textual collection containing more than four million sentences, participants were tasked with retrieving sentences relevant to the 18 symptoms defined in the Adult ADHD Self-Report Scale (ASRS-v1.1) questionnaire [4]. Unlike traditional keyword retrieval, the challenge required systems to capture the broader semantic meaning of symptom descriptions, as relevant expressions may be contextual, or phrased differently from the formal symptom definitions.

Several characteristics make this task particularly challenging. First, no labeled training data were provided during system development, requiring participants to rely on unsupervised or weakly supervised retrieval strategies. Second, the large scale of the corpus introduces substantial computational con-

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

*Corresponding author.

teodora.berca@s.unibuc.ro (T. Berca); daria.horga@s.unibuc.ro (D. Horga); raluca-marina.negoita-cretu@s.unibuc.ro (R. Negoiță-Crețu); antonia-maria.popeanga@s.unibuc.ro (A. Popeangă)

straints, making exhaustive manual inspection infeasible and motivating efficient filtering approaches. Third, the colloquial and informal nature of social media language creates a semantic gap between symptom descriptions derived from standardized clinical questionnaires and naturally occurring online expressions. Additionally, the overlapping nature of ADHD symptomatology further complicates symptom-specific retrieval, as many behavioral expressions may be relevant to multiple symptoms simultaneously. In this work, we present the NeuroRankUB systems submitted to Task 3 at eRisk 2026. Our approaches explore multiple retrieval pipelines combining sentence embeddings for candidate retrieval, cross-encoder models for re-ranking, and Large Language Models (LLMs) for refining challenging candidate sentences.

2. Related Work

Research on the use of Natural Language Processing (NLP) for ADHD-related assessment remains comparatively limited relative to other mental health conditions such as depression. Barrios et al. [5] explored ADHD detection through NLP and stylometric analysis of adolescent narratives. Their study analyzed self-defining memory narratives written by adolescents diagnosed with ADHD and control participants, extracting stylometric features such as word count, lexical diversity, cohesion, and linguistic density to identify behavioral differences between groups. Their findings demonstrate the potential of language-based analysis for identifying ADHD-related patterns. However, their experimental setting relied on prompted narratives collected under controlled conditions, whereas the eRisk Task 3 corpus consists of highly informal, noisy social media content that does not follow a predefined prompt, making stylometric analysis substantially more challenging.

Recent advances in Large Language Models (LLMs) have further expanded the potential of computational approaches for ADHD-related assessment. Zhu et al. [6] investigated the use of LLMs and traditional machine learning techniques for ADHD detection from narrative transcripts. Their approach combined multiple language models with a Support Vector Machine (SVM) classifier through ensemble voting to determine whether a transcript exhibited ADHD-related behavioral patterns, highlighting the effectiveness of combining contextual language understanding with traditional predictive methods.

Closer to our retrieval setting, the REBECCA team at eRisk 2024 proposed a system for depression symptom retrieval based on sentence embeddings and prompt-based filtering [7]. Their methodology addressed a task structurally similar to ADHD symptom retrieval, requiring the identification of symptom-relevant evidence from large textual collections. In particular, their filtering strategy based on first-person pronoun usage motivated part of our initial candidate selection process.

3. Methodology

3.1. Dataset and Preprocessing

Participants were provided with a `.trec` file containing 4,170,875 sentences extracted from social media posts and comments. As the corpus included multilingual content, an initial language identification step was performed to ensure linguistic consistency across the retrieval process. We employed the FastText language identification model (`lid.176.ftz`) for automatic language detection [8].

To standardize the corpus, non-English sentences were translated into English using the multilingual machine translation model `Helsinki-NLP/opus-mt-mul-en` [9]. Following language detection, the corpus was found to consist predominantly of English content, with the most frequent non-English languages including French, Spanish, German, and Italian, as summarized in Table 1.

Additionally, the corpus contained duplicate entries. After removing duplicate sentences, the dataset was reduced to 3,923,739 unique sentences, which served as the basis for subsequent retrieval and ranking experiments.

Table 1
Top detected languages in the eRisk ADHD corpus.

