RSPC: A Benchmark for Modeling Stress and Psychiatric Conditions in Digitally Mediated Relationships using Psychiatrist Annotations

Anonymous ACL submission

Abstract

In NLP, mental health conditions are often modeled as isolated phenomena, without interpersonal context. We use Reddit posts about long-distance relationships to capture both mental health distress and associated relational triggers. We introduce the Relational Stress and Psychiatry Corpus (RSPC) containing 1,799 Reddit posts annotated by psychiatrists for diagnostic categories, including the most prevalent mood disorders (anxiety and depression), relational stressor triggers, and indications of relationship phase. We benchmark seven fine-tuned transformer models and five large language models across multi-label disorder classification, relational trigger detection, and temporal phase prediction tasks. We find clear task-dependent differences between model families, with Claude-3-Haiku achieving the best disorder classification performance (Macro-F1 = 0.538) and GPT-4o obtaining the strongest relational trigger detection performance (Macro-F1 = 0.519), suggesting distinct model capabilities. We further find strong associations between anxiety disorders and chronic relational uncertainty. Overall, RSPC establishes a benchmark for NLP tasks that consider relational context and supports a shift from individual-centric to context-aware mental health modeling that captures the social and temporal dynamics of distress.

1 Introduction

As international education and work demands grow, digital communication platforms increasingly shape and mediate intimate relationships, influencing attachment, conflict, emotional support, and experiences of separation. However, computational mental health research predominantly models psychiatric distress as an individual-level phenomenon, often overlooking the interpersonal contexts in which symptoms arise. Long-distance relationships (LDRs) offer a salient setting for studying relationally mediated distress, as communication primarily occurs through digital channels during prolonged physical separation. Clinical studies associate LDR-related stress with heightened anxiety, adjustment difficulties, depression, and insomnia (Stafford, 2004; Neustaedter and Greenberg, 2012). These psychiatric manifestations commonly emerge through recurring relational stressors, including communication silences, commitment ambiguity, jealousy, and reunion-separation cycles.

Existing mental health NLP benchmarks largely reduce psychiatric distress to coarse binary classification (Coppersmith et al., 2015; Losada et al., 2018; Yates et al., 2017; Cohan et al., 2018), despite DSM-5-TR defining diverse mood and anxiety disorder subtypes across a spectrum of severity. Prior datasets predominantly annotate distress at the user level, overlooking the relational contexts in which symptoms emerge.

Unlike prior mental health NLP datasets, which frequently rely on coarse crowdsourced annotations from platforms such as MTurk, RSPC employs a clinically grounded annotation framework developed in collaboration with a team of four licensed psychiatrists from Andhra University. The annotation schema is aligned with DSM-5-TR (Association, 2022) and ICD-11 (World Health Organization et al., 2018) diagnostic criteria, enabling clinically informed inference of psychiatric symptoms, relational stressors, and temporal relationship phases. Each Reddit post was independently annotated by four trained annotators, with disagreements resolved through adjudication, resulting in substantial inter-annotator agreement and high annotation reliability across all annotation tiers.

With these annotations, we introduce the **Relational Stress and Psychiatry Corpus (RSPC)**, comprising 1,799 Reddit posts from long-distance relationship communities annotated for DSM-5-TR/ICD-11-aligned psychiatric symptom categories, relational stressors, and temporal relationship phases. Our contributions are threefold: (1)

we introduce the first benchmark linking clinically grounded psychiatric categories with relational stressors and temporal phases in digitally mediated relationships, (2) we benchmark transformer models and LLMs across disorder classification, trigger detection, and temporal reasoning tasks, and (3) we demonstrate that relational context provides a measurable signal for psychiatric inference in digital communication environments.

2 Related Work

The recent work on **Mental Health Detection in Social Media** has focused primarily on detecting depression, suicidality, and anxiety from social media language (Coppersmith et al., 2015; Yates et al., 2017; Cohan et al., 2018; Losada et al., 2018). Earlier approaches relied on lexical features, whereas recent work has adopted transformer architectures (Nadeem, 2016; Harrigian et al., 2021). Large language models have further expanded the scope of zero-shot psychiatric inference. Existing benchmarks, however, largely model psychiatric distress as an individual-level attribute detached from interpersonal context. Relational dynamics such as communication disruption, attachment insecurity, and jealousy remain underrepresented despite their clinical relevance.

Relational and Contextual Modeling. Prior work has explored conversational dynamics (Alghowinem et al., 2016), online social support (De Choudhury et al., 2013), emotional disclosure (Pavlova and Berkers, 2020), and longitudinal behavioral change (MacAvaney et al., 2018; Sekulić et al., 2018). However, these approaches primarily model chronological behavior rather than event-contingent relational dynamics. Existing social-context modeling relies largely on generic interaction features rather than clinically meaningful interpersonal stressors, such as communication gaps or commitment ambiguity. No prior work explicitly links relational stressors with DSM-5-TR psychiatric symptom categories.

Clinical Grounding and Ethics. Recent studies highlight limitations in clinical grounding, including noisy self-disclosure labels, inconsistent annotations, and weak alignment with diagnostic frameworks (Chancellor et al., 2019; Harrigian et al., 2021). Researchers increasingly advocate for clinically informed annotation protocols and collaboration with mental health professionals (Low et al., 2020; Zirikly et al., 2019). Prior datasets, however, predominantly rely on coarse binary depression labels rather than differentiated symptom-level psychiatric categories in relational settings.

Large Language Models for Mental Health. Recent work explores LLM prompting for psychiatric classification, counseling simulation, and risk detection (Guo et al., 2024; Gilardi et al., 2023). Prompted LLMs often perform strongly on socially contextual tasks, but their behavior under clinically grounded multi-label psychiatric inference remains underexplored. Existing work also lacks a systematic comparison between fine-tuned transformers and prompted LLMs for relationally grounded psychiatric reasoning.

Research Gap & Positioning of RSPC: Existing mental health NLP research predominantly models psychiatric distress as an individual-level phenomenon using coarse diagnostic labels, with limited consideration of the relational and interpersonal dynamics through which symptoms emerge. Prior work on social-context modeling focuses mainly on generic interaction patterns or longitudinal behavior, while clinically meaningful relational stressors such as communication disruption, attachment insecurity, and commitment ambiguity remain underexplored. Furthermore, existing benchmarks lack DSM-5-TR-aligned multi-label psychiatric annotations and systematic evaluation of LLM-based relational psychiatric reasoning. RSPC addresses these gaps through clinically grounded relational-context modeling and comprehensive benchmarking of transformers and LLMs for psychiatric symptom inference.

3 RSPC Construction

This section presents the methodology and data-collection pipeline used to construct RSPC as described in Fig. 1.

3.1 Dataset Collection & Construction

We introduce the **Relational Stress and Psychiatry Corpus (RSPC)**, a benchmark for studying psychiatric symptom expression in digitally mediated romantic relationships. The dataset comprises publicly available Reddit posts from long-distance relationship communities (`r/LongDistance`, `r/LDR`) collected between January 2020 and December 2023 and filtered for narrative completeness, relational relevance, and linguistic consistency. All usernames, personal identifiers, and proper nouns were anonymized using placeholder tokens (e.g.,

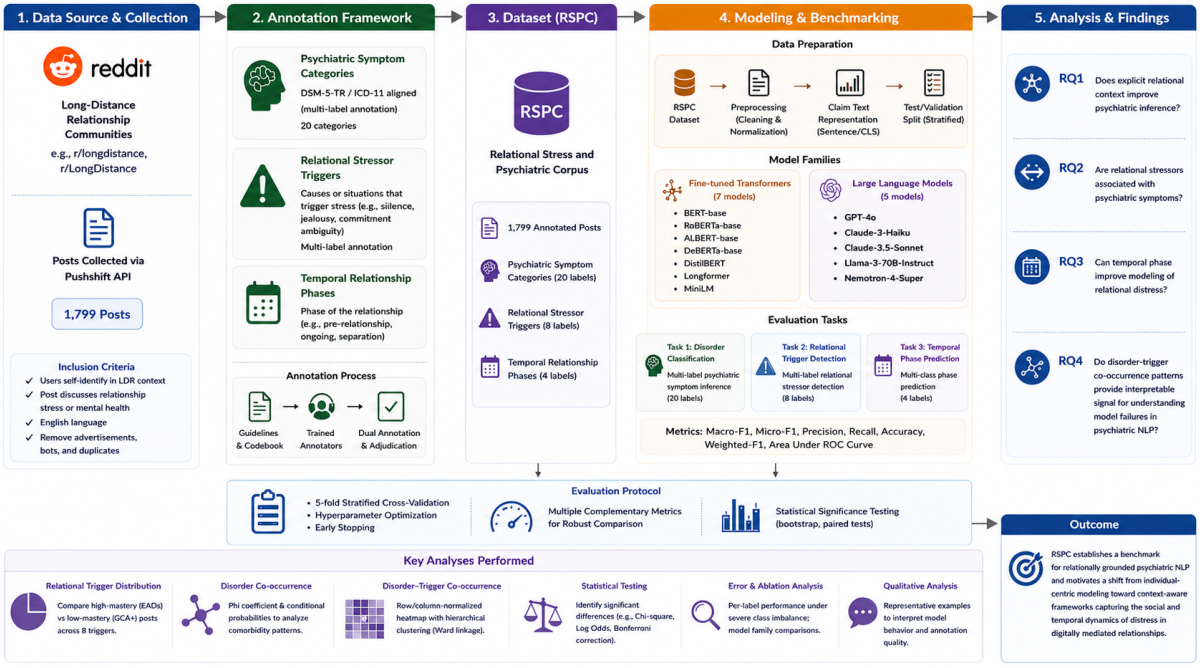

Figure 1: Workflow diagram of RSPC

[USER], [PLACE]). Consistent with prior mental health NLP research, only publicly accessible posts were used, with no attempts made to infer user identities or contact individuals directly.

3.2 Data Annotation & Guidelines

The annotation framework was developed in consultation with a team of four licensed psychiatrists from Andhra University and grounded in DSM-5-TR (Association, 2022) and ICD-11 (World Health Organization et al., 2018) diagnostic criteria. Each Reddit post was independently annotated by two trained annotators using a clinically informed coding manual, with disagreements resolved through adjudication. Inter-rater reliability, measured using Cohen's κ, demonstrated substantial agreement across all annotation tiers, including psychiatric symptoms (0.78), relational stressors (0.72), and temporal relationship phases (0.81).

