INSA-Lyon at eRisk 2026 Task 3: CADRE, a Five-Run Pipeline for ADHD Symptom Sentence Ranking

Notebook for the eRisk Lab at CLEF 2026

Diana Nurbakova[1,*]

[1]INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69100 Villeurbanne, France

Abstract

Attention-deficit/hyperactivity disorder (ADHD) is a neurodevelopmental disorder whose symptoms surface in everyday self-description rather than through diagnostic vocabulary, making sentence-level evidence detection both clinically motivated and linguistically challenging. eRisk 2026 Task 3 introduces the first shared task on ADHD symptom sentence ranking: given the 18 symptoms of the Adult ADHD Self-Report Scale (ASRS-v1.1) and a corpus of approximately 4.17 million Reddit sentences, systems must rank up to 1,000 sentences per symptom by relevance, with no training data and a relevance definition that is polarity-agnostic. We present CADRE, INSA-Lyon's submission to eRisk 2026 Task 3. CADRE is a three-stage pipeline combining bi-encoder retrieval (`all-mpnet-base-v2` with query expansion and a first-person filter), LLM scoring with `Llama-3.1-8B-Instruct` guided by four-layer clinical symptom definitions, and a trained symptom-conditioned bi-encoder ensemble (MentalRoBERTa, ClinicalBERT, and `all-mpnet-base-v2`) optimised with a composite loss combining ordinal calibration, pairwise ranking, and hierarchy regularisation. Our best of five submitted runs reaches MAP 0.159 under majority qrels (rank 7 of 45 runs) and 0.134 under unanimity (rank 5 of 45). Per-symptom analysis shows that four of the five runs win at least one symptom substantively, with the LLM cascade concentrating its wins in the ASRS Verbal Hyperactivity-Impulsivity factor. We report two negative findings: a plain retrieval-only bi-encoder beats the trained reranker by 6 to 3 substantive wins, and three symptoms (approximately 14% of relevant documents) are missed entirely at the candidate stage. Candidate-stage recall and learnt symptom-routed combination of the runs (whose per-symptom oracle leaves +0.032 to +0.038 of MAP headroom) are the priority directions for future work.

Keywords

ADHD, symptom sentence ranking, ASRS, LLM-based relevance scoring, bi-encoder reranking, mental health information retrieval

1. Introduction

Attention-deficit/hyperactivity disorder (ADHD) is a neurodevelopmental disorder characterised by persistent inattention and/or hyperactivity-impulsivity that interferes with daily functioning. Under the criteria of DSM-5-TR (the Diagnostic and Statistical Manual of Mental Disorders, fifth edition, Text Revision), adult diagnosis requires at least five symptoms in either dimension, persisting for six months and present across two or more settings, with clear functional impairment [1]. In adults, ADHD is routinely screened with the Adult ADHD Self-Report Scale (ASRS-v1.1), an 18-item self-report instrument developed by the World Health Organisation in collaboration with Harvard, whose 6-item Part A serves as the validated screener (positive at four or more "shaded" responses, equivalently a Part A sum of at least 14 out of 24) [2]. Mind that the ASRS is a screener rather than a diagnostic instrument: a positive screen warrants further evaluation, not a diagnosis. The instrument also does not assign a severity rating. The mild / moderate / severe specifiers in DSM-5-TR are applied clinically on the basis of symptom count and functional impairment, and are not derived from the ASRS score.

eRisk 2026 Task 3 is the first shared task on ADHD symptom sentence ranking [3, 4]. For each of the 18 ASRS symptoms, the task is to rank up to 1000 candidate sentences from a Reddit test corpus by relevance. The aim is not to assign a diagnosis or a severity band. Rather, it is to surface sentence-level

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

*Corresponding author.

diana.nurbakova@insa-lyon.fr (D. Nurbakova)

0000-0002-6620-7771 (D. Nurbakova)

evidence informative about the writer's state with respect to the symptom. Three properties make this hard. First, ADHD symptoms tend to surface indirectly in everyday self-description rather than through surface keywords [5]: a writer who describes recurring difficulties consistent with the ASRS items themselves (misplacing items, leaving small tasks unfinished) may give stronger evidence than one who uses the term "ADHD" explicitly. Second, the organisers' relevance definition is polarity-agnostic: a sentence in which the writer denies a symptom is no less informative about their state than one affirming it, which rules out a pure presence-detection approach. Third, no training data is released for this edition. Together with the polarity-agnostic definition, this rules out conventional supervised fine-tuning on symptom-labelled data.

We participated in eRisk 2026 Task 3 as INSA-Lyon with CADRE (Clinically-grounded ADHD-symptom Detection and Retrieval Ensemble), a three-stage pipeline. Stage 1 retrieves candidate sentences for each symptom with a bi-encoder, augmented by query expansion, a first-person filter, and a keyword boost. Stage 2 scores each candidate with a large language model (LLM) using a structured output schema and four-layer clinical definitions of the symptom. Stage 3 reranks the resulting silver labels with an ensemble of three symptom-conditioned bi-encoder backbones (MentalRoBERTa, ClinicalBERT, and `all-mpnet-base-v2`), trained with a composite loss combining ordinal calibration, pairwise ranking, and a hierarchy regulariser. Five complementary runs were submitted: three trace the main flow at successive depths (Stage 1 alone, Stages 1–2, Stages 1–2–3), one fuses two of them by reciprocal rank fusion, and one substitutes a depression-transfer baseline.

In this paper, we present and evaluate CADRE in the eRisk 2026 Task 3 shared-task setting. Our contributions are threefold. First, we describe a five-run pipeline combining bi-encoder retrieval, LLM scoring grounded in four-layer clinical definitions, and a trained bi-encoder ensemble reranker. Second, a per-symptom analysis demonstrates that the runs are complementary: four of them win at least one symptom substantively, and wins concentrate in different parts of the ASRS bifactor structure. Third, two negative results follow from the same analysis, namely that a plain retrieval-only bi-encoder outranks the trained reranker on six symptoms versus three, and that three symptoms (approximately 14% of relevant documents in the majority qrels) are missed entirely at the candidate stage rather than mis-ranked. Evaluated against the official qrels, our best run reaches mean average precision (MAP) of 0.159 under majority assessment (rank 7 of 45 runs) and 0.134 under unanimity (rank 5 of 45). Our analysis identifies candidate-stage recall and symptom-routed combination of the runs as the priority directions for future work.