Language	Number of Sentences
English (en)	4,023,643
French (fr)	23,742
Spanish (es)	20,463
German (de)	14,170
Italian (it)	12,888

3.2. Candidate Sentence Filtering

Given the large scale of the corpus, an initial filtering stage was applied to reduce the search space prior to retrieval and re-ranking. The objective was to remove noisy or low-information content while preserving potentially relevant symptom-related evidence.

Following insights from the REBECCA system at eRisk 2024 [7], we first applied a first-person pronoun filter, motivated by the self-report nature of ADHD symptoms in the ASRS-v1.1 questionnaire. We retained only sentences containing first-person pronouns (e.g., *I*, *me*, *my*) and subsequently removed hyperlinks, which were often associated with low-information or external content. Sentence length filtering was applied to exclude excessively short or long sentences, while duplicate removal and alphanumeric filtering further reduced noisy social media content.

Table 2
Corpus size after each filtering step.

Filtering Step	Number of Sentences
Initial corpus	4,170,875
First-person filter	1,642,543
Link removal	1,614,693
Length filter	1,535,170
Duplicate removal	1,500,871
Alphanumeric filtering	1,498,792

We submitted a total of four runs. A mapping between submission file names and methods is provided in Appendix A, while the following subsections describe each approach in detail.

3.3. MiniLM Semantic Filtering with Cross-Encoder Re-ranking and LLM Refinement

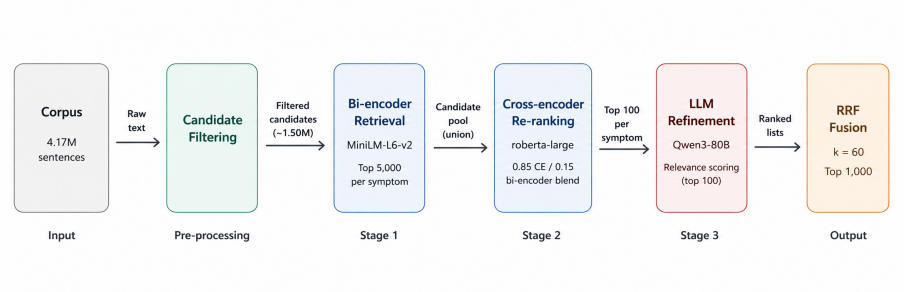

Figure 1: Overview of the retrieval pipeline.

This approach followed a multi-stage pipeline combining semantic candidate filtering, cross-encoder re-ranking, and LLM-based refinement. Following preprocessing and rule-based filtering, sentence representations were generated using the `all-MiniLM-L6-v2` sentence transformer model [10]. To

better align formal ASRS-v1.1 symptom descriptions with informal social media language, each symptom was enriched with multiple first-person reformulations (see Table 6).

Cosine similarity was computed between sentence embeddings and symptom reformulations, and for each symptom the top 5,000 candidates were retained based on their maximum similarity score. The union of these symptom-specific candidate sets formed a shared candidate pool for ranking.

The selected candidates were subsequently re-ranked using the `stsb-roberta-large` cross-encoder model [11]. Each symptom was associated with an enriched query derived from the original ASRS-v1.1 questionnaire item and adapted with the assistance of ChatGPT to better reflect natural ADHD-related expressions (see Appendix B.2). Candidate sentences were jointly scored against the enriched query and its reformulations, with the final cross-encoder score computed as the mean score across all query variants. This score was combined with the original bi-encoder similarity score using a weighted scheme (0.85 cross-encoder, 0.15 bi-encoder).

Finally, the top 100 candidates for each symptom were refined using the `Qwen3-Next-80B-A3B-Instruct` model [12], prompted to assign numerical relevance scores to symptom–sentence pairs under a constrained scoring format (see Appendix B.3). The final ranking was generated using Reciprocal Rank Fusion (RRF), combining the ranking outputs of the cross-encoder and LLM refinement stages using the smoothing constant $k = 60$.

