erisk-cedri at eRisk 2026 Task 3: A Zero-Shot Ensemble Pipeline with Multi-Query Retrieval, NLI and Cross-Encoder Re-Ranking for ADHD Symptom Sentence Ranking

Filipa Afonso[1,*], Rui Pedro Lopes[1]

[1]CeDRI – Research Centre in Digitalization and Intelligent Robotics, Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

Abstract

This paper describes the participation of the erisk-cedri team in Task 3 of eRisk 2026, devoted to the ranking of candidate sentences according to their relevance for each of the 18 symptoms of the Adult ADHD Self-Report Scale (ASRS-v1.1). Since the task is in its first edition and no annotated training data was released, a fully zero-shot strategy was adopted. The pipeline combines four complementary components: a multi-query semantic similarity stage based on all-mpnet-base-v2 with three to five query reformulations per symptom and max-pooling; a Natural Language Inference (NLI) stage using DeBERTa-v3-large fine-tuned on MNLI, FEVER, ANLI, LingNLI and WANLI; a keyword module with high and medium confidence dictionaries derived from the ASRS-v1.1 items; and a cross-encoder re-ranking stage (ms-marco-MiniLM-L-12-v2) applied to the top-100 candidates per symptom. Five runs were submitted, spanning the spectrum from single-component baselines to weighted ensembles that progressively shift more weight to NLI and add the cross-encoder boost. The best run (EnsembleV3) achieved AP = 0.190, R-PREC = 0.262, P@10 = 0.400 and NDCG@1000 = 0.461 under majority-vote relevance judgements, placing the team in second position among the 12 participating teams. The results indicate that, in the absence of annotated training data, careful combination of semantic retrieval, entailment-based scoring and lexical signals is a strong baseline, and that NLI plays a more central role than initially expected.

Keywords

ADHD, ASRS-v1.1, Information Retrieval, Sentence Ranking, Zero-Shot Learning, Natural Language Inference, Cross-Encoder Re-Ranking, eRisk 2026

1. Introduction

Mental health screening from social media writings has been a long-standing line of work in the eRisk Lab [3, 1, 2]. Recent editions of the lab have moved from coarse user-level detection to fine-grained, sentence-level ranking of evidence aligned with clinical questionnaires. The 2024 and 2025 editions of Task 1 used the Beck Depression Inventory (BDI-II) and its 21 symptoms as the structure for the ranking targets. In 2026, the lab introduces a new symptom ranking task focused on Attention-Deficit/Hyperactivity Disorder (ADHD), using the 18 items of the Adult ADHD Self-Report Scale (ASRS-v1.1) [4] as queries.

The task is framed as an Information Retrieval (IR) problem. Given a corpus of approximately 4.17 million sentences extracted from Reddit [1, 2], participants must rank, for each of the 18 ASRS-v1.1 symptoms, up to 1000 candidate sentences in decreasing order of relevance. A sentence is considered relevant if it conveys information about the user's state with respect to that symptom, regardless of polarity. Relevance assessments were collected by top-k pooling with three expert annotators and reported under two settings: majority vote (at least two of three) and unanimity (all three).

A central characteristic of this edition is that, being the first iteration of the ADHD ranking task, no annotated training data was provided. This precludes any directly supervised approach unless synthetic labels are generated. The team had previously developed a supervised pipeline for the depression

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany

*Corresponding author.

a61183@alunos.ipb.pt (F. Afonso); rlopes@ipb.pt (R. P. Lopes)

task, based on a multilabel DeBERTa-v3-large model with 21 LoRA adapters [6, 12], complemented by semantic retrieval and lexical rules. When that architecture was transferred to ADHD using synthetic data generated with a large language model, the supervised component failed to discriminate between the 18 symptoms. After consultation with the supervisor, the team pivoted to a fully zero-shot pipeline.

This paper describes the resulting system and the five runs submitted for evaluation. Three main contributions are made: (i) it is shown that a careful combination of multi-query semantic retrieval, NLI-based scoring and cross-encoder re-ranking is competitive in the absence of training data, ranking the team second out of twelve in the official evaluation; (ii) an ablation across the five runs is provided that quantifies the contribution of each component; (iii) lessons learned from the failed attempt at supervised fine-tuning with synthetic data are reported, which are believed to be useful for future participants.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the pipeline in detail. Section 4 reports the official results and Section 5 discusses them. Section 6 concludes.

