A Comparative Analysis of In-Context Learning and Fine-Tuning for Biomedical Information Retrieval and Sentence Extraction Using Research Domain Criteria

Athlene V. Jones
University of North Florida
School of Computing
Jacksonville, FL, USA
N00168025@unf.edu

Khanh Linh Lieu
University of North Florida
School of Computing
Jacksonville, FL, USA
N01583210@unf.edu

Indika Kahanda
University of North Florida
School of Computing
Jacksonville, FL, USA
indika.kahanda@unf.edu

Abstract

Research Domain Criteria (RDoC) is a National Institute of Mental Health framework for studying mental disorders by integrating information across genetics, circuits, and behavior. Manually curating biomedical abstracts relevant to RDoC is a significant challenge due to semantically overlapping construct definitions (e.g., "Acute Threat," "Potential Threat," and "Sustained Threat") and the exponential growth of biomedical literature. This study compares two modeling strategies, domain-adapted fine-tuning and in-context prompting, across two RDoC-related subtasks from the official BioNLP-OST 2019 RDoC shared task. For Task 1, unlabeled PubMed abstracts are retrieved and ranked by relevance to eight of the RDoC constructs. We compare a TF-IDF baseline against ModernBERT and Llama (zero-shot and five-shot) using Mean Average Precision (MAP). For Task 2, the objective is to identify the single most relevant sentence from an abstract for a given construct, evaluated using per-construct accuracy. The fine-tuning track performs end-to-end fine-tuning of BioBERT, PubMedBERT, ModernBERT, and RoBERTa using a cross-encoder input format and per-construct grid search. These are compared against the in-context learning of several open-source language models. Both our approaches are competitive against the best-performing team's score from the BioNLP-OST 2019 RDoC shared task. Taken together, these findings suggest that five-shot prompted LLMs and domain-adapted fine-tuned transformers are viable tools for semi-automating the expert annotation in RDoC curation.

1 Introduction

The exponential growth of biomedical literature poses a significant challenge for mental health and psychiatric researchers and clinicians. PubMed searches for foundational concepts yield results at a scale that makes manual review unmanageable. For example, as of April 2026, the keyword "brain" returns over 2.5 million articles, "depression" returns 724k articles, and "anxiety" yields 397k articles. This challenge is further intensified by the National Institute of Mental Health's adoption of the Research Domain Criteria (RDoC) framework (Insel et al., 2010), which marks a paradigm shift from traditional diagnostic models such as the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases (ICD), to a dimensional, biologically-informed approach to understanding mental disorders. RDoC is comprehensive yet complex, encompassing six major domains of human functioning subdivided into constructs and sub-constructs. Because most of the existing biomedical literature reports (or have been curated) within DSM-ICD terminology, it must be systematically remapped to RDoC concepts, a process that is neither practical nor scalable without computational support.

Effective RDoC-based literature curation requires at least two complementary capabilities. The first is information retrieval (IR): given an RDoC construct, identifying and ranking abstracts from a large corpus that are relevant to it. The second is extracting the most relevant evidence, i.e., sentence extraction (SE): given a relevant abstract, pinpoint the specific sentence(s) that most directly support that construct. These capabilities correspond to the two subtasks established at the BioNLP-OST 2019 shared task (Anani et al., 2019) Task 1 (abstract ranking, evaluated via Mean Average Precision) and Task 2 (sentence extraction, evaluated via per-abstract accuracy), which together define a benchmark for end-to-end RDoC-based literature.

This current study evaluates two computational paradigms across the two aforementioned tasks. For Task 1, the *fine-tuning track* compares pre-trained and fine-tuned BERT family encoders as semantic relevance rankers, while the *in-context learning track* evaluates LLaMA 3.1 (8B) under

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 644–655

zero-shot and few-shot strategies.

For Task 2, the *in-context learning track* applies five small open-source large language models: LLaMA 3.1 (8B), Mistral (7B), Qwen 2.5 (7B), Phi-3 (3.8B), and Gemma 2 (9B) under zero-shot and few-shot strategies with dynamic example selection including based-on TF-IDF, BM25, and sentence embeddings; and the *fine-tuning track* trains four pre-trained transformer encoders: BioBERT, PubMedBERT, RoBERTa, and ModernBERT as per-construct cross-encoder classifiers that jointly encode a construct description and a candidate sentence to produce a binary relevance judgment.

By systematically comparing in-context prompting and supervised fine-tuning, we aim to provide actionable guidance for practitioners seeking to automate RDoC literature curation and to assess the extent to which modern LLMs and fine-tuned transformers can help reduce the expert-annotation bottleneck that has historically constrained progress in this domain (Anani et al., 2019). The RDoC benchmark presents particular challenges not addressed in general-domain NLP settings. The relatively small per-construct training sets and highly overlapping construct semantics make it a demanding low-resource evaluation environment.

This paper makes three main contributions: (1) We present the first systematic comparison of in-context learning and supervised fine-tuning across both RDoC Task 1 (abstract ranking) and Task 2 (sentence extraction), using the original BioNLP-OST 2019 evaluation protocol, (2) We show that few-shot prompting with small open-source LLMs achieves competitive performance with domain-adapted transformers, narrowing the gap to the shared-task best system with no parameter updates, and (3) We provide an empirical analysis demonstrating that lexical example selection (TF-IDF/BM25) consistently competes with dense embeddings for construct-specific prompting in the RDoC setting.