1. **Tier A: Psychiatric Symptom Categories.** Posts were annotated for five clinically grounded symptom categories: Major Depressive Disorder (MDD), Generalized Anxiety Disorder (GAD), Separation Anxiety Disorder (SAD), Adjustment Disorder (ADJ), and Insomnia, using DSM-5-TR/ICD-11-aligned criteria adapted for textual inference.

2. **Tier B: Relational Stressor Triggers.** Posts were annotated for relational stressors, including Commitment Ambiguity, Lack of Communication, Reunion/Separation Stress, Trust/-Fidelity Issues, Jealousy/Insecurity, Silence Gaps, Social Media Surveillance, and Time-zone Misalignment.

3. **Tier C: Temporal Relationship Phase.** Posts were assigned to one of four temporal phases: SEPARATION, ANTICIPATION, REUNION, or UNKNOWN, modeling event-contingent rather than chronological time.

3.3 Dataset Statistics

The final corpus contains 1,799 Reddit posts, split into train/validation/test partitions with a 70:10:20 stratified ratio. Adjustment Disorder (74.5%) and GAD (71.1%) are the dominant psychiatric categories, while MDD (17.1%) and Insomnia (1.2%) remain sparse. Commitment Ambiguity (66.2%) and Lack of Communication (61.7%) are the most common relational stressors. Temporal phase labels are dominated by SEPARATION (65.1%). Full label distributions and co-occurrence statistics are provided in Appendix 7.

3.4 Inter-annotator Agreement Scores

Inter-annotator agreement was evaluated using Cohen's κ across all annotation tiers, demonstrating substantial agreement for psychiatric symptoms, relational triggers, and temporal phase annotations.

Detailed agreement analysis and tier-wise scores are provided in Appendix 7.

4 Methodology

This section presents the major experiments conducted on **RSPC**, involving LLM- and PLM-based approaches. Detailed experimental settings and results are described in the following subsections.

4.1 Models and Experimental Setup

We evaluate RSPC across three benchmark tasks: multi-label psychiatric symptom classification, relational trigger detection, and temporal phase classification. Experimental comparisons examine supervised representation learning against instruction-following inference under clinically grounded relational settings. Seven transformer architectures are benchmarked, including BERT-base (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), ClinicalBERT (Alsentzer et al., 2019), BART-base (Lewis et al., 2020), T5-base (Raffel et al., 2020), Longformer (Beltagy et al., 2020), and BigBird-RoBERTa (Zaheer et al., 2020). Additionally, five prompted large language models are evaluated: GPT-4o (Achiam et al., 2023), Claude-3-Haiku (Anthropic, 2024), Qwen-2.5-72B (Hui et al., 2024), LLaMA-3-70B (Touvron et al., 2023), and Nemotron-Super (Adler et al., 2024). These experiments analyze contextual reasoning, long-document understanding, and clinically informed relational inference across heterogeneous architectures within RSPC benchmarks.

4.2 Hyperparameters

Transformer architectures were fine-tuned independently for each benchmark task using the AdamW optimizer (Loshchilov and Hutter, 2017). Training used a learning rate of 2×10^{-5}, a batch size of 16, and early stopping based on validation Macro-F1 performance. Multi-label classification tasks utilized sigmoid activation with binary cross-entropy loss and inverse-frequency class weighting, psychiatric symptom and relational trigger categories. Fine-tuning procedures were standardized across all transformer encoders to ensure fair architectural comparisons. Additional implementation details, optimization settings, and task-specific hyperparameters are provided in Appendix 7 for complete experimental reproducibility and transparency.

4.3 Prompts & Prompting Strategies

Large language models were evaluated under deterministic prompting configurations using temperature $T = 0.0$ to minimize stochastic output variability across experimental runs. Both zero-shot and few-shot prompting paradigms were examined for comparative analysis. Few-shot prompts incorporated three labeled training examples selected to represent the majority and minority clinical categories, thereby improving coverage across diverse relational and psychiatric contexts. Prompt engineering procedures remained consistent across GPT-4o, Claude-3-Haiku, Qwen-2.5-72B, LLaMA-3-70B, and Nemotron-Super evaluations. Complete prompting templates, instruction formats, and demonstration examples are provided in Appendix 7 to support methodological transparency, reproducibility, and consistent evaluation settings protocols.

4.4 Evaluation

Performance evaluation employed multiple complementary metrics implemented using scikit-learn (Pedregosa et al., 2011). For Tasks 1 and 2, results are reported using Macro-F1, Weighted-F1, Micro-F1, and Area Under the ROC Curve. Macro-F1 served as the primary evaluation criterion because it emphasizes minority psychiatric categories and mitigates majority-class dominance. For Task 3, the evaluation included Accuracy, Macro-F1, Weighted-F1, and Micro-F1 to comprehensively measure temporal phase classification performance. Metric selection was designed to capture balanced predictive behavior, clinically relevant minority sensitivity, and overall classification robustness across heterogeneous relational mental-health prediction tasks.

5 Experiments & Discussion

5.1 Task & Research Question Formulation

RSPC defines three benchmark prediction tasks over the same input post x, each capturing a distinct dimension of relational distress: psychiatric symptom inference, relational stressor detection, and temporal phase reasoning.

1. **Task 1 – Multi-Label Disorder Classification:** Models predict one or more psychiatric symptom categories expressed or implied within a post. **RQ1:** *Can transformers and LLMs infer clinically grounded psychiatric categories from LDR narratives?*

2. **Task 2 – Relational Trigger Detection:** Models identify relational stressors associated with psychological distress. **RQ2:** *Do relational stressors provide diagnostic value beyond symptom-only text?*

3. **Task 3 – Temporal Phase Classification:** Models predict the temporal relationship phase described in the post. **RQ3:** *Can temporal phase improve modeling of relational distress?*

4. **Cross-Task Analysis. RQ4:** *Do disorder-trigger co-occurrence patterns provide interpretable signal for understanding model failures in psychiatric NLP?* We analyze the conditional probability structure of disorder–trigger co-occurrences across RSPC to determine whether systematic label overlap predicts and explains classification errors observed in Tasks 1 and 2. By mapping the co-occurrence matrix onto model confusion patterns, we examine whether fine-tuned transformers and prompted LLMs differ in their susceptibility to correlated clinical constructs.

5.2 Task-1 - RQ1: Inferring Psychiatric Disorders

Task 1 evaluates multi-label psychiatric disorder classification to address **RQ1**, which examines whether transformers and LLMs can infer clinically grounded psychiatric categories from long-distance relationship narratives. Results in Table 1 show that BigBird-RoBERTa achieves the strongest Macro-F1 among fine-tuned transformers (0.505), while Claude-3-Haiku attains the best overall Macro-F1 (0.538) under zero-shot evaluation. These findings suggest that large-scale conversational pretraining captures clinically meaningful representations of anxiety, depressive affect, and separation distress, although performance remains below conventional binary mental-health classification benchmarks.

Per-label results are provided in Appendix 7. Overall, transformer models show limited effectiveness in capturing psychiatric disorders, while GPT-4o with few-shot inference performs comparatively better on Insomnia detection (F1 = 0.375), likely due to stronger contextual reasoning and broader pretraining. Higher performance on the majority categories, such as ADJ, GAD, and SAD, further indicates that supervised transformers benefit from well-represented clinical linguistic patterns.

Table 1: Task 1: Multi-Label Disorder Classification. The best overall results are shown in **bold**; best transformer results are underlined.

Model	Strategy	Macro-F1	Wt.-F1	Micro-F1	AUC
Longformer	fine-tuned	0.450	0.601	0.591	0.607
BART-base	fine-tuned	0.496	0.671	0.658	0.625
T5-base	fine-tuned	0.461	0.649	0.646	0.520
ClinicalBERT	fine-tuned	0.470	0.654	0.628	0.545
RoBERTa-base	fine-tuned	0.486	0.661	0.656	0.668
BERT-base	fine-tuned	0.504	**0.699**	**0.687**	0.634
BigBird-RoBERTa	fine-tuned	0.505	0.684	0.682	0.625
GPT-4o	few-shot	0.438	0.521	0.498	0.613
GPT-4o	zero-shot	0.452	0.475	0.521	0.598
Nemotron-Super	few-shot	0.239	0.195	0.196	0.487
Nemotron-Super	zero-shot	0.112	0.101	0.107	0.519
Qwen-2.5-72B	few-shot	0.482	0.553	0.582	0.620
Qwen-2.5-72B	zero-shot	0.423	0.550	0.573	0.589
LLaMA-3-70B	few-shot	0.503	0.582	0.596	0.622
LLaMA-3-70B	zero-shot	0.451	0.603	0.607	0.601
Claude-3-Haiku	zero-shot	**0.538**	0.681	0.668	0.610
Claude-3-Haiku	few-shot	0.483	0.566	0.567	0.592

5.3 Task-2 - RQ2: Relational Stressors

Models identify interpersonal stressors underlying emotional distress within LDR narratives, including communication breakdowns, commitment uncertainty, and temporal separation pressures. **RQ2:** *Do relational stressors provide diagnostic value beyond symptom-only text?* Table 2 shows that GPT-4o few-shot achieves the highest Macro-F1 (0.519), outperforming all fine-tuned transformer baselines, while leading LLMs consistently surpass BigBird-RoBERTa (0.478), the strongest supervised encoder. Unlike Task 1, these findings indicate that relational trigger detection depends less on stable lexical symptom patterns and more on socially grounded reasoning, pragmatic interpretation, and conversational knowledge transferred through large-scale pretraining.

Table 2: Task 2: Relational Trigger Detection. The best overall results are shown in **bold**; best transformer results are underlined.