2. Related Work

Symptom ranking at eRisk. The symptom ranking task was introduced in the eRisk lab in 2023 and consolidated in the 2024 and 2025 editions, all focused on depression and BDI-II. Participating systems have explored a wide range of approaches, including dense retrieval with sentence encoders, cross-encoder re-ranking, supervised binary classifiers per symptom, multilabel transformer models and ensemble combinations. The 2026 edition extends this line of work to ADHD and to the 18 items of the ASRS-v1.1 [4, 1, 2].

Sentence encoders and dense retrieval. Sentence-Transformers [5] project sentences into a fixed-dimensional vector space where cosine similarity correlates with semantic equivalence. The all-mpnet-base-v2 model used in the pipeline produces 768-dimensional embeddings and has been shown to be a strong general-purpose baseline across semantic textual similarity benchmarks. In the IR literature, bi-encoders are typically used for first-stage retrieval and combined with more expensive cross-encoder re-rankers in the final stage [10].

Cross-encoder re-ranking. Cross-encoders process the query and the document jointly through a transformer and produce a single relevance score. They are typically more accurate than bi-encoders for fine-grained ranking but more computationally expensive, which is why they are usually applied only to a shortlist of candidates from a first retrieval stage. The ms-marco-MiniLM family of models [11] has become a standard re-ranker due to its favourable accuracy-efficiency trade-off.

Zero-shot classification via NLI. Yin et al. [9] showed that NLI models trained on large corpora such as MNLI [7] can be used as zero-shot text classifiers by reformulating the candidate label as a hypothesis and computing entailment with respect to the input sentence. Subsequent models, such as DeBERTa-v3-large-mnli-fever-anli-ling-wanli used in the pipeline, are trained on a combination of MNLI, FEVER, ANLI [8], LingNLI and WANLI, which makes them robust to adversarial and out-of-distribution inputs. This paradigm is used in the present system to assign an entailment score to each candidate sentence given a symptom-specific hypothesis derived from the ASRS-v1.1.

3. System Description

3.1. Pipeline Overview

The pipeline is organised as a cascade of four zero-shot components. The first three are scoring components that produce, for each pair (sentence, symptom), a normalised score in $[0, 1]$. The fourth is a

re-ranking component applied only to the top of the candidate list. The data flow can be summarised as follows: starting from the raw TREC corpus, sentences are extracted and deduplicated, embeddings are precomputed, multi-query similarity is run to obtain a top-5000 shortlist per symptom, NLI and keyword scores are computed on that shortlist, the top-100 are optionally re-ranked with a cross-encoder, and the scores are finally combined with run-specific weights.

3.2. Corpus Preprocessing

The released corpus contains 4,170,875 sentences in TREC format, where each document encodes a single sentence with its identifier and the immediate textual context (PRE, TEXT, POST). The TREC files were parsed into a tabular representation preserving the original sentence identifiers, which are the identifiers expected in the submission files. A full-text field was then constructed by concatenating the preceding context, the target sentence and the following context, which serves as the input for all downstream components.

After exact deduplication on the full-text field, the working corpus contains 3,734,663 unique sentences (a reduction of 9.6%). All retrieval and scoring operations are performed on the deduplicated corpus. Submissions are produced using the original sentence identifiers, so the deduplication is transparent to the evaluation.

3.3. Multi-Query Semantic Similarity

The all-mpnet-base-v2 model [5] is used to obtain 768-dimensional embeddings for every sentence in the deduplicated corpus. Embeddings are computed in fp16 on an NVIDIA RTX 3090 (24 GB VRAM), processed in 38 chunks of 100,000 sentences each, and concatenated into a single 5.4 GB array. All embeddings are L2-normalised, so that cosine similarity reduces to a dot product.

For each of the 18 symptoms, between three and five query formulations were manually authored covering distinct registers: a clinical paraphrase of the ASRS-v1.1 item, two or three colloquial reformulations in the style of Reddit self-reports, and an alternative phrasing that captures synonyms or related expressions. For instance, for symptom 1 (*Difficulty wrapping up the final details of a project*) the queries ranged from "I have trouble finishing projects. I struggle to wrap up the final details of tasks once the challenging parts are done" to "i always abandon projects right before they're done, the final details are so boring i just can't push through".

