Overview of the PsyDefDetect Shared Task at BioNLP 2026: Detecting Levels of Psychological Defense Mechanisms in Supportive Conversations

Hongbin Na[1,*] **Zimu Wang**[2,*] **Zhaoming Chen**[3] **Yining Hua**[4]
Rena Gao[5] **Kailai Yang**[6] **Ling Chen**[1] **Wei Wang**[2] **Shaoxiong Ji**[7,8]
John Torous[4] **Sophia Ananiadou**[6]

[1]University of Technology Sydney [2]Xi'an Jiaotong-Liverpool University
[3]University of Utah [4]Harvard University [5]The University of Melbourne
[6]The University of Manchester [7]ELLIS Institute Finland [8]University of Turku
Hongbin.Na@student.uts.edu.au, Zimu.Wang@liverpool.ac.uk

Abstract

We present an overview of PSYDEFDETECT, the shared task on detecting levels of psychological defense mechanisms in emotional support dialogues, co-located with BioNLP@ACL 2026. Grounded in the clinically validated Defense Mechanism Rating Scales (DMRS) framework, the task asks systems to classify a target seeker utterance, given its preceding dialogue context, into one of nine categories: seven hierarchical DMRS levels plus two auxiliary labels. Participants worked on PSYDEFCONV, a newly released corpus of 200 dialogues and 2,336 help-seeker utterances annotated under DMRS with substantial inter-annotator agreement. The task attracted 172 participants on CodaBench who produced 563 submissions, with 21 teams officially registering their results for the final ranking. The best system achieved a macro F1-score of 0.420, surpassing the strongest fine-tuned baseline reported in the dataset paper by a notable margin, yet leaving clear headroom. Our analysis highlights (i) a persistent tendency to over-predict the majority *High-Adaptive* class, (ii) a widening gap between accuracy and macro-F1 that reveals class-imbalance sensitivity, and (iii) the value of theory-aware and LLM-based approaches for fine-grained defensive-function classification. We release all task materials and invite the community to continue work on this novel intersection of clinical psychology and NLP.[1]

1 Introduction

Psychological defenses are, in Winnicott's words, the means by which the *False Self*, if successful in its function, hides the *True Self* (Winnicott, 2018). They are automatic strategies people use to regulate distress, and when used rigidly or excessively they are linked to poorer mental health outcomes and to interpersonal difficulties (Perry and Henry, 2004; Di Giuseppe et al., 2024). In emotional support conversations (ESC) (Liu et al., 2021), defenses shape what help-seekers disclose and how they accept or resist help—yet, despite rapid progress on empathy modeling (Hua et al., 2025b; Cai et al., 2024; Sorin et al., 2024), strategy selection (Kang et al., 2024; Hua et al., 2025a; Na et al., 2025), and affect understanding (Wang et al., 2024a, 2025; Ma et al., 2025; Zhao et al., 2025), the defensive function of seeker utterances remains largely unmodeled in current ESC systems (Di Giuseppe et al., 2024).

To catalyze research on this under-explored dimension of supportive dialogue, we organized **PSYDEFDETECT**, a shared task co-located with BioNLP@ACL 2026 on detecting *levels* of psychological defense mechanisms. The task is grounded in the clinically validated Defense Mechanism Rating Scales (DMRS) (Perry and Henry, 2004; Vaillant, 2012), and asks systems to classify each help-seeker utterance—given its preceding dialogue context—into one of nine categories: the seven hierarchical DMRS levels, augmented with two auxiliary labels for phatic and under-specified turns. The task is built on PSYDEFCONV (Na et al., 2026), the first conversational corpus annotated with DMRS-based defense levels. PSYDEFCONV is derived from ESConv (Liu et al., 2021) via stratified sampling over problem types and emotions, and was double-blind annotated to substantial agreement.

Participation. PSYDEFDETECT attracted strong community engagement. The task was hosted on CodaBench (Xu et al., 2022) from December 2025 through April 2026. Over the course of the evaluation period, the competition registered **172 participants** who together produced **563 submissions** to the leaderboard. By the final registration deadline, **21 teams** had officially submitted their results for inclusion in the ranking. The best system reached

[*]Co-leads of the shared task organization.