2. Task and Data

2.1. Problem statement

The task can be formalised as follows:
Given:

- The set $S = \{s_1, \dots, s_{18}\}$ of 18 ASRS-v1.1 self-report symptoms, each described by its questionnaire item.
- A test corpus C of 4,170,875 social-media sentences with their (PRE, TEXT, POST) context triplets.

Produce: for each $s \in S$, a ranking $R_s = (c_1, \dots, c_k)$ of up to $k \leq 1000$ sentences $c_i \in C$, ordered by decreasing estimated relevance to s.
Constraints: the task is zero-shot from the symptom descriptions (no training data is released), and each team submits up to 5 runs.
Evaluation: MAP (primary), nDCG (with nDCG@1000), and P@10. Relevance is assessed by top-k pooling with three expert assessors, yielding two qrels: majority ($\geq 2/3$) and unanimity ($3/3$).

The organisers define a sentence as relevant when it is *informative about the user's state with respect to the symptom, regardless of polarity* [3]. Evidence of absence counts equally with evidence of presence, which rules out presence-detection as the design target for our trained bi-encoder reranker in §3.5.

2.2. Corpus, qrels, and metrics

The test corpus C comprises 4,170,875 sentences from Reddit, with no training data released for the task. Each sentence is delivered with a (PRE, TEXT, POST) context triplet and a unique identifier. The corpus is part of the eRisk benchmark series and was assembled under the data-use terms summarised in [6, 7]. The official assessment used top-k pooling of submitted run output, with 5,224 sentences in aggregate judged by three expert assessors. Each pooled sentence received a majority qrels label (relevant when at least 2 of 3 assessors agreed) and a unanimity qrels label (3 of 3), yielding 751 and 483 relevant sentences respectively. Three metrics are reported: MAP (the primary one), nDCG (including nDCG@1000), and P@10. We report against both qrels in §4 because they are not interchangeable, namely the relative ordering of our runs shifts between them.

3. System: CADRE

3.1. Overview

Our system, CADRE[1] (Clinically-grounded ADHD-symptom Detection and Retrieval Ensemble), is a three-stage pipeline whose architecture is shown in Figure 1. Candidate selection (§3.2) retrieves a small set of sentences per symptom from the test corpus. LLM scoring (§3.3–§3.4) labels each candidate against a clinically-grounded symptom representation, producing a silver supervision signal. A trained bi-encoder ensemble (§3.5) then re-scores the candidates one final time. Five runs were submitted that probe different points in the pipeline: three trace the main flow at successive depths (R5 stops after retrieval, R2 stops after LLM scoring, R1 runs the full pipeline), one fuses two of them by reciprocal rank fusion (R3 combines R1 and R2), and one replaces the ADHD-specific scoring with a cross-disorder transfer baseline (R4). The per-run derivation, including R4's partial coverage and fallback behaviour, is given in §3.6.

3.2. Candidate selection

Stage 1 retrieves a candidate set for each symptom from the 4.17M-sentence test corpus using a general-purpose bi-encoder, `all-mpnet-base-v2` [8]. The retrieval is shaped by three knobs applied in sequence. First, the symptom query is expanded with empirical discussion topics drawn from clinical descriptions of how the symptom manifests, supplementing the brief ASRS item phrasing. Second, a first-person filter retains only sentences expressing self-disclosure, removing roughly 60% of the corpus and biasing the candidate set toward target writers' own statements. Third, an additive keyword boost (+0.05 per matched keyword, with 16–18 symptom-specific keywords per item) is summed onto the cosine similarity to favour candidates that contain symptom-salient terms. Each symptom receives a top-5000 candidate list passed to §3.4 for LLM scoring. The 5000 cardinality balances LLM scoring throughput against recall headroom for the Stage-3 trained ranker.

3.3. Symptom representation

Four-layer clinical definitions. The ASRS items are Likert-style screening prompts asking how often a symptom occurs (e.g. item 13: "How often do you feel restless or fidgety?"), not descriptions of how the symptom manifests in everyday text. Feeding such a prompt directly to the LLM would push it toward matching on the prompt itself rather than reasoning about how the symptom surfaces in writing. We therefore augment each ASRS item with a four-layer template that the LLM sees in place of the bare item. The first layer reproduces the DSM-5-TR criterion text for the corresponding ADHD symptom [1], anchoring the LLM in the clinical definition. The second layer describes adult behavioural manifestation in plain language, naming how the symptom typically presents in adults

[1]The system was developed under the working name *HiPerT* and submitted under that label. We use CADRE in this paper throughout; the trained-reranker run we refer to as "CADRE full" appears in the official Perez et al. (2026) [3] results table under its submission label "HiPerT full".