3.4. Dense Retrieval with BM25, Cross-Encoder Re-ranking, and LLM Scoring

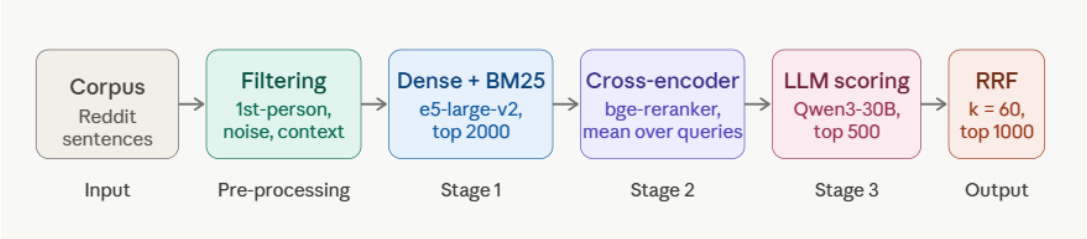

Figure 2: Overview of the retrieval pipeline.

The proposed system employed a multi-stage retrieval pipeline combining dense retrieval, BM25 lexical retrieval, neural reranking, and LLM-based refinement.

Following preprocessing and rule-based filtering, sentence embeddings were generated using the `intfloat/e5-large-v2` model. Cosine similarity was computed between symptom reformulations and sentence embeddings, and the top 2,000 candidates were retrieved for each symptom. In parallel, BM25 retrieval was applied over the enriched sentence corpus in order to preserve exact lexical matches that may not be captured by dense retrieval alone.

The retrieved candidates were subsequently reranked using the `BAAI/bge-reranker-large` cross-encoder model. Final cross-encoder scores were computed as the mean score across all symptom reformulations.

To combine dense retrieval, BM25, and cross-encoder rankings, the system employed Reciprocal Rank Fusion (RRF).

$$\mathrm{RRF}(d) = 0.35 \cdot \frac{1}{k + r_{dense}(d)} + 0.50 \cdot \frac{1}{k + r_{cross}(d)} + 0.15 \cdot \frac{1}{k + r_{bm25}(d)}$$

where $r_{dense}(d)$, $r_{cross}(d)$, and $r_{bm25}(d)$ denote the ranking positions assigned by the dense retriever, cross-encoder, and BM25 retriever, respectively. The smoothing parameter was set to $k = 60$.

Finally, the top-ranked candidates were refined using the `qwen/qwen3-30b-a3b` LLM, prompted to assign numerical relevance scores to symptom–sentence pairs based on ADHD symptom definitions and examples.

3.5. GTE-Based Dense Retrieval with Query Reformulation, Rank Fusion, and LLM Re-ranking

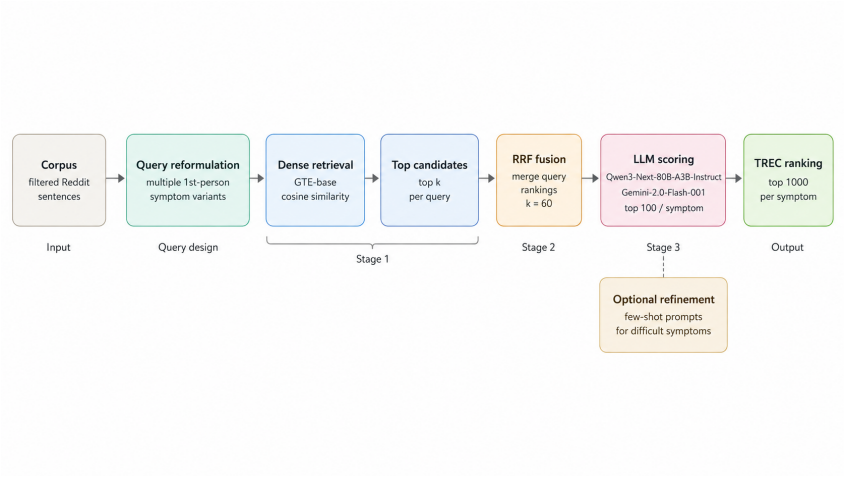

Figure 3: Overview of the GTE-based query reformulation, fusion, and reranking pipeline.

The proposed system followed a multi-stage dense retrieval pipeline combining symptom-specific query reformulation, dense semantic retrieval, rank fusion, and Large Language Model (LLM)-based re-ranking.

For each of the 18 ADHD symptoms, we manually constructed multiple first-person natural language reformulations of the original ASRS-v1.1 symptom description. These reformulations were designed to better match the informal and highly variable language used in Reddit posts while preserving the clinical meaning of each symptom.