[1]Task website: https://psydefdetect-shared-task.github.io/; CodaBench: https://www.codabench.org/competitions/12124/.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 932–943

a macro F1-score of 0.420, substantially surpassing the strongest fine-tuned baseline reported in the dataset paper (≈ 0.315 macro-F1 (Na et al., 2026)), while still leaving ample headroom for future research.

Contributions. This paper reports on the organization, dataset, participating systems, and results of PSYDEFDETECT. Our contributions are as follows:

- We introduce PSYDEFDETECT, the first shared task that operationalizes DMRS-based defensive functioning as an utterance-level classification problem over emotional support dialogue.
- We report task-level participation that reflects substantial community engagement, with 172 CodaBench participants, 563 leaderboard submissions, and 21 officially ranked teams.
- We present a methodological taxonomy of participating systems together with a full leaderboard analysis that benchmarks submissions against both zero-shot and fine-tuned baselines from the dataset paper.
- We characterize the dominant failure modes observed across the leaderboard, in particular the severe sensitivity to class imbalance and the systematic over-prediction of the High-Adaptive level, and use them to motivate concrete directions for future work at the intersection of clinical psychology and NLP.

2 Background and Related Work

Defense mechanisms and DMRS. Defense mechanisms are a cornerstone of psychodynamic theory (Freud, 1936). The *Defense Mechanism Rating Scales* (DMRS) organize roughly thirty mechanisms into seven hierarchical levels of defensive maturity, from *Action Defenses* (Level 1) to *High-Adaptive Defenses* (Level 7) (Perry and Henry, 2004; Vaillant, 2012). DMRS was originally designed for longitudinal clinical case formulation rather than for single-utterance classification; PSYDEFCONV (Na et al., 2026) adapts the scheme to conversational text by annotating at the *level* rather than the *mechanism* granularity, increasing identifiability and reliability.

Emotional support conversations. Research on ESC has built increasingly capable dialogue agents that alleviate user distress through multi-turn, strategy-grounded interaction, starting with the ESConv corpus (Liu et al., 2021) and extending to multi-strategy (Bai et al., 2025), synthetic (Zheng et al., 2023; Wang et al., 2024a), and reasoning-aware (Zhang et al., 2024) variants. These efforts have largely focused on *supporter* strategies, empathy, and affect modeling (Wang et al., 2024a; Cai et al., 2024; Sorin et al., 2024; Hua et al., 2025b; Xu et al., 2026), while the *defensive function* of seeker utterances has remained unmodeled.

Mental health classification. Prior shared tasks at BioNLP and CLPsych have addressed clinically-motivated classification problems such as suicide risk assessment, depression detection, and empathy prediction. BioNLP has covered a wide spectrum of medical-related challenges, such as clinical question-answering and summarization (Colelough et al., 2025; Soni et al., 2025; Xiao et al., 2025). However, tasks specifically addressing mental health have remained notably underrepresented. CLPsych shared tasks have addressed a range of mental-health-related problems, including depression and PTSD (Coppersmith et al., 2015), suicide risk assessment (Zirikly et al., 2019; Chim et al., 2024), online peer support (Milne et al., 2016), developmental mental-health prediction (Lynn et al., 2018), and longitudinal affective modeling (Tsakalidis et al., 2022; Tseriotou et al., 2025). However, these efforts have primarily centered on social media platforms such as X and Reddit, with comparatively limited attention to ESC. PSYDEFDETECT complements this line of work by introducing a theory-grounded, fine-grained classification problem that requires both local linguistic reasoning and context-aware interpretation.

3 Task Description

3.1 Problem Formulation

Given a multi-turn emotional support dialogue $D = (u_1, u_2, \ldots, u_t)$ between a help-*seeker* and a *supporter*, and a target help-seeker utterance u_t, the task is to predict a defense level label $y \in \mathcal{Y}$, where $\mathcal{Y}$ comprises the seven DMRS levels and two auxiliary labels:

- **Level 0 – No Defenses:** phatic or functional utterances that do not engage with psychological conflict.