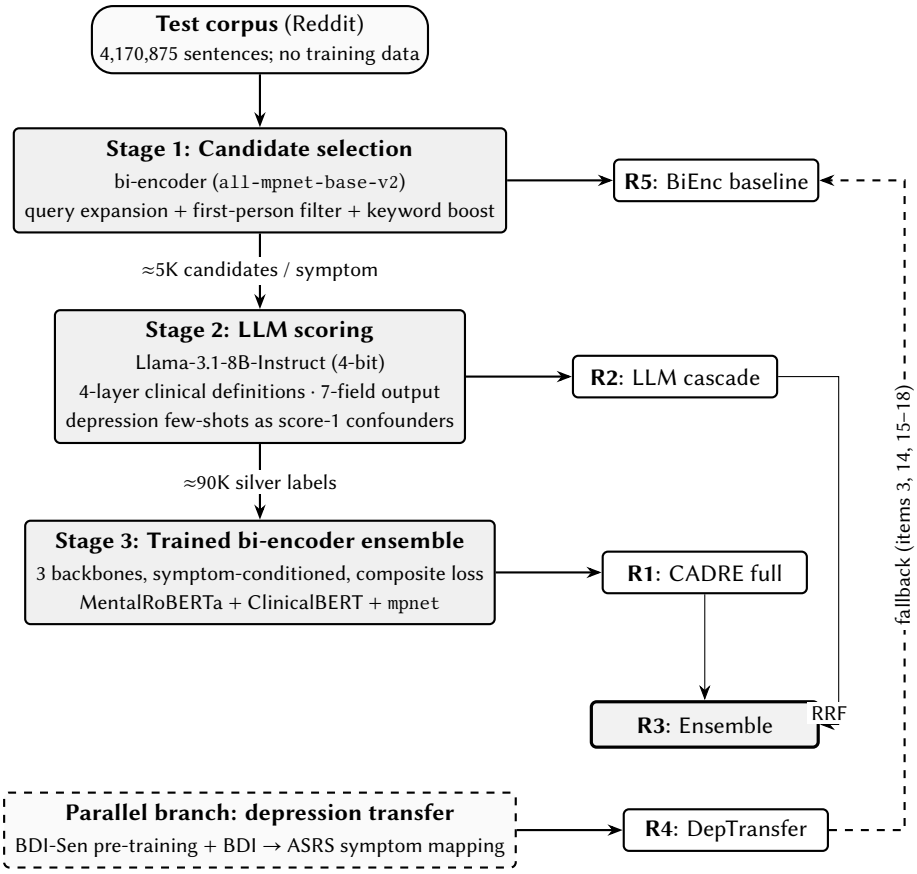

Figure 1: The CADRE pipeline. Five submitted runs are derived from a three-stage pipeline: R5 from Stage 1 retrieval, R2 from Stage 2 LLM scoring, and R1 from Stage 3 trained-bi-encoder reranking; R3 is the reciprocal-rank fusion of R1 and R2; R4 is a parallel depression-transfer branch using BDI-Sen pre-training and a BDI-II to ASRS symptom mapping covering 12 of the 18 symptoms, with the 6 unmapped items (3, 14, 15–18) falling back to R5 (dashed arrow). The system was developed under the working name *HiPerT* and submitted under that label; we refer to it as CADRE throughout this paper, including for the trained-reranker run *CADRE full* (which appears in the official Perez et al. 2026 [3] results table as *HiPerT full*).

rather than children. The third layer characterises empirical social-media language patterns associated with the symptom: register, voice, and recurring patterns of self-reference. These patterns are derived from clinical descriptions of how adults discuss the symptom [5], not from inspection of the test corpus. The fourth layer lists differential markers that separate the symptom from confounding disorders the writer might also discuss, most prominently depression and anxiety. Mind that the augmentation does not replace the ASRS item itself: the symptom intent comes from the item, and the four layers elaborate around it.

Schematically, the template takes the following shape (with tier-dependent omissions described in the token-budget allocation below; a fully worked example for a representative symptom is given in Appendix A):

```
SYMPTOM <n>: <short label>                                    [e.g. 13: restless/fidgety]
L1 (DSM-5-TR criterion):  <criterion text from APA 2022>      [always present]
L2 (adult manifestation): <plain-language description>        [omitted: compressed, minimal]
L3 (empirical language):  <register, voice, self-reference>   [always present]
L4 (confounders):         <differential markers>              [omitted: minimal]
```

Token-budget allocation. Not all 18 symptoms receive the same template. We allocate template layers per symptom by a token budget tied to cross-diagnostic ambiguity. Symptoms at higher risk of confusion with depression or anxiety receive the full four-layer template (the *full* tier, 5 items: 7–11). Symptoms at intermediate risk drop the adult-behavioural-manifestation layer (the *compressed* tier, 7 items: 1–4, 6, 13–14). Symptoms

naming concrete observable behaviours retain only the DSM-5-TR criterion and the empirical-language layer (the *minimal* tier, 6 items: 5, 12, 15–18).

3.4. LLM scoring cascade

Stage 2 scores every candidate sentence against the augmented symptom representation from §3.3 using a single LLM, `Llama-3.1-8B-Instruct` quantised to 4-bit precision (temperature 0.1, maximum 512 output tokens, served via HuggingFace on Colab). The scorer produces a 7-field structured output: SYMPTOM_MATCH, SELF_REFERENCE (DIRECT/INDIRECT/NONE), DETAIL_LEVEL, CONFOUNDERS, SCORE (a 0–3 relevance label), CONFIDENCE (a 1–5 self-rating), and REASONING. Depression sentences from BDI-Sen [9] are included as in-prompt confounders that anchor the score-1 level, namely as examples of sentences that are informative about a writer's state but not about ADHD specifically. We designed an escalation cascade to route uncertain or low-confidence candidates to a stronger scorer. At submission time, however, no stronger model was available in the runtime budget, so escalation collapsed to the same Llama-3.1-8B-4bit. After scoring, a resolution step combines the per-candidate label and confidence into a confidence-weighted silver label (weight in $[0.15, 0.85]$) for use in Stage-3 ranker training.

3.5. Trained bi-encoder ranker

Stage 3 reranks the candidate pool with a trained bi-encoder ensemble. Three backbones are fine-tuned in parallel: MentalRoBERTa [10], ClinicalBERT [11], and `all-mpnet-base-v2` [8]. The three backbones provide complementary domain coverage: MentalRoBERTa specialises in mental-health-Reddit register, ClinicalBERT brings clinical-terminology grounding, and `all-mpnet-base-v2` contributes general-purpose semantic generalisation. Each backbone is a symptom-conditioned encoder: the sentence and the symptom representation from §3.3 are encoded together, and the candidate's relevance is read out from a small ordinal head with four bins matching the 0–3 silver-label scale.