After preprocessing and rule-based filtering, sentence embeddings were generated using the `thenlper/gte-base` sentence transformer model. Cosine similarity was then computed between each symptom reformulation and the candidate sentence corpus, producing multiple ranked lists for the same symptom. For each reformulation, the highest-ranked candidates were retained for subsequent fusion.

To aggregate evidence across reformulations, we applied Reciprocal Rank Fusion (RRF). This step promoted sentences that were retrieved consistently by several semantically related reformulations, thereby reducing sensitivity to any single query wording.

Finally, the top-ranked candidates from the fused retrieval stage were further refined using an LLM-based scoring step. We used symptom-specific prompts consisting of a short symptom definition together with positive and negative relevance criteria. For more difficult symptoms, we additionally tested prompt refinement and few-shot examples in order to better separate closely related behaviors. Depending on the symptom-specific configuration, this stage was performed using `Qwen3-Next-80B-A3B-Instruct` or `Gemini-2.0-Flash-001`. The resulting LLM scores were used to refine the highest-ranked candidates, and the final system output was a ranked list of the top 1,000 sentences for each symptom. Detailed prompt templates and examples are provided in Appendix C.

3.6. All-Mpnet-Base-V2 Semantic Filtering with Cross-Encoder Re-ranking

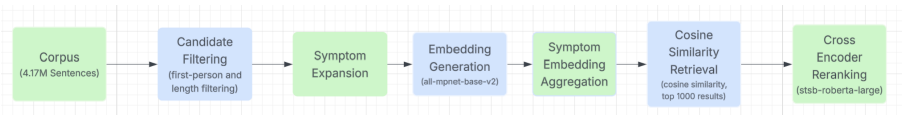

Figure 4: Overview of the retrieval pipeline.

This approach is based on a semantic retrieval and reranking pipeline. The architecture contains five major stages: symptom expansion, sentence embedding generation, symptom prototype construction, cosine similarity retrieval and cross-encoder reranking.

Each of the 18 ADHD symptoms was expanded into 12 representative sentences, resulting in a symptom dataset containing 216 sentences in total (see Table 7). These representative sentences were generated with ChatGPT and were designed to reflect natural Reddit style communication and informal self-expression patterns commonly found in online discussions. The main objectives of this expansion step are: improving semantic coverage, reducing wording bias and capturing the variability of natural language expressions used by Reddit users.

By introducing multiple paraphrased formulations for each symptom, the system becomes more robust to linguistic variation and better able to retrieve semantically related content.

Sentence embeddings were generated using the pretrained all-mpnet-base-v2 model from Sentence Transformers [13, 14]. After embedding generation, all vectors were L2-normalized to ensure that cosine similarity is equivalent to dot-product similarity.

For each symptom, a prototype embedding was constructed by grouping together the embeddings of its 12 representative sentences and applying mean pooling across them. This process produces a centroid embedding that represents the semantic space associated with that symptom. The resulting centroid vectors were normalized again after pooling.

Finally, all symptom prototype embeddings were stacked into a matrix where each row corresponds to a single ADHD symptom representation.

The retrieval stage computes semantic similarity between Reddit sentence embeddings and the symptom prototype embeddings using cosine similarity. Reddit embeddings were precomputed using the same embedding model to maintain consistency within the semantic space.

For each symptom, the top 1000 Reddit sentences with the highest cosine similarity scores were retrieved. This stage serves as an efficient dense retrieval mechanism that generates candidate sentences for the reranking phase.

To improve ranking precision, a cross-encoder reranking stage was applied to the retrieved candidates. The reranking model used was `stsb-roberta-large` cross-encoder model [11].

Unlike bi-encoder retrieval, the cross-encoder jointly processes the symptom query and candidate sentence, allowing deeper semantic interaction between the two inputs. The reranker produces refined relevance scores that are used to generate the final ranking of Reddit sentences for each ADHD symptom.