2 Related Work

Biomedical literature curation related to mental health and brain sciences remains a recognized bottleneck in scientific research (Anani et al., 2020). While traditional keyword-matching techniques like TF-IDF (Manning et al., 2008) establish strong initial retrieval baselines, domain-specific extraction requires specialized approaches. To address this, the 2019 BioNLP-OST RDoC Task introduced the inaugural benchmark competition to identify the concepts of Research Domain Criteria (RDoC) in biomedical abstracts (Anani et al., 2019). For abstract ranking (Task 1), Chaudhary et al. (Chaudhary et al., 2019) established the top benchmark (MAP 0.86) using a multi-grain neural relevance ranking system that integrated neural topic models with attention-based sentence interactions. For sentence extraction (Task 2), Laden et al. (2020) developed RDoCer, combining TF-IDF document scoring with a supervised ensemble of lexical and structural features. This approach achieved an accuracy of 0.48, ranking first on the *Sustained Threat*.

In the years that followed, research shifted toward domain-specific transformer architectures, notably with Naseem et al. (Naseem et al., 2022), who demonstrated that BioALBERT could surpass several prior supervised benchmarks by leveraging parameter-efficient pretraining on large biomedical corpora. As the field entered the era of Large Language Models (LLMs), Chen et al. (Chen et al., 2024) evaluated the zero-shot capabilities of models such as GPT-4, finding that although these models exhibit deep clinical reasoning, they often require sophisticated prompting to match the ranking precision of specialized systems. However, neither of these studies incorporates RDoC Task data into the benchmarking process.

To the best of our knowledge, no prior work has directly compared in-context prompting with fine-tuned transformer classifiers on both RDoC Task 1 and Task 2 simultaneously, nor has it evaluated dynamic few-shot example selection strategies in this domain. This study fills that gap with a systematic empirical comparison based on the original shared-task evaluation protocol.

3 Methodology

As depicted in Figures 1 and 2, this study employs two experimental tracks, in-context learning and fine-tuning, applied across the two RDoC task formulations. Both tracks are evaluated against the same benchmark dataset using identical train-test splits and official evaluation protocols, but they represent fundamentally different computational paradigms. The in-context track requires no parameter updates and relies on prompt-based reasoning, while the fine-tuning track adapts pre-trained model weights through supervised training on task-specific labeled data.

Figure 1: This pipeline shows the complete workflow for Task 1. From raw inputs (Title, Abstract, Training set, Query with RDoC Construct and NIMH Definition) through baseline modeling (TF-IDF with Cosine Similarity Ranking), implemented IR techniques (Pre-trained vs Fine-tuned BERT family models), implemented reasoning techniques (Zero-shot vs. Five-shot LLaMA Version 3.1), to final evaluation using Mean Average Precision (MAP).

RDoC Construct	Train	+Test	-Test	Gold
Negative Valence System				
Acute Threat (Fear)	39	53	26	53
Potential Threat (Anxiety)	27	124	15	124
Frustrative Nonreward	21	96	48	96
Sustained Threat	18	82	64	82
Loss	28	90	48	90
Arousal & Regulatory Systems				
Arousal	38	97	11	97
Sleep and Wakefulness	48	121	1	121
Circadian Rhythms	47	123	0	123
Total	**266**	**786**	**213**	**786**

Table 1: Dataset distribution across eight RDoC constructs for training and test splits. Task 1 test abstracts include both positive (relevant) and negative (irrelevant) examples; Task 2 test abstracts are gold-labeled at the sentence level.

3.1 Data

This study employs the benchmark dataset from the BioNLP-OST 2019 RDoC Task (Anani et al., 2019), which consists of human-annotated PubMed abstracts covering eight RDoC constructs. These constructs span two domains of the RDoC framework: the Negative Valence System (Acute Threat, Potential Threat, Sustained Threat, Loss, Frustrative Nonreward) and the Arousal and Regulatory Systems (Arousal, Circadian Rhythms, Sleep and Wakefulness). Gold-standard annotations were produced by three independent expert annotators affiliated with the National Alliance on Mental Illness (NAMI) Montana. (Anani et al., 2019).

The structure of the data set differs between the two tasks, each with a dedicated test set. The shared training set contains 266 abstracts annotated as positive examples for their respective constructs, with per-construct counts ranging from 18 (Sustained Threat) to 48 (Sleep and Wakefulness). The Task 1 test set contains 999 abstracts spanning all eight constructs, of which 786 are labeled as construct-relevant (positive) and 213 as construct-irrelevant (negative). The Task 2 test set contains 244 abstracts, each paired with gold-standard, sentence-level labels that mark the single sentence in each abstract that is most relevant to the target construct. Table 1 reports the full per-construct distribution across all three splits (see Table 1).

For the fine-tuning track's Task 2 experiments, the 266 training abstracts were segmented into individual sentences using NLTK's sentence tokenizer (https://www.nltk.org/), yielding 1,570 sentence-level training instances across all constructs (417 relevant and 1,153 non-relevant). A sentence received a positive label (relevant $= 1$) if it occurred as a substring within the gold relevant context annotation for its abstract; otherwise, it was assigned a negative label (irrelevant $= 0$). This programmatic derivation meant that all sentence-level supervision was obtained directly from the existing abstract-level annotations without additional human labeling. The resulting training data is moderately imbalanced, with approximately 73% non-relevant instances, a characteristic that varies across constructs and bears on training stability.