For each symptom s, let $\{q_{s,1}, \ldots, q_{s,n_s}\}$ be the set of n_s query reformulations and let $\mathbf{e}_{q_{s,j}}$ be their normalised embeddings. The similarity score of sentence i with respect to symptom s is computed as the maximum over query variants:

$$\mathrm{sim}(i, s) = \max_{j=1,\ldots,n_s} \mathbf{e}_i^\top \mathbf{e}_{q_{s,j}}. \tag{1}$$

Max-pooling allows each variant to specialise in a different way of expressing the symptom, while ensuring that a sentence which strongly matches any one formulation receives a high score. To keep memory usage bounded, these dot products are computed in chunks of 500,000 sentences. The top 5000 candidates per symptom are retained together with their similarity scores. This shortlist is the input to the subsequent components.

3.4. Natural Language Inference Scoring

The NLI component is implemented with DeBERTa-v3-large-mnli-fever-anli-ling-wanli, a model fine-tuned on MNLI [7], FEVER, ANLI [8], LingNLI and WANLI. For each symptom s, a single hypothesis was written that paraphrases the ASRS-v1.1 item in the third person and that can be unambiguously evaluated against a candidate sentence treated as the premise. For example, the hypothesis for symptom 3 is "The user frequently forgets appointments and obligations".

Given a premise sentence p_i and a hypothesis h_s, the pair is fed to the NLI model and the softmax probability of the *entailment* class is taken as the NLI score:

$$\mathrm{nli}(i, s) = P(\text{entailment} \mid p_i, h_s). \tag{2}$$

Scoring is performed only on the 5000-candidate shortlist produced by the similarity stage. After deduplication across symptoms, the union of shortlists contains 49,808 unique sentences, which keeps the cost of the NLI stage tractable. With fp16 inference and a batch size of 32, the full NLI scoring of all 5000 candidates for the 18 symptoms is completed in approximately four minutes on the RTX 3090.

3.5. Keyword Matching

The keyword component implements a lightweight lexical safety net for sentences that contain explicit ADHD vocabulary but are not necessarily captured by the dense components. For each symptom, a dictionary was curated with two confidence levels: *high confidence* expressions that are strongly indicative of the symptom (for example, "finish sentences" or "can't wait my turn") and *medium confidence* expressions that are broadly compatible but more ambiguous ("interrupt", "patience"). The dictionaries were derived from a careful reading of the ASRS-v1.1 items, supplemented by qualitative inspection of the subreddit `r/ADHD`.

Given a sentence and a symptom, occurrences of the high and medium confidence terms in the lowercased full-text are counted, weighted by 3.0 and 1.0 respectively. The raw score is then normalised by dividing by 15 and clipped to the interval $[0, 1]$:

$$\mathrm{kw}(i, s) = \min\left(\frac{3 \cdot n_{\text{high}}(i, s) + n_{\text{med}}(i, s)}{15},\, 1\right). \tag{3}$$

This normalisation makes the keyword score directly comparable with the similarity and NLI scores after the min-max normalisation step described below.

3.6. Cross-Encoder Re-Ranking

The cross-encoder stage uses ms-marco-MiniLM-L-12-v2, a 12-layer cross-encoder pre-trained on the MS MARCO passage ranking dataset [11]. The model receives the symptom query and the candidate sentence as a single packed input and produces an unnormalised relevance score. The cross-encoder is applied only to the top-100 candidates per symptom after the similarity and NLI stages, which keeps the total inference cost below two seconds for all 18 symptoms.

The raw cross-encoder outputs vary substantially across symptoms, with observed ranges from roughly $[-11.3, -4.3]$ for the symptoms that the MS MARCO domain captures less well (for example, "Misplacing things") to $[-11.3, 4.7]$ for symptoms where the model discriminates strongly ("Waiting Turn", "Distracted by Noise"). Since the cross-encoder is used only as an additive boost on the top-100 of two of the runs, what matters is the relative ordering within each symptom, not the absolute scale. Min-max normalisation is applied per symptom before combining with the other components.