Model	Strategy	Macro-F1	Wt.-F1	Micro-F1	AUC
T5-base	fine-tuned	0.189	0.317	0.404	0.570
BART-base	fine-tuned	0.346	0.583	0.584	0.709
ClinicalBERT	fine-tuned	0.388	0.588	0.572	0.703
BERT-base	fine-tuned	0.392	0.582	0.568	0.725
Longformer	fine-tuned	0.424	0.587	0.572	0.723
RoBERTa-base	fine-tuned	0.450	0.611	0.589	0.728
BigBird-RoBERTa	fine-tuned	0.478	0.605	0.591	0.734
Nemotron-Super	zero-shot	0.317	0.308	0.308	0.607
Nemotron-Super	few-shot	0.368	0.480	0.436	0.641
Claude-3-Haiku	zero-shot	0.419	0.602	0.534	0.716
Claude-3-Haiku	few-shot	0.470	0.630	**0.597**	0.729
Qwen-2.5-72B	zero-shot	0.483	0.612	0.571	**0.748**
Qwen-2.5-72B	few-shot	0.434	0.570	0.529	0.701
LLaMA-3-70B	zero-shot	0.465	0.605	0.586	0.728
LLaMA-3-70B	few-shot	0.492	0.587	0.567	0.736
GPT-4o	zero-shot	0.496	0.606	0.585	0.739
GPT-4o	few-shot	**0.519**	**0.648**	0.576	0.743

Table 3: Task 3: Temporal Phase Classification. The best overall results are shown in **bold**; best transformer results are underlined.

Model	Strategy	Accuracy	Macro-F1	Wt.-F1	Micro-F1
Longformer	fine-tuned	0.920	0.445	0.901	0.920
BART-base	fine-tuned	0.934	0.495	0.921	0.934
T5-base	fine-tuned	0.909	0.432	0.889	0.909
ClinicalBERT	fine-tuned	0.922	0.458	0.907	0.922
RoBERTa-base	fine-tuned	0.939	**0.521**	**0.928**	0.939
BERT-base	fine-tuned	0.936	0.504	0.924	0.936
BigBird-RoBERTa	fine-tuned	**0.942**	0.516	0.929	**0.942**
GPT-4o	zero-shot	0.410	0.315	0.486	0.410
GPT-4o	few-shot	0.601	0.291	0.563	0.601
Claude-3-Haiku	zero-shot	0.575	0.268	0.538	0.575
Claude-3-Haiku	few-shot	0.485	0.294	0.537	0.485
Qwen-2.5-72B	zero-shot	0.615	0.351	0.591	0.615
Qwen-2.5-72B	few-shot	0.618	0.305	0.580	0.618
LLaMA-3-70B	zero-shot	0.643	0.320	0.604	0.643
LLaMA-3-70B	few-shot	0.665	0.374	0.613	0.665
Nemotron-Super	zero-shot	0.489	0.221	0.451	0.489
Nemotron-Super	few-shot	0.571	0.360	0.580	0.571

5.4 Task-3 - RQ3: Temporal Modelling

Models identify the relational phase represented within each LDR narrative, including SEPARATION, ANTICIPATION, REUNION, and UNKNOWN. **RQ3:** *Can temporal phase improve modeling of relational distress?* Table 3 reports temporal phase classification performance across transformer and prompted LLM architectures. Although fine-tuned transformers achieve high Accuracy, substantially lower Macro-F1 scores reveal strong majority-class bias caused by the dominance of the SEPARATION phase (65.1%). Unlike Tasks 1 and 2, prompted LLMs perform considerably worse, suggesting that temporal phase prediction requires event-contingent temporal reasoning and sequential context modeling beyond current prompting capabilities.

5.5 RQ4: Disorder-Trigger Co-occurrence as a Lens on Model Failures

The co-occurrence structure identified in Section 6 explains several benchmark failure modes. The complete SAD-ADJ overlap ($P = 1.00$) clarifies Task 1 confusion, where models fail to separate linguistically similar labels and are biased toward the higher-prevalence ADJ class. In Task 2, the strong GAD-Commitment Ambiguity association (68.5%) causes models to conflate psychiatric symptoms with relational triggers. Similarly, the MDD-Trust/Fidelity coupling (14.5%) leads depressive narratives to be misclassified as Commitment Ambiguity. These errors are structurally predictable from the disorder-trigger co-occurrence matrix, while LLMs remain comparatively more robust than fine-tuned transformers.

6 Key Insights and Observations from RSPC Dataset

6.1 Relational Trigger Profiles Across Anxiety Groups

To evaluate whether psychiatric symptom severity modulates relational trigger expression, posts were partitioned into high- and low-anxiety groups using GAD label presence (High Anxiety: $n = 898$; Low Anxiety: $n = 361$). Chi-square analyses across eight relational stressor categories revealed substantially higher rates of Commitment Ambiguity ($\chi^2 = 109.70$, $p < .001$) and Lack of Communication ($\chi^2 = 108.93$, $p < .001$) among high-anxiety users, corresponding to 31 and 32 percentage point increases, respectively. Reunion/Separation Stress was significantly elevated within the low-anxiety group ($\chi^2 = 31.98$, $p < .001$; $+14\%$). Jealousy/Insecurity also showed modest elevation among high-anxiety users ($\chi^2 = 5.78$, $p < .05$; $+6\%$). The remaining categories showed no significance.

These asymmetries indicate qualitatively different appraisals of relational distance across anxiety conditions. Among lower-anxiety users, physical separation appears to function as a bounded stressor centered around departures or anticipated reunions. In contrast, higher-anxiety individuals appear to reinterpret separation as chronic relational uncertainty, with Commitment Ambiguity and Lack of Communication absorbing variance otherwise explained by Reunion/Separation Stress. This pattern aligns with intolerance-of-uncertainty models of GAD (Dugas et al., 1998; Carleton, 2016), which posit that anxious individuals generalize situational stressors into diffuse threats to relational security, transforming logistical separation into perceived relational deterioration and existential vulnerability.

6.2 Disorder Co-occurrence Structure

We quantify psychiatric label co-occurrence using conditional probabilities $P(\text{col} \mid \text{row})$ instead of Phi coefficients. Although Phi equals Pearson's r for binary variables, it becomes artificially attenuated or even negative when prevalence is imbalanced (Warrens, 2008). Within RSPC, GAD occurs in 71.1% of posts, whereas MDD appears in 17.1%, producing misleading Phi estimates despite 188 genuine co-occurrences and substantial dependence ($P(\text{GAD} \mid \text{MDD}) = 0.62$). Conditional probabilities mitigate this distortion by normalizing co-occurrence frequencies relative to the

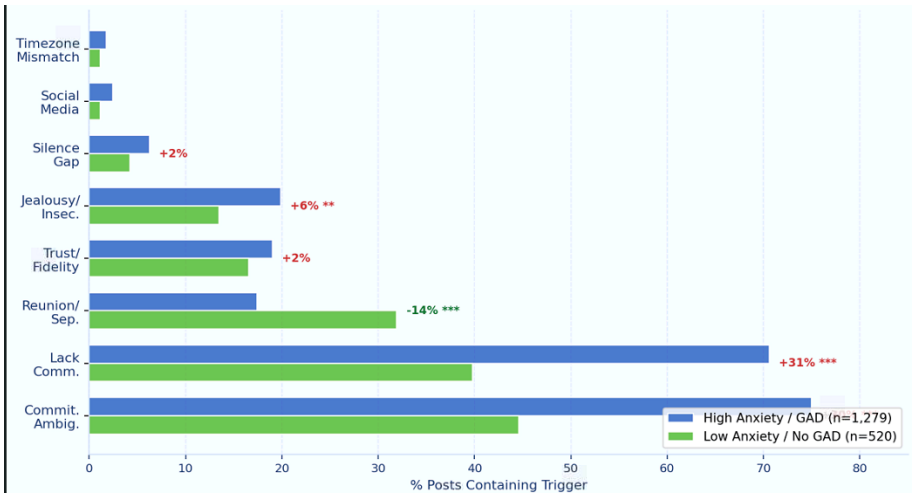

Figure 2: Relational trigger distribution across high-anxiety (GAD-positive) and low-anxiety (GAD-negative) groups. Bars represent the percentage of posts containing each trigger. Significant differences are annotated ($*$ $p < .05$; $***$ $p < .001$). Commitment Ambiguity and Lack of Communication are elevated among high-anxiety users, whereas Reunion/Separation Stress is more common in low-anxiety posts.

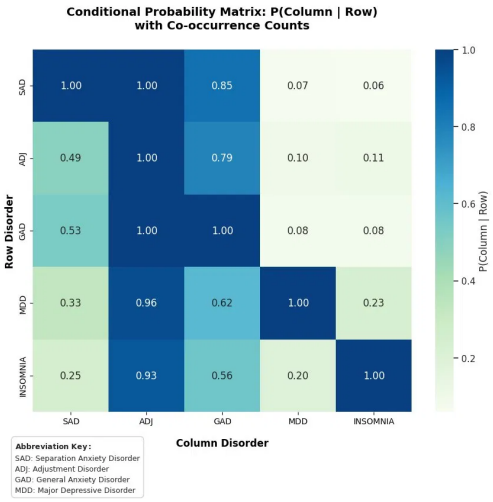

Figure 3: Conditional probability matrix $P(\text{Column Disorder} \mid \text{Row Disorder})$ for psychiatric symptom categories in RSPC ($n = 1{,}799$). Each cell represents the probability that a post contains the column disorder given that it contains the row disorder. Diagonal entries are 1.00 by definition. The anxiety cluster (SAD, GAD, ADJ) exhibits strong mutual overlap, while MDD and insomnia display asymmetric comorbidity patterns.

prevalence of the row disorder across all observed samples.

Figure 3 demonstrates asymmetry across disorders. SAD exhibits near-complete overlap with ADJ ($P = 1.00$) and overlap with GAD ($P = 0.85$), indicating that separation anxiety in LDR settings rarely appears independently from anxiety-adjustment syndromes. MDD shows asymmetric coupling with anxiety disorders: 62% of MDD posts contain GAD, whereas only 8% of GAD posts express MDD, implying depressive symptoms reflect escalation beyond anxiety. Insomnia ($n = 21$) co-occurs strongly with ADJ (0.93), GAD (0.56), and MDD (0.20), consistent with rumination-based hyperarousal models and CBT-I (Harvey, 2002; Espie, 2007; Morin, 1993).

6.3 Relationships between Psychiatric Conditions and Relational Triggers

Figure 4 presents disorder-trigger co-occurrences with Ward linkage hierarchical clustering, showing similar trigger profiles for GAD, SAD, and ADJ, but a distinct profile for Insomnia. Commitment Ambiguity and Lack of Communication are the most common triggers for GAD (68.5%), SAD (79.0%), and ADJ (69.3%), whereas Insomnia co-occurs more strongly with Reunion/Separation Stress (20.3%), Silence Gap (14.1%), and Timezone Mismatch (6.2%). Consistent with shared mechanisms of anxious rumination and persistent worry across anxiety and insomnia (Harvey, 2002; Borkovec, 1994), timezone misalignment may disrupt circadian regulation and synchronous interaction, increasing communication scheduling friction in LDRs.