- **Level 1 – Action Defenses:** passive aggression, help-rejecting complaining, acting out.
- **Level 2 – Major Image-Distorting:** splitting, projective identification.
- **Level 3 – Disavowal:** denial, rationalization, projection, autistic fantasy.
- **Level 4 – Minor Image-Distorting:** devaluation, idealization, omnipotence.
- **Level 5 – Neurotic:** repression, dissociation, reaction formation, displacement.
- **Level 6 – Obsessional:** isolation of affect, intellectualization, undoing.
- **Level 7 – High-Adaptive:** affiliation, altruism, anticipation, humor, self-assertion, self-observation, sublimation, suppression.
- **Level 8 – Needs More Information:** context is insufficient to assign a label with confidence.

Systems have access to the dialogue context preceding and including u_t, but must not use future turns—this mirrors the online nature of the clinical annotation setup (Na et al., 2026).

4 The PSYDEFCONV Dataset

4.1 Construction

PSYDEFCONV (Na et al., 2026) is built on top of ESConv (Liu et al., 2021). The organizers performed stratified sampling over the joint distribution of problem types and emotions in ESConv to obtain a representative 200-dialogue subset, annotated at the *seeker*-utterance level for defense functioning.

Two trained annotators with expertise in both psychology and NLP labeled each seeker turn independently and double-blind, reaching Cohen's $\kappa = 0.639$ (substantial agreement). Disagreements were adjudicated by consensus to form the gold standard. The annotation workflow was supported by DMRS CO-PILOT (Na et al., 2026), a four-stage LLM pipeline that produces stressor hypotheses, screens candidate DMRS items, validates evidence, and synthesizes ranked recommendations; it reduced mean annotation time by 24.0%.

Category	Total	Supporter	Seeker
# Dialogues	200	–	–
# Utterances	4,709	2,373	2,336
Avg. Turns per Dialogue	23.5 ± 6.6	11.9 ± 3.4	11.7 ± 3.3
Avg. Length of Utterances	19.8 ± 16.5	20.9 ± 17.0	18.8 ± 15.8

Table 1: Data statistics of PSYDEFCONV.

4.2 Corpus Statistics

Table 1 summarizes the basic statistics of PSYDEFCONV. The corpus contains 200 dialogues and 4,709 utterances, roughly evenly divided between supporter (2,373) and seeker (2,336) turns. Conversations are of moderate length, with a mean of 23.5 turns per dialogue and an average utterance length of 19.8 tokens. Only the 2,336 seeker turns carry DMRS defense-level annotations; supporter turns are provided as dialogue context but are not part of the evaluation.

4.3 Label Distribution

Table 2 reports the distribution of the 2,336 annotated seeker utterances across the nine defense labels, as well as their aggregation into four broader defensive categories. The distribution is strongly skewed toward the *High-Adaptive* level (Level 7), which alone accounts for 51.8% of all annotated utterances. Non-defensive or contextually ambiguous turns (Levels 0 and 8) jointly cover another 17.4%, leaving only roughly one third of the corpus for the remaining six defense levels. Within that tail, *Immature* defenses (Levels 1–4) aggregate to 18.9% while *Neurotic* defenses (Levels 5–6) account for 11.9%, with the individual minority levels ranging from 2.6% (Neurotic) to 9.2% (Obsessional).

This imbalance reflects the natural prevalence of defensive functioning in supportive dialogue (Na et al., 2026): adaptive coping such as self-assertion, self-observation, and seeking affiliation is, by design, what participants in ESC conversations most often express. However, the imbalance also means that systems optimized for overall accuracy can trivially favor the dominant class at the expense of minority defense levels. We adopt macro-averaged F1 over the positive classes (1–8) as the official ranking metric (Section 5) precisely to penalize such majority-class bias and to encourage systems that discriminate well across the full DMRS hierarchy.