4. Results

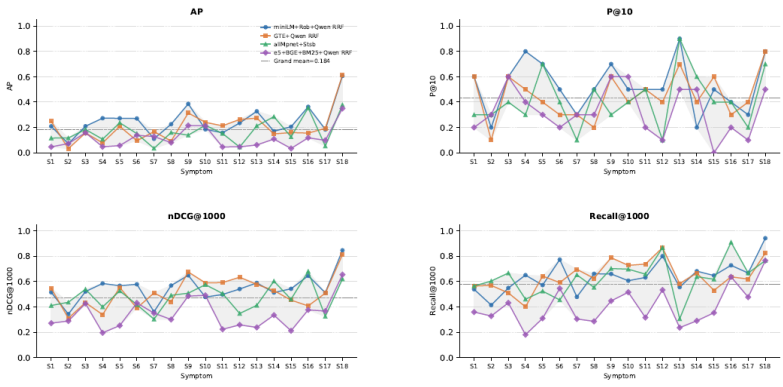

Figure 5: Comparison of all four runs across ADHD symptoms.

We submitted four runs based on different retrieval strategies in the `.trec` format. As this was the first edition of the eRisk ADHD retrieval task, relevance judgements were obtained through expert

annotation of pooled candidate sentences generated across all submissions. Results were evaluated under both majority-vote and unanimity agreement schemes using AP, R-PREC, P@10, and NDCG@1000.

Table 3
Official shared task results for NeuroRankUB submissions, ranked by AP (majority voting)

Run	AP	R-PREC	P@10	NDCG@1000
final ADHD ranking	**0.248**	**0.328**	**0.528**	**0.547**
gte rrf final	0.203	0.264	0.450	0.516
ADHD submission	0.171	0.242	0.417	0.475
final ranking	0.092	0.185	0.283	0.312

Table 4
Official shared task results for NeuroRankUB submissions ranked by AP (unanimity voting)

Run	AP	R-PREC	P@10	NDCG@1000
final ADHD ranking	**0.207**	**0.274**	**0.367**	**0.488**
gte rrf final	0.184	0.233	0.306	0.466
ADHD submission	0.166	0.232	0.311	0.439
final ranking	0.073	0.117	0.183	0.268

Figure 5 shows a stable ranking across metrics, including Recall@1000, with *MiniLM+RoBERTa+Qwen RRF* consistently outperforming other systems. While *GTE+Qwen RRF* achieved comparable Recall@1000 (Figure 5), its lower precision-based scores suggest weaker ranking effectiveness. Performance varied substantially across symptoms: S2, S7, and S17 remained challenging for all runs, whereas S18 achieved markedly stronger results, likely due to more explicit linguistic cues. The similarity in Recall@1000 across systems suggests that performance differences are driven primarily by re-ranking quality rather than candidate retrieval.

5. Conclusions

In this work, we presented the NeuroRankUB systems for ADHD symptom sentence retrieval at eRisk 2026. Our best-performing pipeline combining bi-encoder retrieval, cross-encoder re-ranking, and LLM refinement via RRF, ranked first under both majority and unanimity relevance judgements, with three of our four submissions occupying the top three positions under unanimity evaluation.

Per-symptom analysis revealed that behaviourally explicit symptoms such as S18 (*interrupting others*) are substantially easier to retrieve than symptoms with lexically ambiguous query concepts, pointing to contextual query enrichment and symptom-specific negative supervision as promising directions for future work.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools, including ChatGPT (OpenAI), to support language refinement, including grammar and spelling correction, sentence rephrasing, and the reformulation of prompts and symptom descriptions used in the retrieval and re-ranking stages of the methodology. These tools were also used to improve the clarity and readability of the manuscript.

All AI-assisted outputs were carefully reviewed, validated, and edited by the authors as necessary. The authors take full responsibility for the content, methodology, interpretations, and conclusions presented in this work.

References

[1] C. Wang, L. Hou, H. Zhou, Q. Wang, H. Yan, Y. Lang, J. Tan, Y. Liu, Y. Ge, Y. Lin, C. Fu, J. Liu, The evolving global burden of adhd: A comprehensive analysis and future projections (1990–2046), Journal of Affective Disorders 391 (2025) 120037. URL: https://www.sciencedirect.com/science/article/pii/S016503272501479X. doi:10.1016/j.jad.2025.120037.

[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 (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.

[3] 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.

[4] R. C. Kessler, L. Adler, M. Ames, et al., The world health organization adult adhd self-report scale (asrs): A short screening scale for use in the general population, Psychological Medicine 35 (2005) 245–256.