Training proceeds in three stages of progressive domain transfer. **Stage A1** pre-trains each backbone on the BDI-Sen sentence-level depression labels [9] over the 21-item BDI-II symptom set. **Stage A2** transfers to the eRisk 2025 Task 1 sentence-relevance data [7] and extends the input encoding to include the (PRE, TEXT, POST) context window used at inference. **Stage B** fine-tunes on the approximately 90,000 ADHD silver labels from §3.4, restricted to the 18 ASRS symptoms and with a curriculum schedule that prioritises high-confidence candidates early in training.[2]

All three stages share a composite loss with four components: an ordinal cross-entropy term over the 0–3 relevance scale; a margin ranking loss (margin = 0.5) to enforce pairwise ordering; and a small hierarchy regulariser ($\lambda_{\text{repel}} = 0.01$) that separates symptom-pair representations. Component weights for (ordinal, ranking, hierarchy) are $(1.0, 0.5, 0.05)$ throughout. The ordinal component is swapped between stages to match label reliability: ordinal cross-entropy with label smoothing ($\varepsilon = 0.2$) for the cleaner BDI-Sen and eRisk-2025 labels in A1 and A2, and symmetric cross-entropy [12] ($\alpha = 1.0$, $\beta = 0.5$) for the noisy ADHD silver labels in B. After Stage B, a per-symptom temperature plus Dirichlet calibration is fit on the Stage-B held-out validation split, producing the calibrated relevance distribution $p_{\text{cal}}(r \mid s, q)$ used at inference.

At inference time, each candidate receives a per-backbone score given by the calibrated expected relevance $\phi = \sum_r r \cdot p_{\text{cal}}(r \mid s, q)$ with $r \in \{0, 1, 2, 3\}$. Each backbone scores a candidate pool 2× wider than the final top-1000. The three per-backbone scores are then combined by uniform mean, and the top-1000 of the averaged ranking forms the run output.

3.6. The five runs

The five submitted runs each probe a different point in the pipeline (compare Figure 1). **R5 (BiEnc baseline)** uses the Stage-1 retrieval cosine similarity directly as the ranking, exposing candidate selection alone. **R2 (LLM cascade)** outputs the Stage-2 silver labels directly as the ranking, exposing the LLM scorer at Stages 1+2. **R1 (CADRE full)** outputs the trained-bi-encoder rerank, exposing the full main flow at Stages 1+2+3. **R3 (Ensemble)** fuses R1 and R2 by reciprocal rank fusion (RRF) with the standard parameter $k = 60$ [13], operating

[2]Training hyperparameters across the three stages: AdamW optimiser (weight decay 0.01), max gradient norm 1.0, AMP mixed precision; max sequence length 128 tokens per segment, with A1 encoding text only and A2/B encoding text plus PRE/POST context as three separate segments. LR schedule: linear warmup over 500 optimiser steps followed by cosine annealing with warm restarts ($T_0 = 5$ optimiser steps). Per stage: A1 base LR 2×10^{-5}, max 10 epochs, batch 32, bottom 6 backbone layers frozen for the first 3 epochs then full fine-tuning at 0.1× LR; A2 base LR 1×10^{-5}, max 5 epochs, batch 32, full fine-tuning from step 0; B base LR 1×10^{-5}, max 15 epochs, batch 32, full fine-tuning. Early stopping monitors the composite loss on a 10% held-out validation split per stage.

Table 1
Official eRisk 2026 Task 3 results for our five submitted runs, alongside the field leader (NeuroRankUB) and the second-best team (erisk-cedri's EnsembleV3) for context. Metrics are reported under both the majority ($\geq 2/3$) and unanimity (3/3) qrels. The trained-reranker run referred to as "CADRE full" here appears in the official Perez et al. 2026 [3] results table under its submission label "HiPerT full". In bold, our best run.

	Majority qrels				Unanimity qrels			
Run	MAP	R-PREC	P@10	nDCG@1000	MAP	R-PREC	P@10	nDCG@1000
NeuroRankUB (rank 1)	0.248	0.328	0.528	0.547	0.207	0.274	0.367	488
erisk-cedri EnsembleV3 (2nd)	0.190	0.262	0.400	0.461	0.138	0.193	0.261	0.387
Ensemble (ours, R3, `INSALyon_Ensemble`)	**0.159**	**0.206**	**0.317**	**0.406**	0.129	0.160	0.244	0.363
LLM cascade (ours, R2, `INSALyon_LLM_cascade`)	0.158	0.180	0.344	0.402	**0.134**	**0.167**	**0.272**	**0.369**
CADRE full (ours, R1, `INSALyon_HiPerT_full`)	0.137	0.169	0.294	0.388	0.107	0.131	0.172	0.336
BiEnc baseline (ours, R5, `INSALyon_BiEnc_baseline`)	0.117	0.160	0.311	0.321	0.122	0.154	0.256	0.313
DepTransfer (ours, R4, `INSALyon_DepTransfer`)	0.054	0.087	0.100	0.216	0.054	0.065	0.078	0.202

over the top-1000 ranks of each run. Because RRF combines runs by rank positions rather than score values, no per-run score normalisation is needed. R1 (post-trained-ranker) and R2 (post-LLM-scoring) are chosen as the two end-to-end runs of the main flow, combining the symptom-specific LLM signal with the trained-ranker output. **R4 (DepTransfer)** replaces the ADHD-specific scoring with a bi-encoder pre-trained on BDI-Sen depression labels. Queries are derived from a Beck Depression Inventory-II (BDI-II) to ASRS item mapping that covers 12 of 18 symptoms. The 6 items with no clear BDI-II analogue (3, 14, 15–18; predominantly hyperactivity-impulsivity items) fall back to R5. The runs are complementary in a non-trivial way, as we report in §4.2.