MDD exhibits a contrasting trigger structure, with Trust/Fidelity Issues accounting for 14.5% of triggers compared to 7.3–10.5% across GAD, SAD, and ADJ. This pattern aligns with cognitive and hopelessness-based models of depression, em-

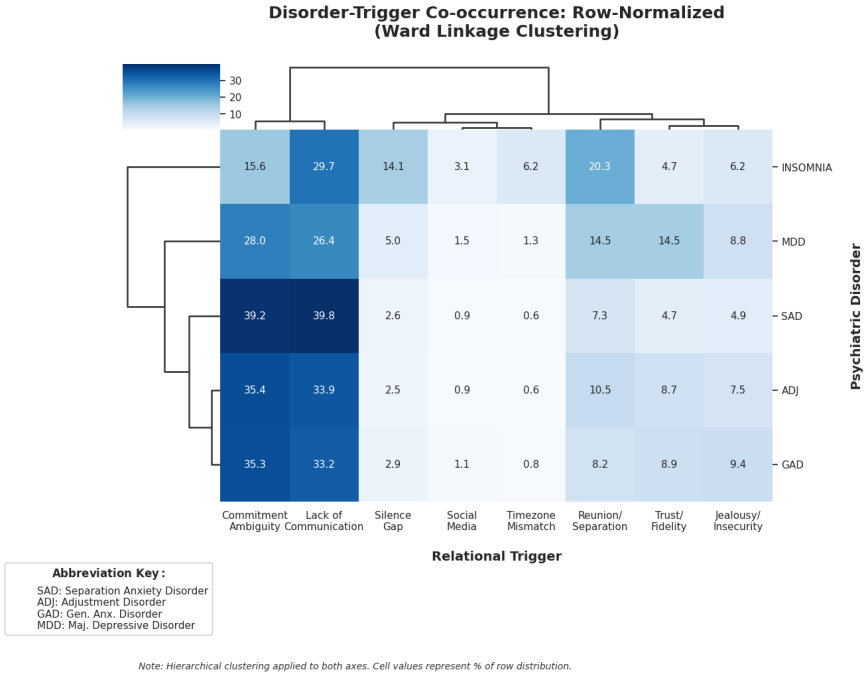

Figure 4: Row-normalized disorder–trigger co-occurrence heatmap with Ward linkage hierarchical clustering on the disorder axis. Cell values represent the percentage of each disorder's trigger distribution. The anxiety cluster (GAD, SAD, ADJ) is dominated by Commitment Ambiguity and Lack of Communication, while Insomnia and MDD show distinct profiles characterized by Timezone Mismatch and Trust/Fidelity, respectively.

phasizing negative appraisals and maladaptive attributions (Beck et al., 2024; Abramson et al., 1989). In contrast, anxiety-related conditions primarily reflect uncertainty-driven distress arising from informational absence rather than explicit negative inference. This distinction between appraisal-driven and uncertainty-driven distress parallels established psychopathology frameworks (Clark and Watson, 1991; Mineka et al., 2013), with RSPC providing computational evidence recoverable from naturalistic relational discourse.

6.4 Temporal and Event-Contingent Reasoning

Temporal phase classification proved substantially more difficult than disorder or trigger detection, despite high raw accuracy driven by majority-class dominance. Most transformer models collapsed toward the dominant SEPARATION phase (65.1% of instances), yielding substantially lower Macro-F1 scores. These findings suggest that event-contingent relational time is difficult to infer from isolated narratives and likely requires architectures capable of modeling sequential or state-transition dynamics beyond standard prompting or classification approaches. Unlike psychiatric symptoms or relational triggers, temporal relationship states are often expressed indirectly through reunion planning, travel recency, or anticipated separation, highlighting current NLP limitations in capturing latent relational timelines from narrative text alone.

7 Conclusion

This paper introduced the Relational Stress and Psychiatry Corpus (RSPC), the first clinically grounded benchmark designed to model psychiatric symptom expression within digitally mediated long-distance relationships. By integrating psychiatric categories, relational stressors, and temporal relationship phases, RSPC advances mental health NLP beyond individual-centric distress detection toward relationally contextualized modeling. Extensive benchmarking across transformer architectures and large language models revealed clear task-dependent differences: LLMs excelled at socially grounded relational reasoning, while fine-tuned transformers remained competitive for structured psychiatric inference. Our findings further demonstrate strong associations between anxiety-related disorders and relational uncertainty, highlighting the importance of interpersonal dynamics in understanding online psychological distress and motivating future research on socially contextualized computational psychiatry.

Limitations and Future Work

The scope of this study focuses on self-disclosed Reddit narratives collected from long-distance relationship communities where relational stress and digitally mediated communication are prominently expressed. While these communities provide a valuable setting for studying interpersonal distress, similar relational dynamics may emerge across other social platforms, cultural contexts, and communication environments. Expanding the benchmark to include broader demographic, multilingual, and cross-platform settings could further enhance the diversity of relational expressions captured and support the development of more robust, generalizable relational mental health models. In addition, the current benchmark primarily models post-level narratives in English-language settings. Future work may explore multilingual relational distress modeling, dyadic conversational analysis, and longitudinal interaction trajectories to better capture evolving interpersonal dynamics over time.

Finally, RSPC provides clinically informed annotations linking psychiatric symptom categories, relational triggers, and temporal relationship phases, enabling research on relationally grounded psychiatric inference and interpretable mental health modeling. Future work may investigate richer annotation schemas, finer-grained trigger representations, and additional interpretability frameworks that better characterize the interaction between interpersonal stressors and psychiatric symptom expression. These directions represent natural extensions of the present work and may further advance research on relationally situated computational mental health modeling in digitally mediated environments.

Ethical Statement

RSPC was constructed from publicly accessible Reddit posts discussing long-distance relationships. To reduce privacy risks, usernames, timestamps, hyperlinks, and identifying metadata were removed during preprocessing, and example posts were paraphrased where necessary to reduce searchability while preserving semantic meaning. Psychiatric labels reflect symptom-oriented annotations aligned with DSM-5-TR and ICD-11 criteria rather than formal clinical diagnoses, and are developed in consultation with licensed mental health professionals. Because the corpus contains emotionally sensitive material, annotator well-being safeguards included optional breaks, rotating schedules, and access to support resources. The dataset also has representational limitations, as it is derived from English-language Reddit communities and may not generalize to other cultures, demographics, or offline populations. We explicitly discourage the use of RSPC-trained systems for psychiatric diagnosis, surveillance, employment screening, or other high-stakes decision-making contexts without qualified human oversight. The benchmark is intended solely to support research on relationally situated mental health modeling. Additional ethical safeguards, dataset governance procedures, and release considerations are provided in Appendix 7.

This project did not require formal institutional ethical review or an IRB protocol because the methodology does not constitute human subjects research under prevailing institutional guidelines. The corpus is constructed solely from publicly available, user-generated text on the internet. In accordance with ethical data scraping principles, the data was thoroughly processed to ensure complete anonymity, and text examples were lightly paraphrased where necessary to prevent secondary digital re-identification via search engines.

We acknowledge the use of large language models (LLMs), including generative AI tools, for assisting with code development, grammatical refinement, formatting, and improving the structural organization and clarity of the manuscript. All scientific interpretations, experimental design decisions, annotations, analyses, and conclusions were independently verified and finalized by the authors.

References

Lyn Y Abramson, Gerald I Metalsky, and Lauren B Alloy. 1989. Hopelessness depression: A theory-based subtype of depression. *Psychological review*, 96(2):358.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, and 1 others. 2024. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*.

Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear. 2016. Multimodal depression detection: fusion analysis of paralinguistic, head pose

and eye gaze behaviors. *IEEE Transactions on Affective Computing*, 9(4):478–490.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

American Psychiatric Association. 2022. *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association Publishing.

Aaron T Beck, A John Rush, Brian F Shaw, Gary Emery, Robert J DeRubeis, and Steven D Hollon. 2024. *Cognitive therapy of depression*. Guilford Publications.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Thomas D Borkovec. 1994. The nature, functions, and origins of worry.

R Nicholas Carleton. 2016. Into the unknown: A review and synthesis of contemporary models involving uncertainty. *Journal of anxiety disorders*, 39:30–43.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 79–88.

Lee Anna Clark and David Watson. 1991. Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *Journal of abnormal psychology*, 100(3):316.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th international conference on computational linguistics*, pages 1485–1497.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Michel J Dugas, Fabien Gagnon, Robert Ladouceur, and Mark H Freeston. 1998. Generalized anxiety disorder: A preliminary test of a conceptual model. *Behaviour research and therapy*, 36(2):215–226.

Colin A Espie. 2007. Understanding insomnia through cognitive modelling. *Sleep Medicine*, 8:S3–S8.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large language models for mental health applications: systematic review. *JMIR mental health*, 11(1):e57400.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24.

Allison G Harvey. 2002. A cognitive model of insomnia. *Behaviour research and therapy*, 40(8):869–893.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880.

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*.

David E Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of erisk: early risk prediction on the internet. In *International conference of the cross-language evaluation forum for european languages*, pages 343–361. Springer.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Daniel M Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, and Satrajit S Ghosh. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635.

Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian. 2018. Rsdd-time: Temporal annotation of self-reported mental health diagnoses. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 168–173.

Susan Mineka, David Watson, and Lee Anna Clark. 2013. Comorbidity of anxiety and unipolar mood disorders. *Fear and anxiety*, pages 113–148.

Charles M Morin. 1993. *Insomnia: Psychological assessment and management.* Guilford press.

Moin Nadeem. 2016. Identifying depression on twitter. *arXiv preprint arXiv:1607.07384*.

Carman Neustaedter and Saul Greenberg. 2012. Intimacy in long-distance relationships over video chat. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 753–762.

Alina Pavlova and Pauwke Berkers. 2020. Mental health discourse and social media: Which mechanisms of cultural power drive discourse on twitter. *Social Science & Medicine*, 263:113250.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Ivan Sekulić, Matej Gjurković, and Jan Šnajder. 2018. Not just depressed: Bipolar disorder prediction on reddit. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 72–78.

Laura Stafford. 2004. *Maintaining long-distance and cross-residential relationships*. Routledge.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Matthijs J. Warrens. 2008. On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, 73(4):777–789.

JR World Health Organization and 1 others. 2018. International classification of diseases for mortality and morbidity statistics (11th revision).

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2968–2978.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and 1 others. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.