3.2 In-Context Learning Track

Task 1: Abstract Ranking LLaMA 3.1-70B (AI@Meta, 2024) was queried via the Together AI API (https://www.mintlify.com) to

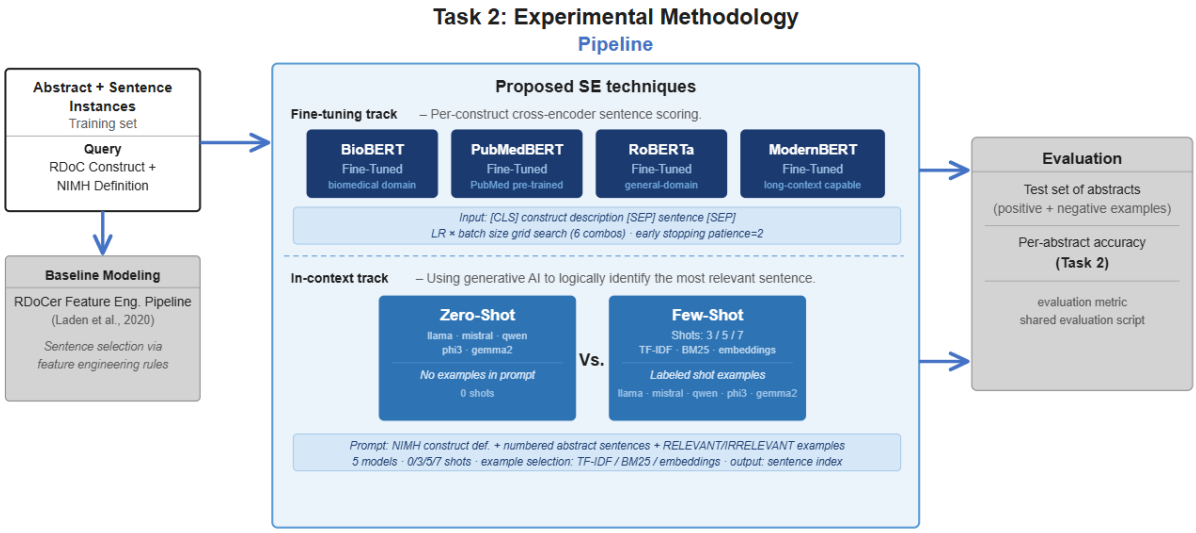

Figure 2: This image depicts the experimental methodology pipeline for Task 2, illustrating two proposed sentence extraction tracks: a fine-tuning track and an in-context learning track comparing zero-shot and few-shot prompting strategies, evaluated using per-abstract accuracy.

assign a relevance score to each test abstract. Two prompting configurations were evaluated. In the zero-shot configuration, the prompt consisted solely of the RDoC construct definition with no labeled examples; five independent trials were conducted, and the results were averaged for robustness. In the five-shot setting, 10 manually selected example abstracts were provided for each construct (five positive and five negative) as in-context demonstrations, enabling the model to discern the linguistic patterns typical of each construct.

Task 2: Sentence Extraction Five small open-source instruction-tuned language models were evaluated locally via Ollama (https://ollama.com/) under zero-shot and few-shot conditions. For each test abstract, the model received a prompt containing the NIMH construct definition and the abstract's sentences enumerated numerically, and was instructed to return a single integer index identifying the most relevant sentence. Under few-shot conditions, labeled RELEVANT and IRRELEVANT sentence examples drawn from the training data were prepended using shot counts of 3, 5, and 7, with demonstrations selected via TF-IDF, BM25, or dense sentence embedding similarity to the test instance. Each configuration was repeated across three random seeds and results were averaged.

3.3 Fine-Tuning Track

Task 1: Abstract Ranking Three BERT-family encoders were evaluated in both pre-trained (frozen) and fine-tuned configurations as abstract relevance rankers, with a binary sequence classification head added to predict whether each abstract is relevant (label $= 1$) or irrelevant (label $= 0$) to a given construct. For fine-tuned variants, a BM25-based pseudo-labeling strategy was first applied to generate construct-specific training subsets. BM25 scores were computed for all training abstracts against an expanded construct query combining the construct name and its NIMH definition. The top-20 abstracts by BM25 score were assigned positive labels; the remaining abstracts received negative labels. This pseudo-labeled subset was used for fine-tuning with early stopping. Pre-trained variants used the same classification head but received no weight updates, serving as a direct comparator.

Task 2: Sentence Extraction. Transformer encoders were fully fine-tuned as per-construct binary classifiers. Rather than encoding the construct and candidate sentence independently, input was formatted using a cross-encoder template that concatenates the NIMH construct definition and the candidate sentence as a single sequence, allowing the model to attend directly across both inputs within a single forward pass. This joint encoding is the core methodological distinction from bi-encoder

or feature-based approaches: the model learns relevance as an interaction between the construct description and the sentence, rather than as a property of either in isolation. A separate classifier was trained for each of the eight RDoC constructs, based on the reasoning that construct-specific fine-tuning is more effective than a single joint model given the semantic distinctiveness of the constructs and the small per-construct training sets.

4 Models

Our approaches were selected to address distinct operational and technical goals within the RDoC annotation pipeline. We chose TF-IDF and BM25 as strong, interpretable lexical baselines to establish a meaningful lower bound for performance. Fine-tuned encoders, including BioBERT, PubMedBERT, ModernBERT, and RoBERTa, were selected to determine if domain-specific pre-training and task-specific supervision could capture subtle semantic distinctions missed by lexical approaches, with the cross-encoder format explicitly conditioning sentence scoring on construct descriptions. Few-shot prompting models evaluated whether instruction-tuned LLMs could generalize RDoC construct identification without task-specific training, representing a low-overhead deployment scenario for resource-constrained clinical settings.

4.1 Baseline

We implement a well-established information retrieval baseline: TF-IDF with Cosine Similarity. For each construct, the query is constructed by concatenating the construct name with its NIMH RDoC construct definition, and all test abstracts are ranked by cosine similarity in the TF-IDF vector space. TF-IDF requires no training and relies solely on keyword overlap between abstracts and construct definitions.