4. Results

4.1. Official results

Our best run reaches mean average precision (MAP) 0.159 under majority assessment (rank 7 of 45 runs) and MAP 0.134 under unanimity (rank 5 of 45), placing our submissions in the upper-middle of the field on both qrels [3]. Table 1 reports the four official metrics across all five runs and both qrels, alongside the field leader NeuroRankUB and the second team erisk-cedri for context. Under majority qrels, `Ensemble` is our strongest run at MAP 0.159, followed closely by `LLM cascade` at 0.158, then `CADRE full` (0.137), `BiEnc baseline` (0.117), and `DepTransfer` (0.054). Under unanimity, the order shifts: `LLM cascade` becomes our strongest at MAP 0.134, followed by `Ensemble` (0.129), `BiEnc baseline` (0.122), `CADRE full` (0.107), and `DepTransfer` (0.054). The leader NeuroRankUB reaches MAP 0.248 / 0.207, putting our best runs 0.089 and 0.073 behind respectively. The second team erisk-cedri reaches 0.190 / 0.138, putting us within 0.031 and 0.004 of second place. The identity of our best run differs between qrels: `Ensemble` under majority, `LLM cascade` under unanimity. We take up this ranking shift, and the per-symptom structure behind it, in §4.2.

4.2. Ranking shift and per-symptom complementarity

Two structural facts behind the aggregate numbers matter. The first is the qrels shift just noted in §4.1: `LLM cascade` overtakes `Ensemble` as our best run when assessment is restricted to sentences on which all three assessors agree, even though the two runs sit within 0.001 MAP of each other under majority assessment. The second is per-symptom complementarity. A best-run-per-symptom breakdown shows that four of our five runs win at least one symptom *substantively* (where a substantive win has the winning run's per-symptom AP strictly above zero): `BiEnc baseline` wins 6, `LLM cascade` wins 4, `CADRE full` wins 3, `Ensemble` wins 2, and `DepTransfer` wins 0. The remaining three symptoms (7, 9, 10) have no substantive winner. We use the three-factor bifactor structure of Stanton et al. [14] to organise the results: **Inattention** (items 1–4, 7–11), **Motor Hyperactivity-Impulsivity** (items 5, 6, 12–14), and **Verbal Hyperactivity-Impulsivity** (items 15–18). Among the substantive wins, the only clear factor cluster is `LLM cascade`'s concentration in Verbal Hyperactivity-Impulsivity items, with 3 of its 4 wins on items 15, 16, and 18. `BiEnc baseline` tilts toward Inattention but with

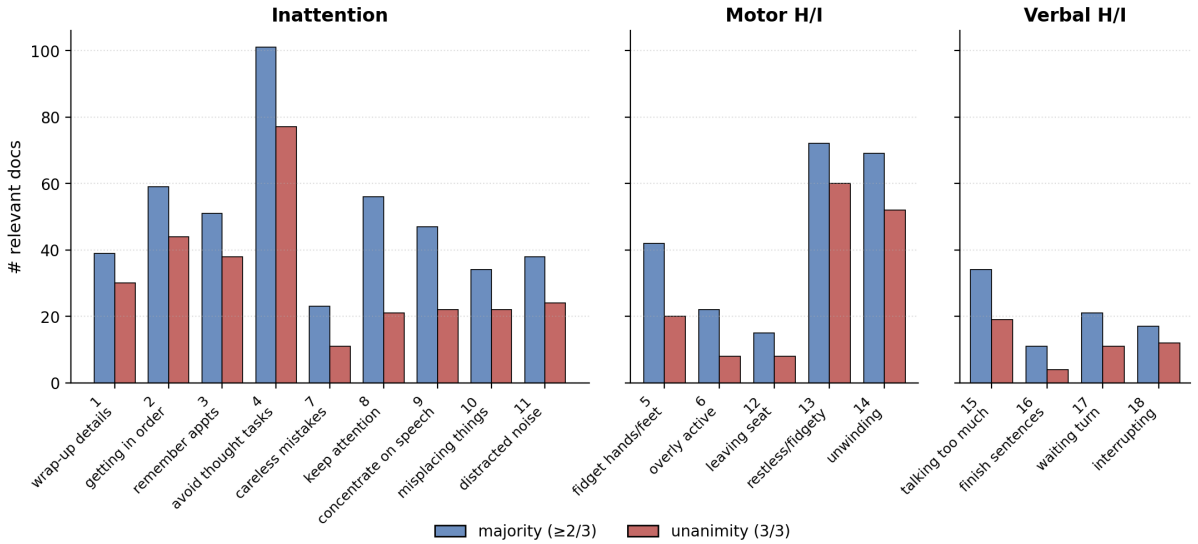

Figure 2: Number of documents judged relevant per ASRS symptom under the two qrels variants (majority: at least 2 of 3 assessors agree; unanimity: 3 of 3), grouped by ASRS bifactor structure (Stanton et al. [14]): Inattention, Motor Hyperactivity-Impulsivity, and Verbal Hyperactivity-Impulsivity. Per-symptom counts vary considerably under both qrels, and the relative shrinkage from majority to unanimity is not uniform: items with concrete observable behaviours (e.g. 13 "restless/fidgety", 4 "avoid thought tasks") retain most relevant docs under the stricter setting, while items with greater interpretive variability (e.g. 8 "keep attention", 5 "fidget hands/feet") lose a larger fraction. Counts are aggregate summary statistics.

two Motor wins. CADRE `full`'s three wins distribute one per factor without clear concentration. Figure 2 shows the per-symptom relevant-document counts under both qrels, grouped by bifactor structure. The per-symptom variation visible there is what drives the complementarity we observe across runs. The visible non-zero bars for items 7, 9, and 10 motivate the candidate-coverage analysis in §5.2.

5. Analysis

5.1. The trained reranker underperforms a plain bi-encoder

Our most counterintuitive finding is that the trained bi-encoder reranker underperforms a plain retrieval-only bi-encoder on a meaningful slice of the task. `BiEnc baseline` wins six symptoms substantively against CADRE `full`'s three (a 2:1 ratio under the substantive-wins count from §4.2). On three of those six wins (items 2, 4, 13), the retrieval-only bi-encoder reaches AP above 0.19, indicating substantial coverage where the trained reranker does not improve upon retrieval. We read the pattern as a sign that training on low-diversity silver labels has added noise on symptoms where semantic similarity from the retrieval stage is already strong. The LLM scoring rubric of §3.4 leaves most candidates with score 0 or 1. The trained reranker therefore sees an essentially binary supervision signal that gives it less room to discriminate than the unsupervised similarity it replaces. The composite-loss design of §3.5 includes a symmetric-cross-entropy term specifically intended to absorb such silver-label noise, but on this 6-vs-3 count the noise-absorption is not enough to clear the diversity ceiling that the silver labels impose.