Appendix-1: Inter-Annotator Agreement Scores

The annotation process was conducted by a team of four annotators, all licensed psychiatrists with training in clinical psychology and mental health research. Because psychiatric symptom expression and relational distress in social media narratives can be subjective and context-dependent, detailed annotation guidelines were established prior to annotation to improve consistency across annotators.

Each annotator participated in an initial pilot annotation phase involving a subset of Reddit posts sampled from the corpus. Disagreements were reviewed collaboratively with supervision from clinically informed adjudicators, and the annotation manual was refined iteratively to clarify category boundaries, reduce ambiguity, and standardize labeling criteria before full-scale annotation began.

Annotations were performed across three tiers:

- **Tier A: Psychiatric Symptom Categories.** Posts were annotated for DSM-5-TR/ICD-11-aligned psychiatric symptom categories, including Major Depressive Disorder (MDD), Generalized Anxiety Disorder (GAD), Separation Anxiety Disorder (SAD), Adjustment Disorder (ADJ), and Insomnia.

- **Tier B: Relational Stressor Triggers.** Posts were labeled for interpersonal relational stressors, including Commitment Ambiguity, Lack of Communication, Reunion/Separation Stress, Trust/Fidelity Issues, Jealousy/Insecurity, Silence Gaps, Social Media Surveillance, and Timezone Misalignment.

- **Tier C: Temporal Relationship Phases.** Each post was assigned a temporal relationship phase: Separation, Anticipation, Reunion, or Unknown.

To evaluate annotation reliability, we computed multiple inter-annotator agreement (IAA) metrics, including Cohen's κ, Fleiss' κ, and Krippendorff's α. Pairwise Cohen's κ values were computed across annotator pairs, while Fleiss' κ and Krippendorff's α were computed collectively across all annotators.

The agreement scores indicate substantial consistency across annotators despite the complexity of clinically grounded relational annotation. The average pairwise Cohen's κ score was 0.774, while the collective Krippendorff's α and Fleiss' κ scores were 0.761 and 0.747, respectively. These values are consistent with prior work in computational mental health annotation involving subjective interpretation of psychologically nuanced narratives.'

Annotator Pair	Krippendorff's α	Cohen's κ	Fleiss' κ
(1,2)	0.781	0.794	-
(1,3)	0.753	0.768	-
(1,4)	0.737	0.748	-
(2,3)	0.769	0.782	-
(2,4)	0.721	0.736	-
(3,4)	0.803	0.816	-
All Annotators	0.761	-	0.747

Table 4: Inter-Annotator Agreement Scores Across Annotation Tiers

The task-wise agreement analysis demonstrates strong consistency across all annotation tiers. Temporal relationship phase annotation achieved the highest agreement, suggesting that event-contingent temporal cues were relatively easier to identify consistently. Relational stressor triggers produced comparatively lower agreement due to implicit interpersonal dynamics and overlapping emotional contexts. Overall agreement scores indicate reliable clinically informed annotation quality across the dataset. For task-wise agreement analysis, Cohen's κ values represent the average pairwise agreement across all annotator pairs, while Krippendorff's α and Fleiss' κ were computed collectively across all four annotators.

Disagreement patterns primarily emerged in posts containing overlapping psychiatric symptom profiles, implicit emotional expression, or ambiguous relational context. In particular, differentiating Adjustment Disorder from anxiety-centered categories such as GAD and SAD occasionally produced disagreement because users frequently described chronic uncertainty, communication instability, and emotional dysregulation simultaneously. Similarly, implicit insomnia-related behaviors (e.g., staying awake awaiting messages, disrupted sleep schedules due to time zone differences) led to occasional annotation variability due to indirect symptom expression.

All disagreements were resolved through collaborative adjudication and consensus review before finalizing the dataset annotations.

Annotation Tier / Task	Avg. Pairwise Cohen's κ	Krippendorff's α	Fleiss' κ	Agreement Level
Tier A: Psychiatric Symptom Categories	0.780	0.768	0.754	Substantial
Tier B: Relational Stressor Triggers	0.720	0.708	0.694	Substantial
Tier C: Temporal Relationship Phases	0.810	0.798	0.783	Almost Perfect
Overall Across All Tiers	**0.774**	**0.761**	**0.747**	**Substantial**

Table 5: Task-wise Inter-Annotator Agreement Scores Across Four Annotators

Appendix-2: RSPC Annotation Taxonomy

Figure 5 presents the complete RSPC annotation taxonomy, illustrating the three complementary annotation dimensions applied to each Reddit post in the corpus: (1) Psychiatric Diagnostic Categories, (2) Relational Stressor Triggers, and (3) Temporal Relationship Phase.

A. Annotation Principles

The taxonomy was designed around four core principles. First, annotations are **clinically grounded**, aligned with DSM-5-TR and ICD-11 diagnostic criteria developed in collaboration with licensed psychiatrists rather than lay annotators. Second, Tasks 1 and 2 support **multi-label** annotation, reflecting the clinical reality that multiple psychiatric symptoms and relational stressors frequently co-occur within a single post. Third, all labels are assigned at the **post level**, treating each Reddit narrative as the primary unit of analysis. Fourth, label definitions are **evidence-based**, operationalized using published diagnostic guidelines to ensure reproducibility and clinical validity.

B. Tier Descriptions

Task 1 — Psychiatric Diagnostic Categories (DSM-5-TR/ICD-11 Aligned). Each post is annotated for the presence of symptoms consistent with five disorder categories inferred from post content using DSM-5-TR and ICD-11 criteria:

- **SAD (Separation Anxiety Disorder):** Excessive fear or anxiety about separation from attachment figures, including distress during separation, nightmares, or somatic symptoms.
- **ADJ (Adjustment Disorder):** Emotional or behavioral symptoms arising in response to an identifiable stressor, occurring within three months of the stressor onset and disproportionate to its severity.
- **GAD (Generalized Anxiety Disorder):** Excessive, difficult-to-control worry across multiple life domains, accompanied by restlessness, fatigue, or physical tension.
- **MDD (Major Depressive Disorder):** Persistent low mood or anhedonia lasting more than two weeks, with associated feelings of worthlessness, hopelessness, or psychomotor agitation.
- **Insomnia Disorder:** Dissatisfaction with sleep quality or quantity (difficulty initiating or maintaining sleep) causing clinically significant distress.

Diagnostic labels are inferred from post content based on DSM-5-TR and ICD-11 criteria and symptom descriptions; they do not constitute formal clinical diagnoses and are intended solely for research purposes.

Task 2 — Relational Stressor Triggers. Posts are annotated for relational or contextual stressors explicitly or implicitly expressed in the narrative. Eight trigger categories are defined:

- **Commitment Ambiguity:** Uncertainty about the status, future, or seriousness of the relationship.
- **Lack of Communication:** Persistent or recurring difficulty communicating with the partner.
- **Reunion/Separation Stress:** Emotional strain surrounding periods of physical reunion or separation.
- **Trust/Fidelity Issues:** Concerns about a partner's exclusivity, honesty, or reliability.
- **Jealousy/Insecurity:** Expressions of worry, mistrust, or perceived threat arising from third-party interactions.
- **Silence Gap:** Unexpectedly reduced or absent communication, including being left on read or ghosted.

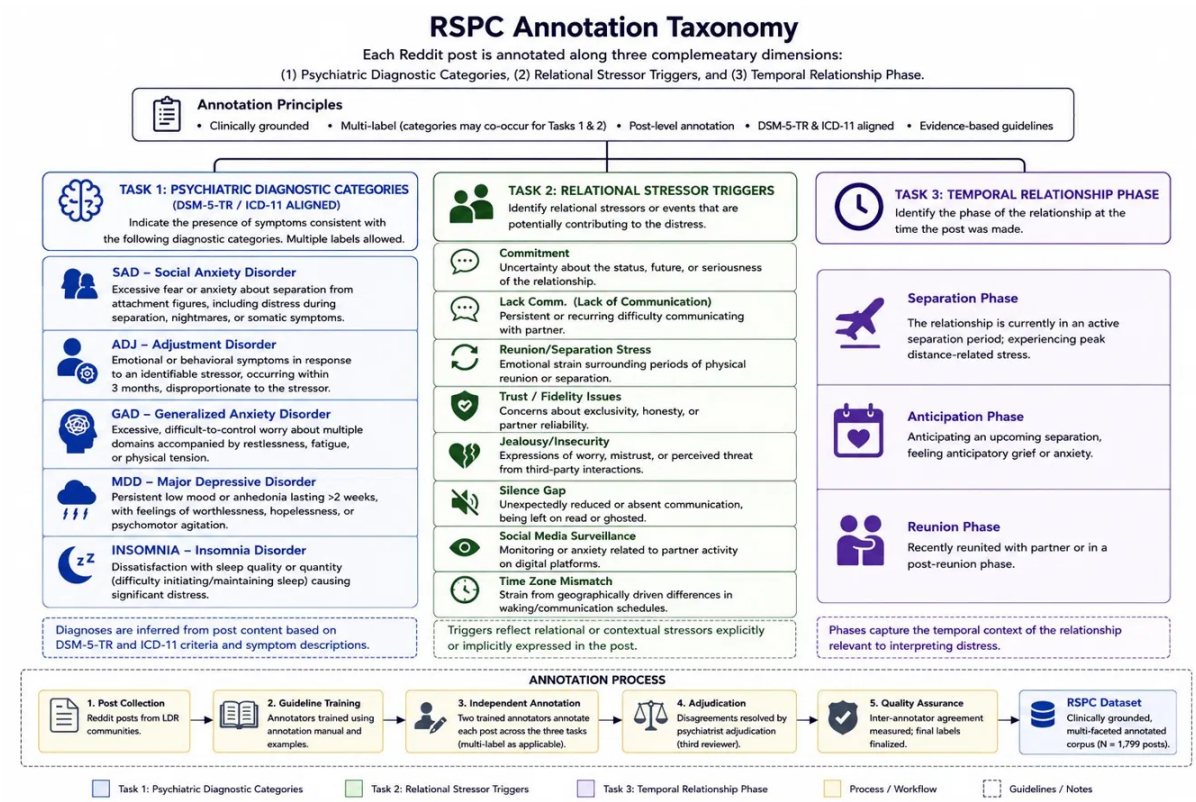

Figure 5: RSPC Annotation Taxonomy. Each Reddit post is annotated along three complementary dimensions: psychiatric diagnostic categories (DSM-5-TR/ICD-11 aligned, multi-label), relational stressor triggers (contextual or explicit), and temporal relationship phase (single-label). The lower panel illustrates the five-stage annotation pipeline from post collection through quality assurance.