4.2 BM25

BM25 extends TF-IDF with term frequency saturation and document length normalization, producing more stable relevance scores across abstracts that vary substantially in length. In this study, BM25 fulfills two functions: it acts as the retrieval baseline for Task 1 and also provides pseudo-labels used as training supervision for the fine-tuned Task 1 models and for in-context prompting in Task 2.

4.3 In-context Learning Track

4.3.1 Task 1: Abstract Ranking

LLaMA 3.1-70B (AI@Meta, 2024) was queried via the Together AI API (`https://api.together.xyz/`). Its 128K-token context window enables processing of full abstracts alongside construct definitions and in-context examples without truncation. However, its limited exposure to RDoC-specific psychiatric terminology may constrain its ability to distinguish conceptually overlapping constructs, particularly the Threat cluster (Acute Threat, Potential Threat, Sustained Threat).

4.3.2 Task 2: Sentence Extraction

Five open-weight instruction-tuned models were evaluated locally via Ollama: LLaMA 3.1 (8B) (Dubey et al., 2024), Mistral (7B) (Jiang et al., 2023), Qwen 2.5 (7B) (Team, 2024), Phi-3 (3.8B) (Abdin et al., 2024), and Gemma 2 (9B) (Team et al., 2024). These models were selected to represent a range of architectures, parameter scales, and training corpora at the small open-source frontier. LLaMA 3.1 uses grouped query attention with a 128K context window; Mistral employs sliding window attention for memory efficient inference; Qwen 2.5 was pre-trained on over 18 trillion tokens with strong scientific coverage; Phi-3 prioritizes high-quality curated training data over scale; and Gemma 2 uses alternating local and global attention with knowledge distillation from a larger model.

4.4 Fine-tuning Track

4.4.1 Task 1: Abstract Ranking

Three BERT encoders were evaluated in both pre-trained and fine-tuned configurations: ModernBERT (Warner et al., 2024), PubMedBERT (Gu et al., 2021), and BioBERT (Lee et al., 2020). Pre-trained variants compute cosine similarity between construct definition and abstract representations without weight updates. Fine-tuned variants are adapted using BM25 pseudo-labeled training data.

4.4.2 Task 2: Sentence Extraction

The fine-tuning track approached Task 2 as a per-construct binary sentence classification problem. Four pre-trained transformer models were fine-tuned independently: BioBERT (**?**), PubMedBERT (Gu et al., 2021), RoBERTa (Liu et al., 2019), and ModernBERT (Warner et al., 2024), representing biomedical domain-specific, PubMed-specific, general-domain, and

long-context architectures, respectively. All four were fine-tuned using the Hugging Face `AutoModelForSequenceClassification` head with two output labels.

BioBERT and PubMedBERT provide biomedical domain specialization through continued and from-scratch pre-training on PubMed and PMC literature, with BioBERT's retention of the NSP objective offering additional structural alignment with the cross-encoder format. RoBERTa serves as a general-domain baseline, while ModernBERT represents an architecturally modern encoder with an extended 8,192-token context window, trading biomedical depth for up-to-date design choices.

5 Experimental Setup

5.1 Task 1: Abstract Ranking

All models were evaluated on the test set using Mean Average Precision (MAP) as the primary metric. MAP is considered a standard metric for information retrieval tasks, computed as the macro-average across all RDoC constructs. This metric assesses the quality of retrieved abstract rankings, with emphasis on early retrieval of relevant abstracts.

Following the original RDoC Task (Anani et al., 2019), strict separation between training data (Set 1) and test data (Set 2) was maintained, with test data containing positive and negative examples in all constructs examined. 20% of the training data was used as a validation set for parameter tuning. Mean Average Precision (MAP) is computed per construct and then macro-averaged, capturing both precision and ranking quality, which is crucial for real-world curation workflows.

Baseline Comparison: Models were compared against a TF-IDF + cosine similarity baseline, using construct definitions as queries. This baseline represents the state of practice in many curation pipelines. LLaMA inference uses standard sampling to balance consistency and diversity. Both zero-shot and few-shot prompting strategies leverage the model's weights without gradient-based updates. LLaMA was used with default parameters. For few-shot, the optimal number of examples were chosen using a validation set.

For the BERT models, a hyperparameter search algorithm systematically evaluated all combinations of learning rates {1e-5, 2e-5, 3e-5} and epoch counts (2, 3, 4) for each construct independently, resulting in 9 total hyperparameter configurations. Validation performance on construct-specific validation sets guided the selection of the following optimal hyperparameters: Max Sequence Length: 512 tokens, Batch Sizes: 8 (training), 16 (evaluation), Weight Decay: 0.01, Warmup Steps: 50, Learning Rate: Hyperparameter search space (primary: 1e-5 to 3e-5), Number of Epochs: Hyperparameter search space (primary: 2-4 epochs), Early Stopping: EarlyStoppingCallback with patience=2, Optimizer: AdamW.

5.2 Task 2: Sentence Extraction

For the in-context track, five models were evaluated under zero-shot and few-shot conditions. In the zero-shot condition, no labeled examples were included. In the few-shot condition, shot counts of 3, 5, and 7 were evaluated, with examples drawn from the training data using three selection strategies: TF-IDF retrieval, BM25 retrieval, and dense sentence embedding similarity. Each few-shot configuration was repeated on three random seeds, and the results were averaged. The complete experimental grid consisted of 5 models $\times 3$ shot counts $\times 3$ selection strategies $\times 3$ seeds = 135 few-shot runs, plus 5 zero-shot runs.