5.2. Candidate-coverage failure on three symptoms

A second negative finding concerns three symptoms that together account for 14% of relevant documents in the majority qrels (104 of 751): item 7 (careless mistakes), item 9 (concentrate on speech), and item 10 (misplacing things). None of our five runs surfaces them substantively. For items 7 and 9, the per-symptom AP is exactly zero across all five runs, not only for the retrieval-only `BiEnc baseline` but also for the LLM scorer and the trained bi-encoder reranker. For item 10, only `DepTransfer` produces a non-zero score, and its AP is 0.0001 (a depression-transfer artefact at the depth of the ranked list, not a meaningful signal). The miss is not a base-rate effect: item 9 has the 8th-highest relevance rate of all 18 symptoms in the qrels, so other teams' runs surfaced relevant sentences ours did not. The pattern is consistent with two mechanisms operating against the candidate-selection

stage in §3.2. First, the first-person filter retains only sentences expressing self-disclosure, which excludes content where ADHD-related struggles are described in third-person, generalised, or narrative-past framings. Second, the LLM scoring rubric in §3.4 leans on presence-evidence. The `SYMPTOM_MATCH` and `SELF_REFERENCE` fields anchor on explicit first-person presence cues, which competes with the polarity-agnostic relevance definition from §2.1. The result is that the candidate set produced for these three symptoms misses relevant sentences not by mis-ranking them but by never including them in the top 1000. Candidate-stage recall is therefore the most actionable improvement lever for future editions. Two interventions are consistent with the failure mode: relaxing the first-person filter, and redesigning the LLM rubric to score informativeness rather than presence.

5.3. Ensemble underperformance and the development-evaluation gap

Two further observations follow from the analysis. The first is that `Ensemble`'s reciprocal-rank fusion averages per-symptom winners with losers and therefore underperforms a symptom-routing oracle that picks the best of our five runs on each symptom. The oracle reaches MAP 0.191 under majority qrels and 0.172 under unanimity, against our actual best of 0.159 (`Ensemble`) and 0.134 (`LLM cascade`). The headroom is +0.032 and +0.038 respectively. The unanimity headroom is the larger figure, consistent with the qrels shift noted in §4.2: RRF behaves differently under broader and stricter relevance because the per-symptom rank positions it averages diverge more under unanimity. The oracle is an upper bound on a system we did not build, not a method we propose. The practical question for future work is whether a learnt symptom-routing scheme can recover any meaningful fraction of this headroom.

The second observation concerns our internal development evaluation, which used a Llama-3.3-70B judge to score `LLM cascade`'s outputs and produced a P@10 of 94.4%. The official P@10 for the same run under the majority qrels is 0.344. The dev-evaluation thus overestimated quality by roughly 60 percentage points, even though the judge model was both different from and more capable than the scorer it evaluated. The lesson is not about same-model circularity but about LLM-as-judge reliability more generally: in a zero-shot regime without held-out validation data, an LLM judge can systematically agree with the system's own scoring biases and produce wildly optimistic estimates. Future work should pair zero-shot system development with an independent validation signal: a small held-out human-annotated set, a different evaluation protocol, or a non-LLM judge. The within-paper LLM-as-judge agreement we used is not a reliable substitute.

6. Conclusion and Future Work

We have presented CADRE, a five-run pipeline for eRisk 2026 Task 3 combining bi-encoder retrieval, LLM scoring grounded in four-layer clinical definitions, and a trained bi-encoder ensemble reranker. Our best run reaches MAP 0.159 under majority qrels (rank 7 of 45 runs) and 0.134 under unanimity (rank 5 of 45). Two findings beyond the headline are worth carrying forward: the five runs are complementary across the ASRS bifactor structure, and a plain retrieval-only bi-encoder is harder to beat than the trained reranker would suggest. Future work will target two priority directions. First, candidate-stage recall on the three symptoms missed by all runs (together about 14% of relevant documents in the majority qrels). Second, symptom-routed combination of the runs: a per-symptom oracle reaches MAP 0.191 under majority and 0.172 under unanimity, leaving +0.032 to +0.038 of headroom over our best actual run. Code and configurations for our five runs are released at https://github.com/diana-nurbakova/depression-detection.

Declaration on Generative AI

Two distinct uses of generative AI are disclosed here. First, generative AI is a methodological component of the system under study: `Llama-3.1-8B-Instruct` (4-bit, via HuggingFace on Colab) is used as the LLM scorer in Stage 2 of CADRE (§3.4), and `Llama-3.3-70B-Instruct` is used as the development-time evaluation judge whose dev-evaluation P@10 figure is reported and analysed in §5.3. Second, the authors used Anthropic Claude (Opus 4.7) for the following purposes: *drafting content* of some paragraphs based on the detailed outline; editing content including *paraphrasing and rewording*, and structural editing; and generation of portions of the pipeline and analysis code described in Sections 3 and 4. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content. The methodological design, the experimental decisions, the numerical claims, the analyses, and the conclusions are the authors' own.

References

[1] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, dsm-5-tr ed., American Psychiatric Association Publishing, 2022. URL: https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425787. doi:10.1176/appi.books.9780890425787.

[2] R. C. Kessler, L. Adler, M. Ames, O. Demler, S. Faraone, E. Hiripi, M. J. Howes, R. Jin, K. Secnik, T. Spencer, T. B. Ustun, E. E. Walters, 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. URL: https://www.cambridge.org/core/product/identifier/S0033291704002892/type/journal_article. doi:10.1017/S0033291704002892.

[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 (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.

[4] 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.

[5] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses, in: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 1–10. URL: http://aclweb.org/anthology/W15-1201. doi:10.3115/v1/W15-1201.