[5] J. Barrios, E. Poznyak, J. Lee 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) 1519753. doi:10.3389/frcha.2025.1519753.

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

[7] A. Barachanou, F. Tsalakanidou, S. Papadopoulos, Rebecca at erisk 2024: Search for symptoms of depression using sentence embeddings and prompt-based filtering, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, CLEF 2024 Labs and Workshops, Notebook Papers, 2024. Notebook for the eRisk Lab at CLEF 2024.

[8] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL) (2017) 427–431.

[9] J. Tiedemann, S. Thottingal, Opus-mt — building open translation services for the world, in: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, 2020, pp. 479–480.

[10] 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, 2019, pp. 3982–3992.

[11] R. Nogueira, K. Cho, Passage re-ranking with bert, arXiv preprint arXiv:1901.04085 (2019).

[12] Q. Team, Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).

[13] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 16857–16867.

[14] Sentence Transformers, all-mpnet-base-v2, https://huggingface.co/sentence-transformers/all-mpnet-base-v2, 2021. Hugging Face model card.

A. Submitted runs

Table 5
Submitted runs and corresponding methods.

Run	Method
`final ADHD ranking`	MiniLM + RoBERTa + Qwen + RRF
`gte rrf final`	GTE-base + RRF + Qwen/Gemini
`ADHD submission`	E5-large-v2 + BM25 + BGE reranker + Qwen
`final ranking`	MPNet + RoBERTa reranking

B. Retrieval and Re-ranking Details

B.1. ADHD Symptom Reformulations

Table 6
Examples of symptom reformulations used for retrieval.

ASRS Symptom	Representative Reformulations
Difficulty sustaining attention	• "I can't focus for long" • "I get distracted easily" • "My mind wanders constantly" • "I struggle to pay attention"
Forgetfulness	• "I keep forgetting things" • "I forget appointments" • "I'm always losing stuff" • "I forget what I was supposed to do"
Task avoidance	• "I procrastinate everything" • "I avoid difficult tasks" • "I put things off until the last minute"

Table 7
Examples of ChatGPT-generated symptom reformulations used for semantic retrieval.

ASRS Symptom	Representative Reformulations
Difficulty wrapping up the final details of a project	• "I struggle to finish the final details of a project" • "I leave tasks almost done but don't complete the last steps" • "I lose motivation when only small finishing tasks remain" • "I have trouble wrapping things up at the end of a project"
Difficulty organizing tasks	• "I have difficulty organizing tasks that require planning" • "I struggle to put tasks in a logical order" • "I don't know how to start organizing complex tasks" • "I find it hard to break tasks into clear steps"
Forgetfulness	• "I often forget appointments I have scheduled" • "I forget obligations or commitments I need to attend" • "I miss meetings because I forget about them"

B.2. Enriched Retrieval Queries

Table 8
Representative enriched symptom queries used for retrieval and re-ranking.

Symptom Type	Enriched Query Example
Task completion difficulty	"I can do the difficult and interesting parts of a project, but when only the final details are left my motivation drops and I can't make myself finish."
Organization difficulties	"I know what needs to be done, but I can't organize the steps in my head or decide what order to do things in, which leaves me overwhelmed."
Forgetfulness	"I genuinely forget appointments or commitments unless I set multiple reminders or alarms, and even then I still miss things sometimes."
Distractibility	"Any background noise, movement, or activity around me instantly pulls my attention away from what I'm doing and I can't get back on track."
Impulsivity	"I automatically finish other people's sentences before they can finish them. I do it without thinking, and even when I notice it bothers others, I struggle to stop myself."
Interrupting others	"I interrupt people even when they're busy because I feel like I'll lose my thought if I don't say it immediately."

B.3. Qwen Re-ranking Prompt

The model was instructed to act as a strict scoring engine and assign a relevance score between 0 and 10 to each candidate sentence with respect to a target ADHD symptom.

```
You are a strict scoring engine.

Task:
Score each sentence from 0 to 10.

Symptom:
[symptom]

CRITICAL RULES:
- Output ONLY numbers
- EXACTLY N scores
- No explanations
- No extra text

Return format example:
[1,2,3,...]

Sentences:
1. sentence_1
2. sentence_2
...
```

C. LLM Prompting Templates and Reranking Setup

This section of the appendix provides the prompt template used in the LLM-based reranking stage, together with examples of symptom-specific prompt refinements for difficult symptoms.