The prompt structure followed a consistent template across all conditions. The construct definition was provided as a task context block, followed by the abstract sentences enumerated numerically, with the instruction to return a single integer. In the few-shot condition, Example sentences labeled as RELEVANT or IRRELEVANT were inserted with the construct definition and the abstract. (See Figure 5 in the Appendix for a prompt example).

Performance was measured using the official per-abstract accuracy metric via script_task2.py (Anani et al., 2019). A prediction is correct if the selected sentence appears as a substring of the gold Relevant Context annotation. Macro-average accuracy across the 8 constructs is the primary metric.

For the fine-tuning track, Four transformer encoders were fine-tuned per-construct using the Hugging Face Trainer (Wolf et al., 2020) on a Google Colab A100 GPU with mixed-precision (fp16) enabled and a global random seed of 42. For each construct, sentence-level training instances were partitioned 80/20 into training and validation subsets using stratified sampling on the relevance label. Hyperparameters were selected via a grid search over three learning rates $\{1\times10^{-5}, 2\times10^{-5}, 5\times10^{-5}\}$. and two batch sizes (8, 16), yielding six combinations per construct. Early stopping with a patience

of 2 (max epochs of 10) was applied to every grid search run, evaluated once per epoch against validation accuracy. The combination achieving the highest validation accuracy was selected and used to retrain the model on the full training set, again with early stopping.

All runs used AdamW optimization with weight decay of 0.01 and no learning rate warmup. The maximum input sequence length was set to 128 tokens with padding and truncation applied where necessary. During inference, the fine-tuned classifier produced a relevance probability score (the softmax probability of class 1) for each candidate sentence in a test abstract, and the sentence with the highest score was chosen as the prediction. Predictions were evaluated using the official per-abstract accuracy metric via `script_task2.py`, and macro-average accuracy across all eight constructs is the primary reported metric.

6 Results and Discussion

6.1 Task 1: Abstract Ranking

As shown in Figure 3, evaluation across eight distinct approaches revealed substantial differences in performance. The biomedically pre-trained models proved to be competitive but not superior. Across our models, the PubMedBERT models, both the pre-trained and fine-tuned, achieved the highest performance at 0.83 MAP, along with the pre-trained BioBERT. The ModernBERT models performed similarly with MAP scores of 0.82 for the pre-trained and 0.80 for the fine-tuned approach. As for the in-context models, prompting with LLaMA 3.1-70B achieved the lowest performance at 0.79 MAP (zero-shot) and 0.76 MAP (few-shot).

The performance decline from zero-shot to five-shot LLaMA is attributed to selection bias in the manually chosen demonstrations and to lexical overlap within the Threat construct cluster (Acute, Potential, and Sustained Threat), where few-shot examples from one construct inadvertently reinforce vocabulary from the others. In the zero-shot setting, the model relies solely on the NIMH definition, which provides a cleaner disambiguation signal. This suggests that construct-aware, automated example selection is needed for semantically overlapping constructs.

Both our generative reasoning with few examples and discriminative fine-tuning approaches achieve comparable performance on this task. The pre-trained BERT models demonstrated that even without task-specific fine-tuning, the model captures sufficient semantic information for construct classification, where supervised adaptation may not improve results.

Relatively high baseline (TF-IDF) MAP values indicate that, despite high semantic overlap among RDoC constructs, keyword-based retrieval remains competitive. The results demonstrate that while modern general-purpose models are competitive, they do not yet surpass the specialized state-of-the-art for this retrieval task. The RDoC Task's (Anani et al., 2019) best performer (T30) maintains the lead with a MAP of 0.86.

Our models, e.g., ModernBERT, achieve near-perfect results on biologically distinct topics such as Sleep & Wakefulness (0.99) and Potential Threat (0.93), but the performance collapses on overlapping concepts such as Sustained Threat (0.58) and Loss (0.69). This suggests that models effectively capture distinct medical vocabulary but struggle to disentangle nuanced psychological traits such as distinguishing chronic stress from anxiety, dragging down the overall average despite high proficiency in the clearer biological categories.

6.2 Task 2: Sentence Extraction

Neither paradigm, in-context, or fine-tuning reached the RDoCer (Laden et al., 2020), i.e. RDoC Tasks's best (T30) of 0.58, although BioBERT came closest at 0.54. The poor performance of RoBERTa, together with the comparable results of BioBERT and PubMedBERT, suggests that while domain-specific pre-training is required, it alone does not explain performance. Instead, the cross-encoder setup and the use of the construct description as the query appear to be the main methodological factors driving effectiveness.

Fine-tuning The two biomedical domain-specific models produced the strongest results and were nearly identical in performance: BioBERT achieved a macro-average accuracy of 0.543 and PubMedBERT 0.541, both clearing the RDoCer baseline (0.478) by approximately +0.064 and falling within 0.037 of the RDoC Task's best performer (T30) with 0.580. The convergence of these two models supports the conclusion that domain-specific pre-training is the primary differentiating factor, regardless of whether the model is initialized from general BERT weights (BioBERT) or trained from scratch on biomedical text (PubMedBERT). ModernBERT achieved