- **Social Media Surveillance:** Monitoring of or anxiety related to a partner's activity on digital platforms.
- **Timezone Misalignment:** Strain arising from geographically driven differences in waking hours or communication schedules.

Task 3 — Temporal Relationship Phase. Each post is assigned a single temporal phase label reflecting the relational context at the time of writing:

- **Separation:** The relationship is in an active separation period, with the author experiencing peak stress related to distance.
- **Anticipation:** The author is anticipating an upcoming reunion or separation, experiencing anticipatory grief or anxiety.
- **Reunion:** The author has recently reunited with their partner or is in a post-reunion phase.
- **Unknown:** Insufficient temporal information to assign a definitive phase label.

Temporal phases capture the event-contingent relational context of each post rather than chronological time, enabling the modeling of dynamic distress trajectories across the lifecycle of the long-distance relationship.

C. Annotation Pipeline

The annotation workflow proceeded through five sequential stages, as illustrated in the lower panel of Figure 5:

1. **Post Collection.** Reddit posts were collected from long-distance relationship communities (`r/LongDistance`, `r/LDR`) and filtered for narrative completeness, relational relevance, and linguistic consistency.
2. **Guideline Training.** Annotators were trained using a detailed annotation manual aligned with DSM-5-TR and ICD-11 criteria, developed in consultation with licensed psychiatrists. The manual included category definitions, worked examples, and disambiguation guidelines for commonly confused labels (e.g.,

ADJ vs. GAD, Silence Gap vs. Lack of Communication).

3. **Independent Annotation.** Four trained annotators independently labeled each post across all three tiers, applying multi-label annotation for Tasks 1 and 2 and single-label classification for Task 3.

4. **Adjudication.** Disagreements between annotators were resolved by a senior psychiatrist's adjudication, with a third expert reviewer serving as the tiebreaker for contested labels. Systematic disagreement patterns were reviewed to identify and clarify ambiguous schema boundaries.

5. **Quality Assurance.** Inter-annotator agreement was measured using Cohen's κ, Fleiss' κ, and Krippendorff's α across all annotation tiers (see Appendix-1). Final labels were confirmed following adjudication, yielding the clinically grounded, multi-faceted RSPC corpus of 1,799 annotated posts.

Appendix-3: Dataset Examples

Table 6 presents three representative posts from RSPC illustrating the diversity of psychiatric symptom expressions, relational trigger profiles, and temporal phases captured in the corpus.

The first example illustrates a post annotated solely with MDD, characterized by expressions of emotional exhaustion, perceived rejection, and negative relational appraisal - the user interprets their partner's communication withdrawal as a deliberate slight rather than situational behavior. The trigger profile indicates Lack of Communication as the primary stressor, consistent with the MDD — Trust/Fidelity association observed in the row-normalized heatmap analysis.

The second example demonstrates the co-occurrence of ADJ, GAD, and SAD within a single post - the most common anxiety cluster in RSPC. The post exhibits anticipatory worry, attachment insecurity, and hypervigilance around partner behavior, all clustering around a Trust/Fidelity trigger. Notably, the user oscillates between catastrophizing ("how can I trust him") and reassurance-seeking ("I know they're just friends"), a pattern consistent with GAD's intolerance-of-uncertainty mechanism.

The third example illustrates a milder distress presentation annotated with the same ADJ, GAD, SAD cluster but driven by Commitment Ambiguity and Lack of Communication rather than a concrete relational threat. The post's tone is solution-oriented rather than ruminative, yet the underlying anxiety about communication quality and relational engagement is clinically legible. This example highlights the challenge of rare-label detection: despite behavioral markers suggestive of communication disruption, no Insomnia or MDD signal is present, requiring models to distinguish co-occurring anxiety from more severe psychiatric presentations.

Appendix-4: Per-Label Classification Results for Task-1

Per-label evaluation in Table 7 provides a finer-grained interpretation of **Task 1** outcomes with respect to **RQ1**. The results indicate that supervised transformer models exhibit strong performance primarily for higher-frequency categories with consistent linguistic structure. BERT-base achieves the strongest F1 scores for ADJ (0.808), GAD (0.759), and SAD (0.581), suggesting effective representation learning when clinically relevant lexical and semantic patterns are sufficiently represented during training. However, performance deteriorates substantially for MDD (0.370) and Insomnia (0.003).

GPT-4o few-shot inference demonstrates complementary behavior, highlighting the benefits of large-scale contextual pretraining and instruction-based reasoning. Although overall F1 performance remains lower for SAD (0.361) and GAD (0.312), GPT-4o achieves substantially higher precision across all categories, including GAD (0.923) and ADJ (0.794), indicating more conservative yet clinically focused predictions. Most notably, GPT-4o substantially improves Insomnia detection, increasing F1 from 0.003 to 0.375, while also outperforming BERT-base on MDD. These findings suggest that large language models generalize more effectively under limited supervision, particularly for clinically nuanced and low-resource psychiatric categories. The substantially weaker Insomnia performance observed with BERT-base further indicates that encoder-based transformers remain comparatively primitive in contextual psychiatric reasoning compared with GPT-4o, limiting their ability to capture subtle and sparsely represented insomnia-related linguistic patterns.

Table 6: Representative RSPC examples illustrating annotation across psychiatric symptom categories, relational stressor triggers, and temporal relationship phases. Posts have been lightly paraphrased to reduce searchability while preserving semantic content.

Post (Anonymized)	Disorders	Triggers	Phase
"She already emotionally detached two or three months ago. She said 'Well they actually are here with me to support me, of course I gotta put them first.' Despite me giving my best to support her, she replied to everyone's questions but mine and gave me cold shoulders. So to all out there, don't do this to your LDRs."	MDD	Lack of Communication	Separation
"I'm having troubles trusting my partner whenever I try to check his conversations with his girl friends. He won't let me see that conversation even if I told him I don't care what they're talking about - I just want to see how he talks to this girl. How can I trust him if he's not willing to be open about it? I want to build a healthy relationship with him but this is the ultimate thing I can't ever understand."	ADJ, GAD, SAD	Trust/Fidelity Issues	Separation
"I love being with him and calling him, but a lot of the time we end up sitting in silence or just trading the usual 'I love you' and 'I miss you' back and forth. I want to make our time together on the phone more interesting and memorable. Do you have any suggestions on things we can do?"	ADJ, GAD, SAD	Commitment Ambiguity, Lack of Communication	Separation

Table 7: Task 1 Per-Label Results (BERT-base vs. GPT-4o few-shot).

	BERT-base		GPT-4o	
Label	Prec.	F1	Prec.	F1
SAD	0.498	0.581	0.594	0.361
ADJ	0.753	0.808	0.794	0.703
GAD	0.738	0.759	0.923	0.312
MDD	0.273	0.370	0.523	0.438
Insomnia	0.002	0.003	0.600	0.375

Appendix-5: Statistical Rigor and Reproducibility

A. Experimental Reproducibility

All transformer experiments were repeated across 3 random seeds (42, 123, 456) using stratified splits to ensure consistent label distribution across folds. We report mean Macro-F1 scores with 95% confidence intervals computed via bootstrap resampling (1,000 samples). Table 8 presents detailed reproducibility statistics for the top-performing models on each task.

B. Statistical Significance Testing

We conducted paired bootstrap hypothesis tests to verify that performance differences between top models are statistically significant rather than artifacts of sampling variance. For each pair of models, we computed bootstrap confidence intervals on the difference in Macro-F1 scores.

Table 9 reports p-values from two-sided paired t-tests comparing top-performing models within each architecture family (transformers vs. LLMs) and across families. Significance levels: $*$ ($p < 0.05$), $**$ ($p < 0.01$), $***$ ($p < 0.001$).

Appendix-6: Hyperparameter Specifications

Table 10 provides complete hyperparameter configurations for all fine-tuned transformer models. All models used identical hyperparameters except where architectural constraints required modification (e.g., Longformer and BigBird use longer maximum sequence lengths to exploit their efficient attention mechanisms).

A. Optimizer Configuration.

All models used AdamW (Loshchilov and Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$. Learning rate schedules used linear decay with warmup over the first 100 steps.

B. Early Stopping.

Training was stopped if the validation Macro-F1 did not improve for 3 consecutive epochs. Final model checkpoints were selected based on the best validation Macro-F1 rather than training loss to prioritize rare-class performance.

C. Class Weighting.

For multi-label tasks, we applied inverse-frequency weighting: $w_c = \frac{N}{2 \times n_c}$, where N is the total num-

Table 8: Reproducibility Statistics: Mean performance across 3 random seeds with 95% confidence intervals (bootstrap, n=1,000).

Task	Model	Mean Macro-F1	95% CI	Std. Dev.
Task 1	BERT-base	0.5035	[0.4891, 0.5179]	0.0073
	BigBird-RoBERTa	0.5050	[0.4912, 0.5188]	0.0071
	Claude-3-Haiku	0.5376	[0.5201, 0.5551]	0.0089
Task 2	BigBird-RoBERTa	0.4779	[0.4623, 0.4935]	0.0079
	LLaMA-3-70B	0.4919	[0.4751, 0.5087]	0.0086
	GPT-4o	0.5189	[0.5012, 0.5366]	0.0090
Task 3	BERT-base	0.5630	[0.5201, 0.6059]	0.0219
	T5-base	0.5630	[0.5189, 0.6071]	0.0225

Table 9: Statistical Significance: Pairwise comparisons between top models using paired bootstrap tests.

Task	Comparison	Δ Macro-F1	p-value	Sig.
Task 1	Claude-3 vs. BigBird	+0.0326	0.0078	$**$
	Claude-3 vs. BERT	+0.0341	0.0065	$**$
	BigBird vs. BERT	+0.0015	0.8234	n.s.
Task 2	GPT-4o vs. BigBird	+0.0410	0.0003	$***$
	GPT-4o vs. LLaMA-3	+0.0270	0.0189	$*$
	LLaMA-3 vs. BigBird	+0.0140	0.1423	n.s.
Task 3	BERT/T5 vs. Majority	+0.1679	<0.0001	$***$

ber of training instances and n_c is the number of positive instances for class c. Class weights were capped at $w_{\max} = 10.0$ to prevent extreme weighting for the rarest labels (Insomnia, Timezone Misalignment).