Categories	Num	Proportion
Seeker's Defense Levels		
0 No Defenses	371	15.9%
1 Action Defenses	136	5.8%
2 Major Image-Distorting	77	3.3%
3 Disavowal	124	5.3%
4 Minor Image-Distorting	105	4.5%
5 Neurotic	61	2.6%
6 Obsessional	216	9.2%
7 High-Adaptive	1,211	51.8%
8 Needs More Information	35	1.5%
Overall	2,336	100.0%
Seeker's Defense Categories		
Non-Defensive/Ambiguous (0, 8)	406	17.4%
Mature Defenses (7)	1,211	51.8%
Neurotic Defenses (5, 6)	277	11.9%
Immature Defenses (1, 2, 3, 4)	442	18.9%
Overall	2,336	100.0%

Table 2: Distribution of the PSYDEFCONV dataset across annotated defense levels and aggregated defense categories.

5 Evaluation Setup

5.1 Platform and Protocol

The shared task was hosted on CodaBench (Xu et al., 2022), where participants downloaded the dataset, received starter baseline kits, and submitted predictions as zipped JSON files. Each team could register on the official leaderboard after completing the result-registration form.

5.2 Metrics

Following the dataset paper (Na et al., 2026), we report Accuracy, Macro Precision, Macro Recall, and Macro F1-score. Macro F1 evaluated over the positive classes (1–8) is the official ranking metric, chosen because it penalizes majority-class bias, weighs minority defensive levels equally with the dominant High-Adaptive level, and follows established practices for multi-class classification (Wang et al., 2022, 2024b; Chen et al., 2025). Level 0 is excluded from the official macro-F1 because it marks phatic or functional turns without defensive content. Level 8, by contrast, is retained because detecting that the available context is insufficient is an explicit task decision rather than a negative class. Furthermore, we include an additional leaderboard that aggregates performance across all classes, enabling a more comprehensive and holistic comparison of the submitted systems.

5.3 Baselines

The CodaBench page provided zero-shot prompting baselines using strong general-purpose LLMs. These baselines follow the same setup as the dataset paper; for reference, the strongest results reported in (Na et al., 2026) are as follows. Under zero-shot prompting, Gemini 2.5 Pro reaches an accuracy of 0.5636 and a macro-F1 of 0.2599, while DeepSeek-V3.2 with thinking enabled achieves comparable performance at 0.5572 accuracy and 0.2617 F1. Under supervised fine-tuning, the best results come from Ministral-8B (Acc = 0.6483, F1 = 0.3148) and InternLM3-8B (Acc = 0.6398, F1 = 0.3053).

6 Participating Systems

6.1 Participation Statistics

The PSYDEFDETECT shared task drew broad international engagement. The CodaBench competition registered **172 participants** and received **563 submissions** over the course of the evaluation period, with **21 teams** officially registering their results for the final ranking (Table 3). Of these, **15** accompanied their submission with a system description paper contributed to this volume. The author affiliations on these 15 papers span 12 countries—Australia, Bangladesh, Canada, China, Germany, India, Russia, Spain, Switzerland, the UK, the USA, and Vietnam—and 24 distinct institutions, including a number of cross-institution collaborations that combine academic NLP groups with clinical and biomedical informatics partners.

6.2 Methodological Taxonomy

The 15 described systems span a broad methodological space, but gravitate toward six recurring patterns, described below. Almost every team targets the same two task-specific pain points: the heavy skew toward Level 7 ("High-Adaptive"), and the ambiguity of DMRS boundaries for short, context-dependent utterances. Several teams combine two or more strategies, in which case we record the most distinctive contribution here.

(i) Multi-model ensembles and deliberative-agent architectures. Top-ranked systems reject the idea that a stronger single model can solve the task, and instead explicitly engineer *error independence* across voters or agents. NÜRNBERG NLP (ranked 1st; Steigerwald et al., 2026) constructs a 9-voter ensemble along three orthogonal axes—class granularity (9-class *gatekeeper* vs. 8-class *special-*

ists), training paradigm (generative vs. discriminative), and base model (Ministral vs. Phi-4)—with per-axis cross-validation folds. UTS (2nd; Galat and Rizoiu, 2026) operates a multi-phase deliberative council of Gemini 2.5 agents in which class-specific *advocates* rate evidence strength rather than vote, augmented by a targeted override ensemble of three fine-tuned Qwen-family models. TONI-NLP (9th; Paul et al., 2026) likewise finds that ensembles outperform any individual prompting, fine-tuning, or embedding-classifier baseline.