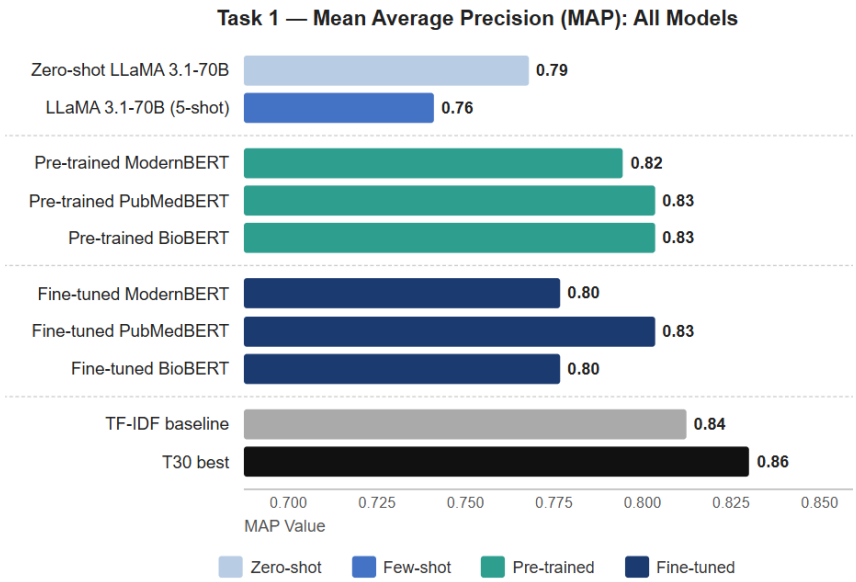

Figure 3: Task 1 MAP results comparing

0.485, marginally clearing the RDoCer baseline (+0.007) but sitting 0.095 below T30. Despite its architectural advantages, its limited biomedical pre-training appears to constrain performance.

RoBERTa was the weakest model at 0.469, falling below both the RDoCer baseline (-0.009) and T30 (-0.111). In four of the eight constructs, the model experienced class collapse: for Arousal, Potential Threat (Anxiety), Acute Threat (Fear), and Frustrative Nonreward, it labeled every test instance as belonging to the non-relevant class. This may be attributable to the removal of the Next Sentence Prediction objective during pre-training, combined with moderately imbalanced training data (73% non-relevant), which together allow the optimizer to converge on a majority-class shortcut. RoBERTa's results should be interpreted with caution, and they are reported for completeness rather than as a reliable performance estimate. Across all models, Acute Threat (Fear) was consistently the hardest construct and Circadian Rhythms the easiest, mirroring patterns from the original 2019 RDoC shared task and consistent with training set size differences across constructs (see Figure 6 in the Appendix) for construct-wise performance.

In-context prompting Zero-shot performance across the five LLMs ranged from 0.380 (LLaMA 3.1 8B) to 0.482 (Gemma 2 9B). All five models fell below the T30 benchmark of 0.580 under zero-shot conditions, and only Gemma 2 exceeded the baseline of 0.40. The spread of more than 10 percentage points between the best and worst zero-shot models suggests that instruction-following capability and model capacity vary meaningfully at this parameter scale for the identification of sentences at the construct-level.

Few-shot prompting produced consistent improvements across all five models, with an average gain of +0.034 over zero-shot. The best overall performance came from Gemma 2 (9B) using 7-shot TF-IDF-based example selection, reaching a macro-average accuracy of 0.517. This is the strongest in-context result and represents a substantial gain over the baseline (+0.117), although it remains 0.063 behind T30. LLaMA 3.1 (8B) showed the largest individual few-shot gain of +0.070, rising from the lowest zero-shot accuracy (0.380) to 0.450 with 3-shot TF-IDF selection, suggesting it is particularly sensitive to prompt context despite weaker zero-shot construct understanding.

Example selection strategy mattered across models. TF-IDF retrieval produced the best few-shot result overall and BM25 was similarly competitive, while sentence embedding selection consistently underperformed both lexical methods. This finding is non-trivial: semantic similarity did not outperform lexical matching for construct-specific example retrieval. We distinguish this finding from prior work on dense retrieval (e.g., MedCPT) (Jin et al., 2023), which targets corpus-level document ranking. Our comparison concerns in-context demonstration selection, where TF-IDF and BM25 excel

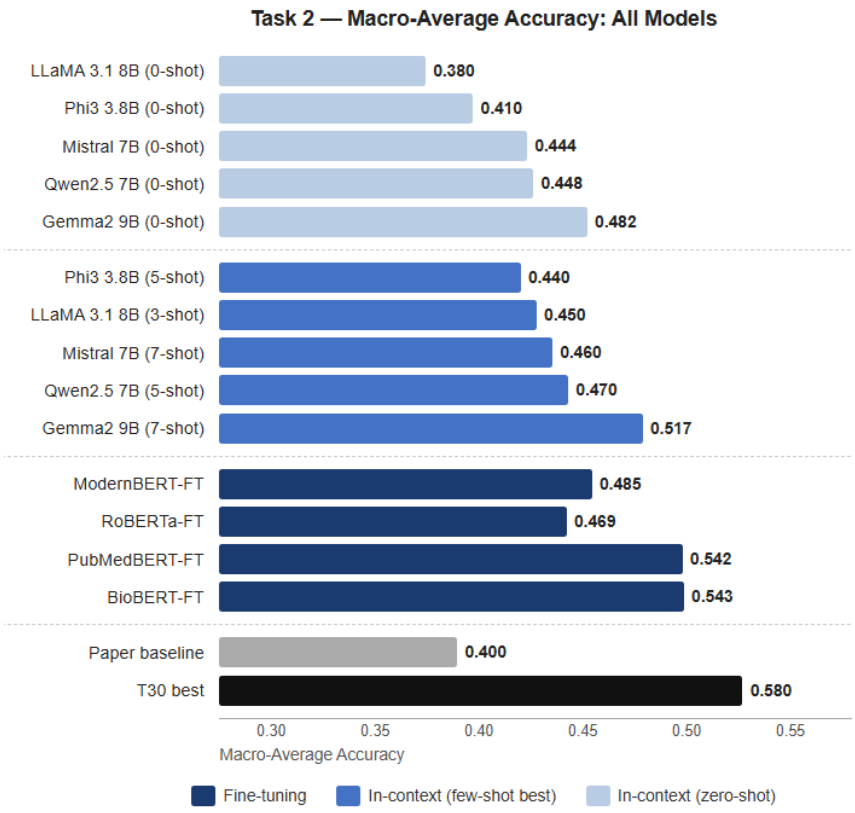

Figure 4: Task 2 Macro-Average Accuracy results

by directly matching National Institute of Mental Health (NIMH) construct-specific terminology through keyword overlap. Conversely, dense embeddings retrieve thematically adjacent examples that often lack this precise, construct-marking vocabulary, ultimately producing less discriminative prompts in this setting.