3.7. Ensemble Strategy

For each run, the final score of a candidate is a weighted linear combination of the normalised component scores:

$$\mathrm{score}(i, s) = w_{\text{sim}} \cdot \widetilde{\mathrm{sim}}(i, s) + w_{\text{nli}} \cdot \widetilde{\mathrm{nli}}(i, s) + w_{\text{kw}} \cdot \mathrm{kw}(i, s) + w_{\text{ce}} \cdot \widetilde{\mathrm{ce}}(i, s), \tag{4}$$

where $\widetilde{\cdot}$ denotes per-symptom min-max normalisation and w_{ce} is zero for candidates that are not in the top-100 shortlist of the cross-encoder. Candidates are then sorted in decreasing order of the final score, and the top 1000 per symptom form the submission ranking in standard TREC format.

3.8. Submitted Runs

Five runs were submitted, probing different points in the trade-off between the components, as summarised in Table 1. The first two are single-component baselines that allow the contribution of each scoring stage to be measured in isolation. The remaining three are ensembles that progressively shift weight from similarity to NLI, and that introduce the cross-encoder boost in the last two.

Table 1
Configuration of the five submitted runs.

Run	w_{sim}	w_{nli}	w_{kw}	Cross-Encoder
VecSearch	1.00	0.00	0.00	No
NLIZeroShot	0.00	1.00	0.00	No
EnsembleV1	0.50	0.35	0.15	No
EnsembleV2	0.35	0.50	0.15	Yes (top-100)
EnsembleV3	0.25	0.55	0.10	Yes (top-100)

The progression from EnsembleV1 to EnsembleV3 reflects two design decisions. First, the weight of NLI was increased relative to similarity, since exploratory inspection of the top-ranked sentences suggested that the entailment signal was more discriminative than expected. Second, the cross-encoder boost was added in the last two runs to test whether a final re-ranking on the top-100 improves the precision at the head of the ranking.

3.9. Implementation Details

The full pipeline was implemented in Python using PyTorch, Sentence-Transformers [5], the Hugging Face Transformers library, and standard NumPy and Pandas for data handling. All inference was carried out on a single NVIDIA RTX 3090 with 24 GB of VRAM. fp16 was used throughout, and the corpus embeddings were chunked into 38 blocks of 100,000 sentences to keep the memory footprint bounded. End-to-end execution of the full pipeline, from raw corpus to final TREC submissions, takes approximately 45 minutes, dominated by the embedding precomputation. Submissions were transferred to the eRisk FTP server in standard TREC format.

4. Results

The official evaluation of Task 3 was performed under two relevance settings derived from three expert annotators: *majority vote*, where a sentence is considered relevant if at least two annotators agree, and *unanimity*, where all three annotators must agree. Tables 2 and 3 report the results for the five submitted runs under both settings, together with their global rank among the 45 runs submitted by the 12 participating teams. It should be noted that, following a correction issued by the organisers, the NDCG metric reported below corresponds to NDCG@1000 rather than the originally announced NDCG@100.

Table 2
Official results of the five submitted runs under **majority-vote** relevance judgements, together with global rank among 45 runs from 12 teams.

Run	AP (rank)	R-PREC (rank)	P@10 (rank)	NDCG@1000 (rank)
EnsembleV3	0.190 (3)	0.262 (3)	0.400 (6)	0.461 (5)
EnsembleV2	0.186 (4)	0.259 (4)	0.406 (5)	0.459 (6)
EnsembleV1	0.167 (6)	0.239 (6)	0.372 (8)	0.446 (7)
VecSearch	0.105 (11)	0.162 (10)	0.278 (13)	0.350 (13)
NLIZeroShot	0.097 (12)	0.162 (10)	0.267 (14)	0.379 (10)

Table 3
Official results of the five submitted runs under **unanimity** relevance judgements, together with global rank among 45 runs from 12 teams.