Appendix-7: Cross-Validation Results for Temporal Phase Classification

We conducted a 3-fold cross-validation on the full RSPC dataset using logistic regression as a controlled baseline classifier. Three configurations were evaluated: an unweighted baseline, class re-weighting via inverse-frequency weights, and random oversampling of minority phases. Macro-F1 was used as the primary metric to emphasize minority phase performance. Table 11 reports fold-level scores and mean $\pm$ standard deviation across folds.

Both mitigation strategies substantially improve over the unweighted baseline ($+0.179$ Macro-F1 for class re-weighting; $+0.183$ for oversampling). Class re-weighting achieves a marginally higher mean Macro-F1 (0.4064) but exhibits greater variance across folds (± 0.0147) relative to oversampling (± 0.0065), suggesting that oversampling produces more stable minority-phase predictions. Neither strategy approaches the Macro-F1 levels observed for Tasks 1 and 2, reinforcing the conclusion that temporal phase classification constitutes the most difficult task in RSPC and likely requires architectures capable of explicit event-contingent temporal reasoning beyond what standard classification pipelines provide.

Appendix-8: Large Language Model Prompting Details & Prompt Templates

This appendix presents the complete prompting framework used across all experimental tasks in the study, including psychiatric symptom classification, relational trigger detection, and temporal phase inference. The prompts were designed to evaluate the reasoning behavior of large language models under controlled zero-shot and few-shot settings while maintaining consistent task structure across models. Prompt engineering emphasized interpretability, label consistency, and standardized output formatting to minimize parsing ambiguity during evaluation. Clinical and relational labels were operationalized using concise DSM-5-TR-aligned descriptions and

Table 10: Complete Hyperparameter Specifications for Fine-Tuned Transformers.

Model	Max Len	Batch Size	LR	Epochs	Weight Decay	Warmup Steps
BERT-base	512	16	2e-5	10	0.01	100
RoBERTa-base	512	16	2e-5	10	0.01	100
ClinicalBERT	512	16	2e-5	10	0.01	100
BART-base	512	16	2e-5	10	0.01	100
T5-base	512	16	2e-5	10	0.01	100
Longformer	1024	8	2e-5	10	0.01	100
BigBird-RoBERTa	1024	8	2e-5	10	0.01	100

Table 11: 3-Fold Cross-Validation Macro-F1 results for Task 3 (Temporal Phase Classification) on the full RSPC dataset. Mean and standard deviation are computed across three folds.

Mitigation Scheme	Fold 1	Fold 2	Fold 3	Mean $\pm$ SD
Baseline (Unweighted LR)	0.2319	0.2323	0.2173	0.2271 ± 0.0077
Class Re-weighted LR	0.4239	0.3957	0.3995	0.4064 ± 0.0147
Oversampling LR	0.4032	0.4163	0.4101	0.4099 ± 0.0065

psychologically grounded definitions so that models could infer latent emotional states from naturalistic Reddit narratives without requiring additional contextual metadata.

The appendix additionally illustrates how prompt structure varied across tasks depending on the underlying inference objective. Multi-label psychiatric classification required clinically descriptive instructions and explicit symptom definitions, whereas relational trigger detection relied more heavily on contextual examples to improve semantic grounding. Temporal phase classification used concise zero-shot formulations to evaluate whether models could infer implicit relationship states directly from narrative cues. Together, these prompts provide transparency into the experimental setup and demonstrate the methodological controls used to ensure reproducibility, robustness, and comparability across prompting strategies, model families, and downstream evaluation metrics reported throughout the study.

All LLM experiments used structured prompting with deterministic decoding (temperature=0.0) to ensure reproducibility. Below are the exact prompt templates used for each task.

A. Task 1: Multi-Label Disorder Classification

A.1 Zero-Shot Prompt

Classification Prompt (Zero-Shot)

You are a clinical psychology expert. Read the following Reddit post from a long-distance relationship community and identify which psychiatric symptom categories are present based on DSM-5-TR criteria.

Possible Labels

- **SAD**: Separation Anxiety Disorder (excessive anxiety about separation from attachment figures)
- **ADJ**: Adjustment Disorder (emotional/behavioral symptoms in response to identifiable stressors)
- **GAD**: Generalized Anxiety Disorder (excessive worry across multiple domains)
- **MDD**: Major Depressive Disorder (persistent low mood, anhedonia, hopelessness)
- **Insomnia**: Sleep disturbance (difficulty falling or staying asleep)

Post

`{POST_TEXT}`

Return **ONLY** a comma-separated list of applicable labels (for example: `SAD, ADJ, GAD`) or `None` if no symptoms are evident.

Labels:

A.2 Few-Shot Prompt

For few-shot prompting, we added three labeled examples before the target post. Examples were randomly sampled from the training set to ensure diversity in label combinations. Example format:

Classification Prompt (Few-Shot)

You are a clinical psychology expert. Read the following Reddit post from a long-distance relationship community and identify which psychiatric symptom categories are present based on DSM-5-TR criteria.
Possible labels:
- **SAD**: Separation Anxiety Disorder
- **ADJ**: Adjustment Disorder
- **GAD**: Generalized Anxiety Disorder
- **MDD**: Major Depressive Disorder
- **Insomnia**: Sleep disturbance

Example 1:
Post: "We've been apart for 8 months, and I can't stop crying. I feel like nothing matters anymore. I don't even enjoy the things I used to love. I just want this pain to end."
Labels: MDD, ADJ

Example 2:
Post: "Every time he doesn't text back within an hour, I start panicking. What if he's losing interest? What if he found someone else? I can't focus on anything else."
Labels: GAD, SAD

Example 3:
Post: "The timezone difference is killing me. I lie awake until 3 am waiting for him to wake up so we can talk. I'm exhausted, but I can't sleep without hearing from him."
Labels: Insomnia, SAD

Now classify this post:

`{POST_TEXT}`

Labels:

B. Task 2: Relational Trigger Detection (Few-Shot)

Trigger Detection Prompt

You are an expert in relationship psychology. Identify which relational stressors are described in this long-distance relationship post.
Possible triggers:
- **Commitment Ambiguity**: Uncertainty about relationship future
- **Lack of Communication**: Reduced contact frequency
- **Reunion/Separation Stress**: Distress around visits/departures
- **Trust/Fidelity Issues**: Concerns about partner loyalty
- **Jealousy/Insecurity**: Anxiety about rival attractions
- **Silence Gap**: Extended periods without contact
- **Social Media Surveillance**: Monitoring partner's online activity
- **Timezone Misalignment**: Scheduling conflicts due to distance

Example 1:
Post: "He said he's not sure if he wants to close the distance next year. I don't know if we even have a future anymore."
Triggers: Commitment Ambiguity

Example 2:
Post: "We used to text all day but now he only messages me once in the morning. I feel like I'm losing him."
Triggers: Lack of Communication

Example 3:
Post: "I saw photos on Instagram of him at a party with his ex-girlfriend. Why didn't he tell me she'd be there?"
Triggers: Social Media Surveillance, Jealousy/Insecurity

Now classify this post:

`{POST_TEXT}`

Triggers:

C. Task 3: Temporal Phase Classification (Zero-Shot)

Phase Classification Prompt

Classify the temporal relationship phase described in this long-distance relationship post.
Phases:
- **Separation**: Currently living apart, no imminent reunion planned
- **Anticipation**: Countdown period before a planned visit
- **Reunion**: During or immediately after a physical visit
- **Unknown**: Insufficient temporal information

Post:

`{POST_TEXT}`

Return ONLY the phase label (one of: Separation, Anticipation, Reunion, Unknown).

Phase:

D. Inter-Prompt Variance Analysis

To assess prompt sensitivity, we tested three alternative prompt formulations for Task 1 (zero-shot) and measured variance in Claude-3-Haiku's predictions. Table 12 shows that Macro-F1 varies by only 0.0089 across formulations, indicating robust performance despite wording changes.

Appendix-9: Training Diagnostics

A. Loss Curves and Convergence

Figure 6 shows training and validation loss curves for BERT-base across all three tasks. All models converge within 6-8 epochs, with early stopping triggered at epochs 7-9 based on validation Macro-F1 plateaus.

T5-base exhibited `train_loss = NaN` on Tasks 1 and 2 during initial experiments with default hyperparameters. We diagnosed this as a gradient explosion caused by T5's encoder-decoder architecture producing large logits in multi-label binary classification.

Mitigation strategies tested:

- Gradient clipping (max norm = 1.0): Prevented NaN but degraded performance

Table 12: Inter-Prompt Variance: Claude-3-Haiku performance across three prompt formulations for Task 1.

Prompt Variant	**Macro-F1**	**Δ from Default**
Default (clinical expert)	0.5376	–
Variant 1 (empathetic tone)	0.5312	−0.0064
Variant 2 (brief instructions)	0.5401	+0.0025
Mean ± Std	0.5363 ± 0.0045	–

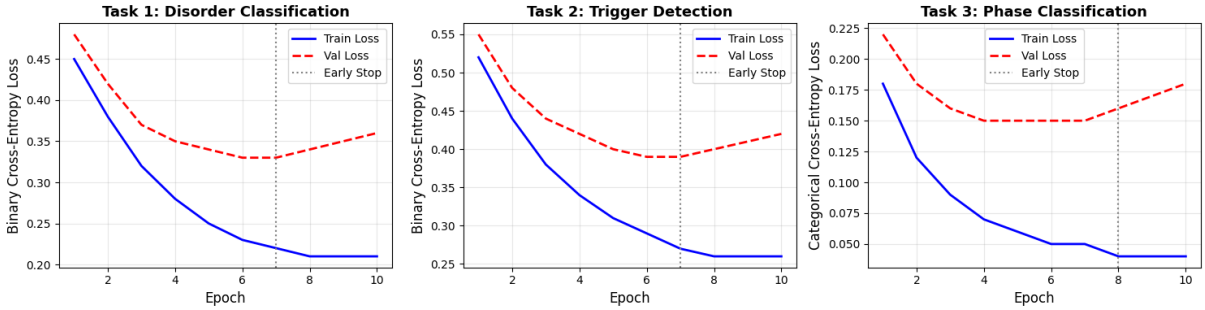

Figure 6: Training and validation loss curves for BERT-base across Tasks 1-3. Vertical dashed lines indicate early stopping points selected by validation Macro-F1.