(ii) Retrieval- and rubric-grounded LLM classification. A second cluster of systems brings the DMRS clinical rubric into the prompt, either dynamically via retrieval or statically via curated boundary cues. PERCEPTIONLAB (3rd; Fahim et al., 2026) pairs dynamic DMRS-Q item retrieval (Gemini 2.5 Pro) with a Gemini 2.5 Flash classifier fine-tuned on reasoning traces distilled from the same Pro model, explicitly targeting "LLM polarization" toward extreme labels. DAL TEAM (10th; Chu et al., 2026) runs a retrieval-augmented LLM pipeline that decomposes prediction into a coarse-to-fine hierarchy and adds summary-based distillation of dialogue context. ZZUCS (6th; Huang et al., 2026) uses its *CoR-QLoRA* approach to encode task contracts, taxonomy definitions, and adjacent-level boundary cues into the prompts used for 8B QLoRA adaptation.

(iii) Parameter-efficient LLM fine-tuning. Most mid-to-top-ranked systems converge on QLoRA or equivalent PEFT on mid-scale open LLMs. LINGUIUTICS (4th; Adib et al., 2026) QLoRA-fine-tunes Qwen3-8B with grouped stratified cross-validation, minority-class lexical augmentation, and a post-hoc logit-bias calibration step that substantially improves the rare "Needs More Information" class. ERASERHEAD (7th; Horaira et al., 2026) fine-tunes Qwen3-14B with clinically informed prompts and iteratively retuned per-class oversampling. TRANSFORMER_1376 (12th; Saha et al., 2026) reformulates the task as conditional text generation and applies 4-bit QLoRA to Gemma-2-2B.

(iv) Encoder-only and domain-specific transformer fine-tuning. Three teams explore encoder backbones. NEURAL NEXUS (11th; Basu, 2026) fine-tunes ROBERTA-BASE with a composite objective combining focal loss, label smoothing, and square-root-dampened class weights, using role-tagged dialogue history and a `[TARGET]` marker on the target utterance. CS_METRO (15th; Rebayet et al., 2026) builds a three-stage pipeline of LLM-based dialogue summarization, domain-specific transformer fine-tuning (including MentalBERT and Mental-RoBERTa variants), and rule-based ensembling; they report that domain-specific encoders outperform generic LLM fine-tuning on this clinical task. KCL-COGSTACK (17th; Agarwal et al., 2026) contrasts flat fine-tuning, few-shot prompting, and a hierarchical coarse-to-fine classifier that exploits the DMRS label tree, finding the hierarchical design the most effective of the three.

(v) Synthetic data and theory-informed augmentation. VISHC (13th; Vu et al., 2026) focuses on data scarcity rather than model scale, proposing stressor-anchored, theory-driven synthetic data generation combined with a hybrid model that fuses language representations with structured clinical features. Lexical augmentation and oversampling are also used as secondary ingredients by LINGUIUTICS, ZZUCS, TRANSFORMER_1376, and ERASERHEAD.

(vi) Systematic exploration and negative-result studies. Two teams emphasize breadth of comparison over a single headline system. ALIEN-ANNOTATORS (19th; Karip and Hossain, 2026) systematically evaluates six open-source small language models ($\leq$ 9B) under zero-shot and fine-tuning regimes, finding that clinically grounded prompts consistently outperform bare label definitions and that model scale alone does not help zero-shot performance; their post-submission configuration (fine-tuning with 5-fold CV and logit averaging) reaches macro-F1 = 0.346, more than double their official submission. EXPLAINATORS (20th; Babakova et al., 2026) similarly runs four exploratory tracks—direct prompting, encoder fine-tuning, novel "state-of-mind" generation, and LLM fine-tuning—across DeepSeek-V3.2, Qwen-family models, and GLM-series models.