Run	AP (rank)	R-PREC (rank)	P@10 (rank)	NDCG@1000 (rank)
EnsembleV3	0.138 (4)	0.193 (5)	0.261 (8)	0.387 (6)
EnsembleV2	0.134 (6)	0.191 (6)	0.278 (4)	0.384 (7)
EnsembleV1	0.122 (9)	0.190 (7)	0.233 (10)	0.372 (10)
VecSearch	0.090 (11)	0.148 (10)	0.200 (11)	0.304 (11)
NLIZeroShot	0.057 (22)	0.093 (20)	0.117 (24)	0.297 (12)

Under majority-vote, the best run (EnsembleV3) achieved AP = 0.190, ranking third among all 45 submissions and second at the team level, only behind the runs of NeuroRankUB. The three ensemble runs occupy the third, fourth and sixth positions in AP, while the two single-component baselines fall to the eleventh and twelfth positions. Under the more restrictive unanimity setting, EnsembleV3 remains the best run from the team with AP = 0.138, again ranking fourth globally. The relative ordering of the five runs is stable across both settings and across all four metrics.

Component ablation. The five runs naturally form an ablation study. The two single-component baselines (VecSearch with AP = 0.105 and NLIZeroShot with AP = 0.097 under majority vote) confirm that neither dense retrieval nor entailment scoring alone is sufficient to obtain competitive performance. The first ensemble, which combines the two scoring components with a small keyword contribution, already achieves AP = 0.167, a relative improvement of 59% over the strongest baseline. Increasing the weight of NLI and adding the cross-encoder boost yields a further improvement to AP = 0.186 (EnsembleV2) and AP = 0.190 (EnsembleV3), confirming the value of both design choices.

Effect of the cross-encoder. The two runs that include the cross-encoder boost (EnsembleV2 and EnsembleV3) consistently outperform EnsembleV1 across all metrics under both relevance settings. The gain is modest in absolute terms (around 0.02 in AP under majority vote) but is observed systematically and is most visible at the head of the ranking, where the cross-encoder operates. This is consistent with the literature on multi-stage retrieval, where cross-encoder re-ranking typically improves precision at the top of the ranking at the cost of additional computation [10].

Effect of NLI weight. Moving from a similarity-heavy ensemble (EnsembleV1, $w_{\text{sim}} = 0.50$) to an NLI-heavy ensemble (EnsembleV3, $w_{\text{nli}} = 0.55$) improves AP from 0.167 to 0.190 under majority vote. This is a non-trivial finding given that NLIZeroShot in isolation is the weakest of the baselines. The implication is that NLI complements similarity rather than substituting for it: similarity provides broad coverage, and NLI filters and re-ranks within that coverage based on entailment with the symptom hypothesis.

5. Discussion

5.1. From Supervised to Zero-Shot

The team's initial plan was to transfer the supervised pipeline developed for the 2024 and 2025 depression tasks (a multilabel DeBERTa-v3-large model with per-symptom LoRA adapters) to the ADHD setting, using synthetic data generated with a large language model. Approximately 150 synthetic sentences per symptom were produced, with a high level of manual curation, but the supervised models failed to discriminate between the 18 symptoms: predictions for a wide variety of inputs collapsed into a narrow score range, and the few synthetic positives could not overcome the breadth of the symptom space. This is interpreted as a consequence of the combination of multi-label granularity (18 symptoms) and

the limited amount of synthetic data: each symptom only had a few hundred positive examples to learn from, and the synthetic data lacked the long-tail variability of real social media writings.

After consultation with the supervisor, the team pivoted to a fully zero-shot pipeline. In retrospect, this decision was the most consequential of the project, and is believed to be the main reason competitive results were achieved despite the limited resources available. The lesson drawn from this experience is that, when annotated training data is unavailable for a fine-grained, multi-symptom IR task, the marginal value of small amounts of synthetic data is often lower than the marginal value of careful prompt engineering, query expansion and ensemble design over already strong pretrained models.

5.2. The Role of NLI in Sentence-Level Relevance

The most surprising finding of the ablation is that NLI, which by itself ranked twelfth among the baselines, ends up being the most heavily weighted component in the best ensemble. Two factors appear to explain this. First, NLI scores are very discriminative on the candidate shortlist: in the present experiments, the entailment probabilities span the full $[0.0001, 0.9995]$ range with no apparent saturation. Second, the failure mode of NLI is different from the failure mode of dense retrieval. Dense retrieval can be misled by surface-level lexical overlap that is semantically unrelated to the symptom, whereas NLI is more sensitive to whether the sentence logically implies the hypothesis. Combining the two reduces the impact of each error mode in isolation.