[6] F. Crestani, D. E. Losada, J. Parapar (Eds.), Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project, volume 1018 of *Studies in Computational Intelligence*, Springer International Publishing, Cham, 2022. URL: https://link.springer.com/10.1007/978-3-031-04431-1. doi:10.1007/978-3-031-04431-1.

[7] J. Parapar, A. Perez, X. Wang, F. Crestani, eRisk 2025: Contextual and Conversational Approaches for Depression Challenges, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, volume 15576, Springer Nature Switzerland, Cham, 2025, pp. 416–424. URL: https://link.springer.com/10.1007/978-3-031-88720-8_62. doi:10.1007/978-3-031-88720-8_62.

[8] 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3980–3990. URL: https://www.aclweb.org/anthology/D19-1410. doi:10.18653/v1/D19-1410.

[9] A. Pérez, J. Parapar, A. Barreiro, S. Lopez-Larrosa, BDI-Sen: A Sentence Dataset for Clinical Symptoms of Depression, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Taipei Taiwan, 2023, pp. 2996–3006. URL: https://dl.acm.org/doi/10.1145/3539618.3591905. doi:10.1145/3539618.3591905.

[10] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7184–7190. URL: https://aclanthology.org/2022.lrec-1.778/.

[11] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly Available Clinical, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. URL: http://aclweb.org/anthology/W19-1909. doi:10.18653/v1/W19-1909.

[12] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, J. Bailey, Symmetric Cross Entropy for Robust Learning With Noisy Labels, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South), 2019, pp. 322–330. URL: https://ieeexplore.ieee.org/document/9010653/. doi:10.1109/ICCV.2019.00041.

[13] G. V. Cormack, C. L. A. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, Boston MA USA, 2009, pp. 758–759. URL: https://dl.acm.org/doi/10.1145/1571941.1572114. doi:10.1145/1571941.1572114.

[14] K. Stanton, M. K. Forbes, M. Zimmerman, Distinct dimensions defining the Adult ADHD Self-Report Scale: Implications for assessing inattentive and hyperactive/impulsive symptoms., Psychological Assessment 30 (2018) 1549–1559. URL: https://doi.apa.org/doi/10.1037/pas0000604. doi:10.1037/pas0000604.

A. Prompt Templates

This appendix shows the full four-layer prompt template used by Stage 2 (LLM scoring, §3.4). Section A.1 gives a schematic example of the template for one symptom. The L1 (Clinical anchor) block has identical structure across all 18 items: the verbatim DSM-5-TR criterion clause for the corresponding ADHD inattention or hyperactivity-impulsivity item [1] together with a one-line adult-adaptation gloss translating the child-oriented criterion phrasing to typical adult presentation; the criterion text itself is not reproduced in this appendix to respect APA copyright. Sections A.2 through A.4 summarise the content structure of layers L2, L3, and L4 across the 18 ASRS items, with explicit honesty notes on where descriptions are post-hoc taxonomic characterisations versus groupings of real prompt content. Section A.5 gives the 7-field structured output schema. No corpus-derived phrasings or example sentences appear anywhere in this appendix.

A.1. Schematic template: full-tier prompt for an Inattention item (item 8, "Sustain attention")

Symptom under scoring. ASRS-v1.1 item 8: "How often do you have difficulty keeping your attention when you are doing boring or repetitive work?"
Token-budget tier. Full (5 items: 7–11; high cross-disorder ambiguity, particularly with depression-driven concentration difficulties).
Prompt body, as supplied to the LLM scorer:

```
SYMPTOM 8: Sustain attention on boring or repetitive work

L1 (Clinical anchor -- DSM-5-TR):
<Verbatim DSM-5-TR criterion clause for the corresponding ADHD
 inattention item (here: criterion 1b, "difficulty sustaining
 attention"), plus a one-line adult-adaptation gloss translating
 the child-oriented criterion phrasing to typical adult
 presentation. The criterion text is omitted from this appendix to
 respect APA copyright; the structural description appears in the
 appendix preamble.>

L2 (Adult behavioural manifestation):
<Concrete daily-life presentations of the symptom in prose, ranging
 over the six life-domain categories (work, home, academic,
 administrative, social, personal) enumerated in A.2. Derived from
 adult-ADHD clinical literature; not derived from inspection of the
 test corpus.>

L3 (Online Discussion Patterns):
<Per-symptom free-text content: a short list of first-person
 expressions characteristic of how adults describe this symptom in
 informal text. Derived from clinical and computational literature
 on adult ADHD language; not derived from inspection of the test
 corpus. See A.2 for a taxonomic synthesis of L3 pattern types
 across all 18 symptoms.>

L4 (Differential-diagnosis markers):
<Structured "vs. <condition>" contrasts across approximately 25
 confounding conditions, plus an "ADHD-specific:" confirming line
 (present in all 18 items). Derived from DSM-5-TR differential
 sections (APA 2022) and standard adult-ADHD clinical literature;
 not derived from inspection of the test corpus. The
```

```
 condition-family grouping is given in A.4.>

CANDIDATE SENTENCE (with surrounding context):
[PRE]:  <preceding sentence from the (PRE, TEXT, POST) triplet>
[TEXT]: <candidate sentence to score>
[POST]: <following sentence from the (PRE, TEXT, POST) triplet>

INSTRUCTIONS:
Score the [TEXT] sentence on its relevance to SYMPTOM 8, using the
four layers above as the symptom representation. Return a structured
JSON object with the seven fields specified below. Polarity-agnostic:
a sentence is relevant if it is informative about the writer's state
with respect to the symptom, whether through affirmation or denial.
```

The full tier (5 items: 7–11) uses all four layers as shown. The compressed tier (7 items: 1–4, 6, 13–14) omits the L2 block. The minimal tier (6 items: 5, 12, 15–18) retains only L1 and L3, omitting L2 and L4.