7 Results

7.1 Official Leaderboard

Table 3 lists the final leaderboard, ranked by macro F1-score on the held-out test set. The top system (Nürnberg NLP, F1 = 0.4200) outperforms the second-ranked team (UTS, F1 = 0.4055) by a small margin, while the overall spread of F1 scores is wide (0.063–0.420), underscoring the difficulty of

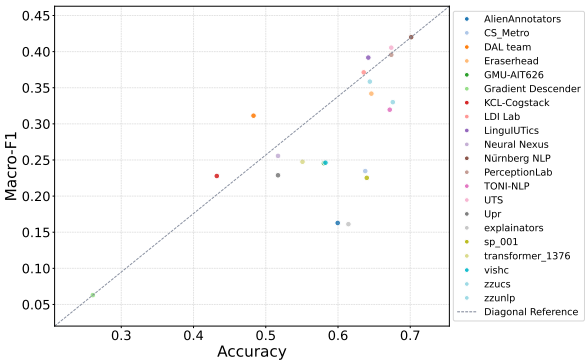


Figure 1: Scatter plot illustrating the relationship between accuracy and macro F1-scores across all submitted systems.

the task.

7.2 Key Observations

Top systems substantially outperform established baselines. The best-performing system, introduced by Nürnberg NLP, achieves a macro F1-score of 0.420, surpassing the best fine-tuned baseline reported in the dataset paper (Ministral-8B, F1 = 0.315) by approximately **10.5 absolute points**, and the best zero-shot baseline by **16 points**. These gains suggest that participants were able to leverage additional task-specific signals beyond standard fine-tuning setups. Nevertheless, the overall performance ceiling remains modest in absolute terms, indicating that the task continues to pose significant challenges.

Wide variability in system performance. System performance exhibits a wide dispersion, with macro F1-scores ranging from 0.063 (GRADIENT DESCENDER) to 0.420 (NÜRNBERG NLP), corresponding to a $6.7\times$ difference. The distribution further reveals a clear quartile structure: the top quartile achieves F1-scores of ≥ 0.37, the median lies ≈ 0.25, and the bottom quartile falls ≤ 0.23. This pronounced spread indicates that, despite the intrinsic difficulty of the task, there remains considerable room for methodological differentiation and performance gains across approaches.

Accuracy is a misleading proxy for F1-score. Several submissions attain relatively high accuracy while exhibiting substantially lower F1-scores, as shown in Figure 1. For instance, ZZUNLP (Acc = 0.676, F1 = 0.330) and TONI-NLP (Acc = 0.672, F1 = 0.320) demonstrate this discrepancy. This mismatch stems from the pronounced class imbalance, with the Level 7 majority class accounting for

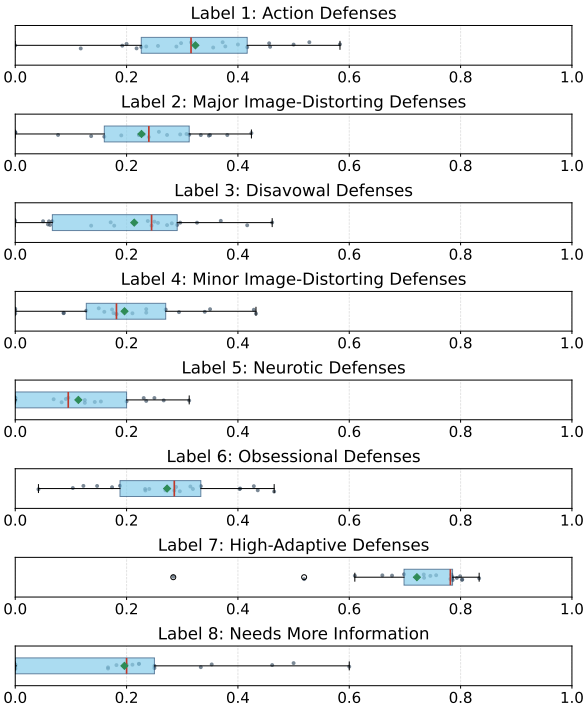


Figure 2: Per-class F1-scores across all submitted systems for the positive classes 1–8. Each box summarizes, for a given defense level, the distribution of F1 values over the 21 leaderboard submissions.