5.3. Limitations

The pipeline has several limitations. First, the keyword dictionaries were hand-crafted from the ASRS-v1.1 items and from a brief inspection of `r/ADHD`. A more systematic procedure, for example based on automatic term extraction from a clinically validated corpus, would likely yield a stronger lexical component. Second, the NLI hypotheses were authored manually and were not systematically tuned. Exploratory experiments with alternative hypotheses suggest that the choice of phrasing can have a measurable effect on the entailment scores, and a more principled selection procedure (for instance, hypothesis ensembles or learned prompts) is a natural direction for future work. Third, the polarity dimension of the relevance definition was not exploited. According to the task description, a sentence indicating the *absence* of a symptom is also relevant; the present pipeline treats all evidence symmetrically, but explicit modelling of polarity could improve the ranking further.

5.4. Comparison with the Depression Task

Although a direct comparison with the 2024 and 2025 depression task results is not possible due to the change of topic and corpus, the gap between the zero-shot ADHD performance reported here (best AP = 0.190 under majority vote) and the supervised depression performance reported in the team's previous internal experiments (P@10 around 0.50) suggests that supervision, when available, remains a strong driver of performance in this family of tasks. The fact that the present zero-shot pipeline still ranks second among 12 teams indicates that the absence of training data shrinks the absolute numbers but not the relative ordering of well-designed systems.

6. Conclusion

This paper described the participation of erisk-cedri in Task 3 of eRisk 2026, focused on sentence-level ranking of evidence for the 18 symptoms of the ASRS-v1.1. In the absence of annotated training data, a zero-shot ensemble pipeline was built that combines multi-query semantic retrieval, NLI-based entailment scoring, keyword matching and cross-encoder re-ranking. The best run (EnsembleV3) achieved AP = 0.190 and ranked second among the 12 participating teams under majority-vote relevance judgements. The ablation across the five submitted runs confirms that no single component is sufficient on its own, and that NLI plays a more central role in the ensemble than its standalone performance

would suggest. The paper also reported on a failed attempt to transfer a supervised depression pipeline to ADHD using synthetic data, and on the decision to pivot to a zero-shot architecture.

For future editions, the natural directions are: systematic optimisation of the NLI hypotheses and query reformulations, explicit modelling of polarity in the relevance criterion, and the use of higher-quality synthetic data (for example, distilled from a much larger language model) as a complement to the zero-shot components rather than a replacement. Now that the official ground-truth labels have been released, a per-symptom error analysis is also planned, in order to identify the items where the pipeline underperforms and to investigate symptom-specific configurations of the ensemble.

Acknowledgments

This work was supported by FCT – Fundação para a Ciência e Tecnologia, I.P. by projects: CeDRI, UID/05757/2025 (DOI: 10.54499/UID/05757/2025) and UID/PRR/05757/2025 (DOI: 10.54499/UID/PRR/05757/2025); SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to: drafting, paraphrasing and grammar and spelling check. 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.

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, 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, Lecture Notes in Computer Science, Springer, 2026.

[3] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: Proceedings of the Conference and Labs of the Evaluation Forum (CLEF 2016), Evora, Portugal, 2016.

[4] 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 (2) (2005) 245–256.

[5] 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 (EMNLP-IJCNLP), 2019, pp. 3982–3992.

[6] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, in: Proceedings of the International Conference on Learning Representations (ICLR 2023), 2023.

[7] A. Williams, N. Nangia, S. R. Bowman, A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2018, pp. 1112–1122.

[8] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, D. Kiela, Adversarial NLI: A New Benchmark for Natural Language Understanding, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 2020, pp. 4885–4901.

[9] W. Yin, J. Hay, D. Roth, Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3914–3923.
[10] R. Nogueira, K. Cho, Passage Re-ranking with BERT, arXiv preprint arXiv:1901.04085 (2019).
[11] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, arXiv preprint arXiv:1611.09268 (2018).
[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, in: Proceedings of the International Conference on Learning Representations (ICLR 2022), 2022.