C.1. Generic Prompt Template

The LLM reranker was instructed to assign numerical relevance scores to sentence candidates for a given ADHD symptom. The generic prompt template is shown below.

```
Rank the candidate sentences by how well they express a first-person
personal experience matching this ADHD symptom: [SYMPTOM DESCRIPTION].
```

```
Prefer direct self-reports of personal experience.

[POSITIVE INSTRUCTIONS]

Do not prefer advice, definitions, questions, third-person statements,
or abstract discussion.

[NEGATIVE INSTRUCTIONS]

[OPTIONAL FEW-SHOT EXAMPLES]

Give each candidate a score from 0 to 100, where 100 is an excellent
match and 0 is irrelevant.

Return ONLY valid JSON in this exact format:
{"scores":[{"idx":0,"score":87,"reason":"short reason"}, ...]}

The score MUST be a number, not text.
Keep each reason very short.

Candidates:
[0] candidate sentence 1
[1] candidate sentence 2
...
```

C.2. Prompting Components

Each symptom-specific prompt was composed of the following elements:

- a short symptom description derived from the ASRS-v1.1 questionnaire,
- positive relevance instructions describing the desired evidence,
- negative relevance instructions suppressing semantically related but irrelevant matches,
- and, for selected difficult symptoms, few-shot positive and negative examples.

C.3. Example: Symptom 12

For Symptom 12 (*getting up or moving around when expected to remain seated*), we used the following refinement:

```
Symptom description:
getting up or moving around when expected to remain seated,
such as in class or meetings

Positive instruction:
Prefer sentences where the author describes needing to get up or move
while expected to stay seated in class or meetings.

Negative instruction:
Do not prefer literal references to seats, vehicles, furniture,
or physical objects.

Few-shot examples:
Relevant example:
"i'm the only one in class who moves around in their chair,
takes a thousand sips of water, and stares at the clock"
```

```
Not relevant:
"i find myself pacing back and forth just to read a recipe"

Not relevant:
"stand up on the train instead of sitting even when there's
plenty of seating available"
```

C.4. Example: Symptom 16

For Symptom 16 (*finishing other people's sentences before they finish speaking*), we used a more constrained prompt in order to distinguish this behavior from generic interrupting or from cases where other people interrupted the speaker.

```
Symptom description:
finishing other people's sentences before they finish speaking

Positive instruction:
Prefer sentences where the author clearly says that they themselves
say the end of another person's sentence before that person finishes speaking.

Negative instruction:
Do not prefer writing, grammar, sentence structure, or finishing one's
own sentence.
Do not prefer sentences where someone else finishes the author's sentence.
Do not prefer sentences primarily about interrupting without mention
of completing what the other person was about to say.
Do not prefer third-person descriptions where the subject is not the author.
Do not prefer vague references like "i do it" or "it happens" without
clear context about finishing sentences.

Few-shot examples:
Relevant example:
"sometimes i get deja vu while i'm having a conversation with somebody
and say the words they were going to say before them"

Not relevant:
"i interrupt people and make them go at my pace"

Not relevant:
"oftentimes, she responds to what i'm saying before i can finish saying it"
```

C.5. Reranking Setup

Dense retrieval was first performed by computing cosine similarity between each symptom reformulation and the sentence corpus encoded with `thenlper/gte-base`. These ranked lists were merged using Reciprocal Rank Fusion.

The LLM reranking stage was then applied only to the highest-ranked candidates rather than to the entire ranking. In the final multi-stage pipeline, we retained the top 300 candidates for standard symptoms and the top 400 candidates for more difficult symptoms after fusion. A shortlist of 100 candidates per symptom was then constructed and reranked by the LLM in batches, with an additional refinement step applied to the highest-ranked subset.

The final submission, however, consisted of the top 1,000 ranked sentences for each symptom. In this final ranking, the highest positions were determined by the LLM-refined candidates, while the remaining positions were filled from the fused retrieval ranking.