A.2. L2: adult-behavioural-manifestation life-domain categories

The L2 block describes concrete daily-life presentations of the symptom, organised across six life-domain categories. The grounding documentation enumerates these explicitly; the deployed configuration flattens the same content to per-symptom prose that ranges over the same six domains:

1. **Work / professional** – task execution, meetings, deadlines, colleague perception.
2. **Home / household** – chores, maintenance, domestic routines.
3. **Academic / learning** – studying, reading, coursework.
4. **Administrative / financial** – bills, forms, taxes, appointments, paperwork.
5. **Social / relational** – conversations, friendships, partner dynamics.
6. **Personal / internal and daily routine** – leisure, sleep, self-regulation, moment-to-moment habits.

Per-symptom coverage of the six domains varies with where the symptom most plausibly manifests; not every symptom touches every domain. The six domains here are explicit in the grounding documentation rather than a post-hoc characterisation; the deployed prose is a re-expression of the same domain structure, not an abstraction above it.

A.3. L3: post-hoc pattern-type synthesis

The L3 block of each per-symptom prompt template (§3.3) is implemented as free-text *Online Discussion Patterns* content rendered from the system's symptom configuration, not as a labeled taxonomy. The categories enumerated below are a *post-hoc* synthesis of pattern types that recur across the 18 per-symptom L3 blocks, abstracted above the concrete first-person phrasings used at runtime. They characterise the L3 design at a taxonomic level rather than reproducing its content.[3]

1. **Capability–selectivity contrast.** Language that juxtaposes preserved ability on novel, challenging, or hyperfocus-eliciting tasks against domain-specific failure on routine, boring, or task-completion contexts. Signals interest-driven rather than global impairment.
2. **Impaired-inhibition / involuntariness.** Framing the behaviour as automatic and resistant to will, locating the deficit in control rather than in desire, knowledge, or fear.
3. **Metacognitive awareness with self-judgment.** Explicit self-monitoring and retrospective regret, often pre-empting an external label such as lazy, rude, careless, or self-centred.
4. **Embodied internal-state report.** First-person description of somatic restlessness, internal mental restlessness, or motor-driven sensation, framed as physical rather than as mood or worry.
5. **Coping-scaffold and treatment talk.** References to external compensations (reminders, alarms, planners, trackers, headphones, body-doubling) and to medication-driven symptom change, used as indirect evidence of the symptom.
6. **Relational / third-party-perception framing.** Surfacing the symptom through its interpersonal or occupational fallout and how others read it (partner, colleagues), using others' reactions as the evidentiary anchor.

[3]The L3 content per symptom was derived from clinical descriptions of how adults discuss ADHD symptoms and from the computational literature on adult-ADHD language patterns [5]; it was not derived from inspection of the test corpus.

A.4. L4: differential-diagnosis condition families

The L4 block enumerates differential-diagnosis contrasts as keyed "vs. CONDITION" lines together with an "ADHD-SPECIFIC:" confirming line; the confirming line is present in all 18 items. The content spans approximately 25 distinct confounding conditions, which we group into six families by competing mechanism:[4]

1. **ADHD-specific signature.** The confirming pattern present in all 18 items: interest-driven engagement, control-deficit framing, lifelong and pervasive presentation. This is a positive cue rather than a confound.
2. **Mood: depression.** The most frequent contrast (appearing in five or more items): global amotivation, cognitive fog, psychomotor agitation, mood-congruent rather than task-specific patterns.
3. **Anxiety spectrum.** Generalised anxiety disorder, social anxiety, post-traumatic stress: worry- or threat-driven, perfectionist or avoidance-based rather than stimulation-driven.
4. **Other psychiatric or neurodevelopmental.** OCD, mania/hypomania, autism spectrum, behavioural addiction, personality traits (e.g. narcissistic, antisocial); each with a distinct competing mechanism.
5. **Medical, neurological, or substance.** Thyroid dysfunction, restless legs, akathisia, hearing or sensory disorders, ageing or dementia, caffeine effects; rule out organic or drug-induced causes.
6. **Normal variation.** Extroversion, ordinary impatience or boredom, flow states; distinguished from ADHD by severity, pervasiveness, and functional impairment.

Unlike the L3 synthesis of §A.3, the families above are a grouping of content that is explicitly present in the prompt: the "vs. CONDITION" lines are real entries that the LLM saw, and the families cluster them by competing mechanism rather than abstracting above runtime content. The "ADHD-SPECIFIC:" confirming line is a distinct positive cue that participates in scoring rather than functioning as a differential.

A.5. Output schema

The LLM scorer returns a structured JSON object for each candidate sentence. The schema is fixed across all symptoms and token-budget tiers:

```
{
  "SYMPTOM_MATCH":  "<true | false>",
  "SELF_REFERENCE": "<DIRECT | INDIRECT | NONE>",
  "DETAIL_LEVEL":   "<low | medium | high>",
  "CONFOUNDERS":    ["<list of differential disorders potentially in play>"],
  "SCORE":          "<0 | 1 | 2 | 3>",
  "CONFIDENCE":     "<1 | 2 | 3 | 4 | 5>",
  "REASONING":      "<short justification, used for diagnostic inspection only>"
}
```

Field semantics. SYMPTOM_MATCH is a coarse gate indicating whether the sentence speaks to the symptom prior to scoring. SELF_REFERENCE distinguishes first-person self-disclosure (DIRECT), implicit self-reference embedded in narrative (INDIRECT), and third-person or generic statements (NONE). DETAIL_LEVEL captures the specificity of the description, from generic mention to detailed self-report. CONFOUNDERS lists any differential disorders the sentence may be invoking (e.g. depression, anxiety); empty list if none. SCORE is the 0–3 relevance label that drives ranking: 0 = irrelevant; 1 = informative about the writer's state but not about ADHD specifically (this is where the BDI-Sen confounder anchoring applies, see §3.4); 2 = relevant; 3 = strongly relevant. CONFIDENCE is the LLM's 1–5 self-rating, used in the §3.4 resolution step to compute the confidence-weighted silver label (weight in $[0.15, 0.85]$). REASONING is a short free-text justification produced for diagnostic inspection during development; it is not consumed by downstream scoring.

[4] The L4 content per symptom was derived from DSM-5-TR differential-diagnosis sections [1] and standard adult-ADHD clinical literature; it was not derived from inspection of the test corpus.