The consistent underperformance of dense embedding selection (all-MiniLM-L6-v2) across 4 out of 5 models supports this interpretation. For example, TF-IDF retrieves Arousal demonstrations containing terms like 'startle reflex' and 'locomotor activity,' while dense embeddings retrieve sentences from adjacent constructs (e.g., Sleep and Wakefulness) due to overlapping themes, weakening the prompt signal. A similar cross-construct confusion occurs in the Threat cluster, where embedding-selected examples blur the boundary between Sustained Threat and Potential Threat through shared 'chronic stress' vocabulary.

7 Conclusion and Future Work

The results demonstrate that In-context Learning and Fine-tuning offer complementary pathways for RDoC construct-based literature retrieval and extraction, each with distinct strengths. For task 1, PubMedBERT provides marginal performance gains over ModernBERT and BioBERT. LLaMA prompting, despite its relatively low MAP values, offers notable data efficiency, rapid adaptability to new constructs, and reasoning transparency through natural-language explanations, which are critical advantages for applications that require construct-specific justifications. The observed relative performance decline from zero-shot to five-shot LLaMA indicates that construct-aware prompt engineering warrants further investigation.

Future work should focus on construct-specific fine-tuning with more strategic example selection against pretrained models to better capture domain-specific terminology and psychiatric construct language. In addition, conducting an in-depth per-construct performance analysis for both In-context Learning and Fine-tuned models using an ensemble of LLMs would enable a granular comparison of construct-specific strengths and weaknesses across the two approaches.

8 Limitations

This study uses the only known curated dataset related to RDoC via the gold-standard abstracts from the RDoC Task, which was held in 2019. Therefore, our study is restricted by the limited nature of this data. For example, this dataset, is only a single, 6-year-old snapshot of the RDoC Landscape; the RDoC Matrix itself has changed (new constructs like "Sensorimotor Systems" have been added, though no constructs have been removed), not to mention the exponential growth of RDoC-related literature in databases such as PubMed since 2019.

Key challenges for developing methods included the limited training data for certain constructs and the computational cost of hyperparameter tuning with LLMs, which require consistent GPU access and extended training time (that limited our ability use larger models with a very high number of parameters). Task 1 fine-tuning was further constrained by the structure of the available training data, which provides only binary positive or negative abstract assignments per construct rather than ranked relevance judgments. Because this is insufficient to train a ranking model directly, we approximated relevance scores using BM25 as a pseudo-labeling strategy; while this allowed us to apply fine-tuned encoders to the retrieval task, the resulting label noise likely suppresses performance relative to what gold-ranked supervision would yield. We report Task 1 results for completeness but do not claim improvement over the TF-IDF baseline, and we identify the absence of ranked training labels as the primary bottleneck for future work on this task. Another limitation is the lack of construct-specific fine-tuning with more strategic example selection to better capture domain-specific terminology and psychiatric construct language. Multi-task learning was avoided due to severe per-construct imbalance.

In addition, open-source BERT-family models and LLaMA do not represent the full capabilities of the current generation of LLMs (especially their commercial counterparts like GPT), which appear to be improving rapidly. Similarly, the implemented IR techniques do not capture some of the more recent IR innovations like retrieval augmented generation (RAG). In addition, dense retrieval methods such as MedCPT (Jin et al., 2023) were not evaluated, specifically on Task 1, representing a meaningful gap for future work. Lastly, we do not utilize the inherent hierarchical structure of the RDoC Matrix nor the construct-specific matrix elements beyond the construct definitions.

9 Ethical considerations

This work involves automated retrieval of biomedical abstracts related to mental health research using the Research Domain Criteria (RDoC) framework. Our work utilizes publicly available PubMed abstracts and the RDoC shared task dataset, which contains only de-identified scientific literature without patient information or clinical records. No individual-level health data was used in this research. While our system retrieves abstracts related to mental health constructs, we emphasize that the RDoC framework is designed for research purposes to advance understanding of mental disorders through dimensional, cross-cutting approaches. Our work does not diagnose, treat, or make claims about individuals with mental health conditions. We acknowledge the potential for language models to perpetuate stigmatizing language or biases present in the biomedical literature and recommend careful human oversight when deploying such systems in practice.

This study focuses on information retrieval for research purposes and is not intended for direct clinical decision-making. Any future deployment of LLM-based retrieval systems in clinical settings would require rigorous validation, regulatory approval, and appropriate safeguards to ensure patient safety. Clinicians and researchers should critically evaluate retrieved abstracts rather than relying solely on automated rankings. We compare open-source models to promote transparency and reproducibility in biomedical NLP. However, we recognize that LLMs may produce errors, hallucinations, or retrieve irrelevant content. Users of such systems should be trained to identify limitations and maintain appropriate skepticism of automated retrieval results.

Improved retrieval of mental health research literature has the potential to accelerate scientific discovery and improve understanding of mental disorders. However, we acknowledge that advances in automated text mining could also facilitate misuse, such as cherry-picking literature to support predetermined conclusions. We advocate for responsible use of these tools within established evidence-based, scientific and ethical frameworks.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

AI@Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mohammad Anani, Nazmul Kazi, Matthew Kuntz, and Indika Kahanda. 2019. RDoC task at BioNLP-OST 2019. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 216–226, Hong Kong, China. Association for Computational Linguistics.