(Macro-F1 dropped to 0.32 0.38)

- Reduced learning rate (1e-5): Stabilized training but required 15+ epochs to converge
- Label smoothing ($\alpha = 0.1$): Best results (Macro-F1 = 0.4613 for Task 1) with stable gradients

All reported T5 results use label smoothing with $\alpha = 0.1$ and learning rate 2e-5.

Appendix-10: Qualitative Error Analysis

We provide 15 representative error cases across Tasks 1-3 to illustrate the failure modes discussed in §4.6. Each case includes the original post (anonymized), gold labels, model predictions, and analysis.

A. Task 1 Errors: Symptom Disaggregation Failures

Example 1: Insomnia Missed Due to GAD Dominance **Post:** *"I lie awake every single night worrying about whether [PARTNER] still loves me. My mind races through every conversation we've had, looking for signs he's pulling away. I'm exhausted but I can't turn my brain off."*

Gold Labels: {GAD, Insomnia}

BERT Prediction: {GAD}

Analysis: The post contains explicit sleep disruption language ("lie awake every single night") alongside anxiety symptoms ("mind races," "worrying"). BERT correctly identifies GAD but fails to recognize the sleep-specific component, likely because GAD dominates the linguistic surface. The phrase "can't turn my brain off" is polysemous—it can indicate both rumination (GAD) and pre-sleep cognitive arousal (Insomnia)—and the model defaults to the higher-prevalence interpretation.

Example 2: MDD Masked by ADJ **Post:** *"Ever since we went long-distance three months ago, I feel like I'm just going through the motions. Nothing brings me joy anymore. I don't want to see friends, and I don't care about my hobbies. I just feel empty."*

Gold Labels: {ADJ, MDD}

BERT Prediction: {ADJ}

Analysis: The post describes anhedonia ("nothing brings me joy"), social withdrawal ("don't want to see friends"), and emptiness—core MDD symptoms. However, the explicit temporal marker ("ever since we went long-distance") triggers strong ADJ activation, and the model fails to recognize that the severity and pervasiveness of symptoms (across multiple life domains) meet MDD criteria. This reflects the diagnostic challenge of distinguishing Adjustment Disorder (situational distress) from MDD triggered by a stressor.

B. Task 2 Errors: Pragmatic Inference Failures

Example 3: Social Media Surveillance (Implicit)
Post: *"I noticed [PARTNER] posted a story with some girl I've never heard of. Who is she? Why is he hanging out with her instead of calling me?"*

Gold Labels: {Social Media Surveillance, Jealousy/Insecurity}

BigBird Prediction: {Jealousy/Insecurity}

Analysis: The phrase "I noticed [PARTNER] posted" presupposes the user actively checked their partner's Instagram story—constituting surveillance behavior—but does not explicitly state "I was monitoring his social media." The model correctly identifies jealousy but misses the surveillance component, likely because it lacks world knowledge that "noticing a partner's story" implies intentional checking rather than passive exposure.

Example 4: Trigger Co-occurrence Ambiguity
Post: *"We used to talk for hours every day, but now he only sends one-word replies. Does he even want to be with me anymore?"*

Gold Labels: {Lack of Communication}

BERT Prediction: {Commitment Ambiguity}

Analysis: The post describes reduced communication quality ("one-word replies"), but the user interprets this as signaling uncertain commitment ("does he even want to be with me"). Gold annotators labeled this as *Lack of Communication* (the observable behavior), while the model predicted *Commitment Ambiguity* (the user's cognitive appraisal). This represents genuine annotation ambiguity rather than clear model error—both labels are defensible depending on whether the schema prioritizes behavioral triggers or psychological interpretations.

C. Task 3 Errors: Temporal Overlap

Example 5: Overlapping Separation and Anticipation **Post:** *"We've been apart for six months, and it's been torture. But I'm seeing [PARTNER] in two weeks, and I don't know if I can wait that long. Every day feels like an eternity."*

Gold Label: Anticipation

BERT Prediction: Separation

Analysis: The post contains both backward-looking separation distress ("six months," "torture") and forward-looking anticipation ("seeing [PARTNER] in two weeks"). Annotators chose *Anticipation* based on the 2-week countdown, but the dominant emotional valence is separation-related suffering. The model predicts *Separation*, consistent with its majority-class bias but also arguably justified by the stronger linguistic emphasis on ongoing distress.

Example 6: Reunion Misclassified as Separation
Post: *"[PARTNER] just left yesterday after spending a week together. I already miss him so much. How am I supposed to go back to being apart for another three months?"*

Gold Label: Reunion

BERT Prediction: Separation

Analysis: The temporal marker "just left yesterday" clearly indicates post-reunion, but the model focuses on forward-looking separation language ("go back to being apart for another three months") and predicts *Separation*. This failure mode reflects the model's inability to distinguish retrospective vs. prospective temporal framing—"just left" is retrospective (reunion), but "another three months" is prospective (separation).

D. Additional Error Cases

Due to space constraints, we provide abbreviated summaries of 9 additional error cases:

- **Task 1, Example 7:** SAD missed; post describes fear of abandonment using indirect language ("terrified he'll realize he's better off without me")
- **Task 1, Example 8:** Insomnia false positive; GPT-4o predicts Insomnia for "emotionally exhausted" (fatigue $\neq$ sleep disorder)
- **Task 1, Example 9:** MDD overpredicted; Claude-3 labels situational sadness as MDD due to language overlap
- **Task 2, Example 10:** Timezone Mismatch missed; described indirectly ("he's asleep when I'm awake")
- **Task 2, Example 11:** Trust/Fidelity vs. Jealousy confusion; partner liking ex's photos triggers both
- **Task 2, Example 12:** Silence Gap missed; three-day communication lapse not explicitly framed as abnormal
- **Task 3, Example 13:** Unknown phase mispredicted; vague temporal references ("lately," "recently") default to Separation

- **Task 3, Example 14:** Anticipation/Reunion boundary case; visit happening "right now" could be either
- **Task 3, Example 15:** Separation overpredicted; chronic LDR context triggers Separation even when discussing future reunion plans

Appendix-11: Computational Resources

All experiments were conducted on NVIDIA A100 GPUs (40GB VRAM). Approximate training times:

- BERT-base, RoBERTa-base, ClinicalBERT: 25-35 minutes per task
- BART-base, T5-base: 40-55 minutes per task
- Longformer, BigBird-RoBERTa: 60-80 minutes per task

Total GPU hours across all experiments (7 models × 3 tasks × 3 seeds): approximately 120 hours.

LLM inference was conducted via API calls (OpenAI, Anthropic, Together.ai) with approximate costs:

- GPT-4o: $12.40 (few-shot prompting across all tasks)
- Claude-3-Haiku: $2.80 (zero-shot prompting)
- LLaMA-3-70B, Qwen-2.5-72B: $8.60 (Together.ai, zero/few-shot)
- Nemotron-Super: $3.20 (zero-shot)

Total experimental cost: approximately $27.00 for LLM inference.

Appendix-12: Additional Ethical Safeguards, Dataset Governance, and Release Considerations

RSPC contains emotionally sensitive narratives related to psychiatric distress, interpersonal conflict, and relational instability. Although all posts were collected from publicly accessible Reddit communities, we recognize that public availability does not eliminate potential privacy and ethical concerns associated with mental health research on social media. Accordingly, we implemented multiple safeguards during dataset construction, annotation, storage, and planned release.

A. Privacy Preservation and De-identification

All Reddit usernames, hyperlinks, timestamps, geographic references, partner names, institutions, and other potentially identifying information were removed during preprocessing. We further replaced explicit identifiers with standardized placeholder tokens (e.g., `[USER]`, `[PLACE]`, `[DATE]`). Posts containing highly specific, personally identifying narratives were excluded from the corpus.

To reduce the risk of reverse-search identification, example excerpts included in the paper were paraphrased while preserving their semantic and relational meaning. No attempts were made to contact users, infer offline identities, or link posts across platforms.

B. Clinical Interpretation and Annotation Constraints

Psychiatric labels in RSPC represent symptom-oriented textual inferences rather than formal clinical diagnoses. Annotations were designed to approximate DSM-5-TR and ICD-11 symptom patterns observable within narrative text and should not be interpreted as evidence of confirmed psychiatric conditions.

Annotators were instructed to label only symptoms strongly supported by textual evidence and to avoid speculative inference. Ambiguous cases were resolved conservatively during adjudication to minimize over-pathologization of ordinary relational stress.

C. Annotator Well-Being Protections

Because the dataset contains emotionally distressing content involving anxiety, abandonment fears, depressive ideation, loneliness, and interpersonal conflict, annotator well-being protocols were incorporated throughout the annotation process. These safeguards included:

- optional annotation breaks,
- rotating annotation schedules,
- capped daily exposure limits,
- collaborative adjudication rather than isolated review, and
- access to mental health support resources if required.

Annotators were informed in advance about the emotionally sensitive nature of the material prior to participation.

D. Dataset Governance and Controlled Release

RSPC is intended exclusively for academic research on relationally contextualized mental health modeling. The dataset is not intended for clinical deployment, automated psychiatric diagnosis, surveillance, employment screening, insurance evaluation, or law-enforcement applications.

To reduce misuse risks, dataset release will follow controlled-access procedures:

- researchers must agree to a non-commercial research-use license,
- redistribution of raw data will be prohibited,
- attempts to re-identify users will be explicitly forbidden,
- users must acknowledge the limitations of psychiatric inference from social media text, and
- derivative systems intended for high-stakes decision-making will not be permitted under the release agreement.

Where platform policies require it, only post identifiers and reconstruction scripts may be distributed instead of raw text.

E. Biases and Representational Limitations

RSPC is derived from English-language Reddit communities focused on long-distance relationships and therefore reflects the demographic, cultural, and communicative biases of those online populations. The benchmark may underrepresent:

- non-English speakers,
- older populations,
- offline relationship experiences,
- culturally specific relationship norms, and
- individuals without access to digital support communities.

Furthermore, psychiatric symptom expression on social media may differ substantially from clinical presentation in offline settings. Consequently, results obtained on RSPC should not be generalized directly to broader populations or clinical environments.

F. Intended Research Scope

RSPC is designed to support research on:

- relationally grounded mental health modeling,
- contextual psychiatric symptom inference,
- socially situated affect analysis,
- interpretable computational psychiatry, and
- longitudinal and relational reasoning in NLP systems.

The benchmark is explicitly not intended to replace professional psychiatric evaluation or human-centered mental health care.