Mohammad Anani, Matt Kuntz, and Indika Kahanda. 2020. Brret: Retrieval of brain research related literature. *AMIA Summits on Translational Science Proceedings*, 2020:53.

Yishai Chaudhary, Pankaj Gupta, and Hinrich Schütze. 2019. Bionlp-ost 2019 rdoc tasks: Multi-grain neural relevance ranking using topics and attention based query-document-sentence interactions. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 227–236. Association for Computational Linguistics.

Qingyu Chen, Jingcheng Du, Yan Hu, and 1 others. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, 171:108189.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. In *ACM Transactions on Computing for Healthcare (HEALTH)*, volume 3, pages 1–23. ACM.

Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. 2010. Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7):748–751.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Daniel Laden, Shyaman Jayasundara, and Indika Kahanda. 2020. Rdocer: Information retrieval and sentence extraction for mental health using research domain criteria. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 226–229.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. In *Bioinformatics*, volume 36, pages 1234–1240. Oxford University Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Usman Naseem, Matloob Khushi, Shahadat Uddin Khan, Kamran Shaukat, and Mohammad Ali Moni. 2022. Benchmarking for biomedical natural language processing tasks with a domain specific albert. *Future Generation Computer Systems*, 129:144–153.

Gemma Team, Morgane Riviere, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix: Prompting Example

Figure 5 shows an example of the few-shot prompting format used for both Task 1 and Task 2 experiments.

Sample Prompt Preview

Which sentence provides the strongest evidence that this abstract discusses Arousal?

Definition: Arousal is a continuum of sensitivity of the organism to stimuli, both external and internal. Arousal facilitates interaction with the environment in a context-specific manner (e.g., under conditions of threat, some stimuli must be ignored while sensitivity to and responses to others is enhanced, as exemplified in the startle reflex). It can be evoked by either external/environmental stimuli or internal stimuli (e.g., emotions and cognition). It varies along a continuum that can be quantified in any behavioral state, including wakefulness and low-arousal states including sleep, anesthesia, and coma.

Key indicators: sensitivity to stimuli, startle reflex, behavioral state, locomotor activity, homeostatic drives, arousal level.

Here are examples of **RELEVANT** and **IRRELEVANT** sentences for this construct:

RELEVANT examples:
(1) "valence, arousal), and neuroscience has more recently sought the neurobiological basis of emotions via functional neuroimaging."
(2) "In the present study, we sought to identify intensive valence and arousal affective states via facial EMG activity."
(3) "CONCLUSIONS: This study is the first to identify arousal-specific retrospective affect report discrepancies over time and suggests retrospective reports also reflect personality differences in affective self-knowledge."

IRRELEVANT examples:
(1) "Seventy-seven students ages 19–30 participated in the study."
(2) "Moral standards, as well as conservative beliefs regarding sexuality, are believed to be involved in the etiology and maintenance of this syndrome."
(3) "However, in a recent study (Ehlers et al."

Now identify the relevant sentence(s) in this abstract:

Sentences:
(1) Previous studies have shown that music is a powerful means to convey affective states, but it remains unclear whether and how social context shape the intensity and quality of emotions perceived in music.
(2) Using a with . . .

Total sentences in abstract: 8

Figure 5: Example of the few-shot prompting template used for RDoC construct classification. The prompt includes the task instruction, construct definition with key indicators, example sentence–label pairs (shots), and the target abstract for sentence identification.

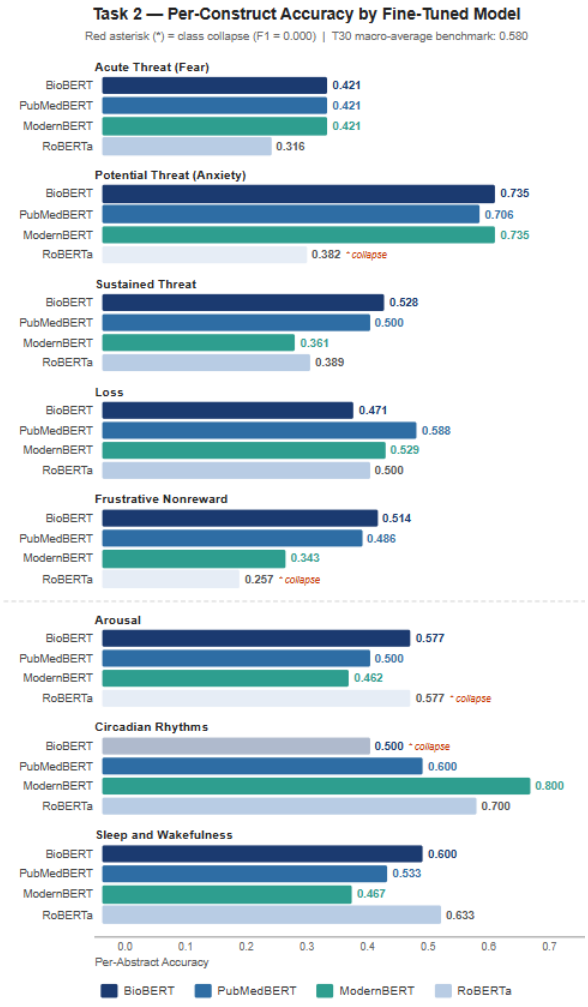

Figure 6: Task 2 Per-construct accuracy for each fine-tuned model across the eight RDoC constructs.

B Appendix: Per-construct Performance