$\approx$52% of the dataset. Models that over-predict the High-Adaptive class can achieve competitive accuracy, but suffer from poor recall on minority classes, thereby degrading macro F1-scores. In contrast, systems such as NÜRNBERG NLP, PERCEPTIONLAB, and LINGUIUTICS achieve a more favorable balance, maintaining strong accuracy alongside improved recall across all classes.

8 Analysis and Discussion

8.1 Per-Class Performance

Figure 2 summarizes per-class F1 across the 21 submitted systems. Level 7 (High-Adaptive) stands apart: its median F1 is by far the highest and the inter-team spread is narrow, reflecting both its majority status ($\approx$52% of the corpus) and the distinctiveness of its surface markers. Every other class has a low median and a wide spread—widest for Levels 2–4, whose linguistic realizations overlap in affective-blame and reality-distortion cues, and for Level 8 (Needs More Information), on which a non-trivial fraction of systems collapse to zero. The narrow box for Level 7 marks a class where surface cues suffice; the wide boxes (Levels 1–6) mark classes where methodological choices translate directly into outcome differences.

			Positive classes (1–8)			All classes (0–8)		
Rank	**Team**	**Acc**	**P**	**R**	**F1**	**P**	**R**	**F1**
1	Nürnberg NLP	0.7013	0.4510	0.4036	0.4200	0.4959	0.4639	**0.4732**
2	UTS	0.6737	0.4607	0.3884	0.4055	0.4915	0.4327	0.4450
3	PerceptionLab	0.6737	0.4263	0.4089	0.3956	0.4721	0.4479	0.4402
4	LinguIUTics	0.6419	0.4003	0.3958	0.3917	0.4416	0.4570	0.4427
5	LDI Lab	0.6356	0.3770	0.3892	0.3713	0.4277	0.4422	0.4244
6	zzucs	0.6441	0.3969	0.3520	0.3585	0.4402	0.4166	0.4135
7	Eraserhead	0.6462	0.4075	0.3193	0.3418	0.4482	0.3801	0.3947
8	zzunlp	0.6758	0.4991	0.2891	0.3300	0.5381	0.3578	0.3909
9	TONI-NLP	0.6716	0.4701	0.2839	0.3196	0.5094	0.3560	0.3813
10	DAL team	0.4831	0.4187	0.2773	0.3113	0.4165	0.3516	0.3391
11	Neural Nexus	0.5169	0.2480	0.2867	0.2556	0.3140	0.3259	0.3080
12	transformer_1376	0.5508	0.2669	0.2351	0.2475	0.3132	0.2890	0.2979
13	VISHC	0.5826	0.2588	0.2503	0.2462	0.3168	0.3069	0.3045
14	GMU-AIT626	0.5805	0.2844	0.2328	0.2455	0.3298	0.2839	0.2952
15	CS_Metro	0.6377	0.3001	0.2572	0.2346	0.3673	0.3131	0.3004
16	Uprm	0.5169	0.2600	0.2161	0.2288	0.3081	0.2854	0.2877
17	KCL-Cogstack	0.4322	0.2913	0.2517	0.2278	0.3359	0.3171	0.2868
18	sp_001	0.6398	0.3057	0.2309	0.2253	0.3617	0.3060	0.2953
19	AlienAnnotators	0.5996	0.1555	0.1975	0.1628	0.2027	0.2823	0.2251
20	explainators	0.6144	0.2366	0.1660	0.1612	0.3041	0.2275	0.2296
21	Gradient Descender	0.2606	0.1383	0.0455	0.0629	0.1465	0.1501	0.0946
Baseline – Ministral-8B (fine-tuned) (Na et al., 2026)		0.6483	0.3397	0.3045	0.3148	0.3978	0.3640	0.3745
Baseline – Gemini 2.5 Pro (zero-shot) (Na et al., 2026)		0.5636	0.2749	0.2612	0.2599	0.2907	0.3107	0.2893

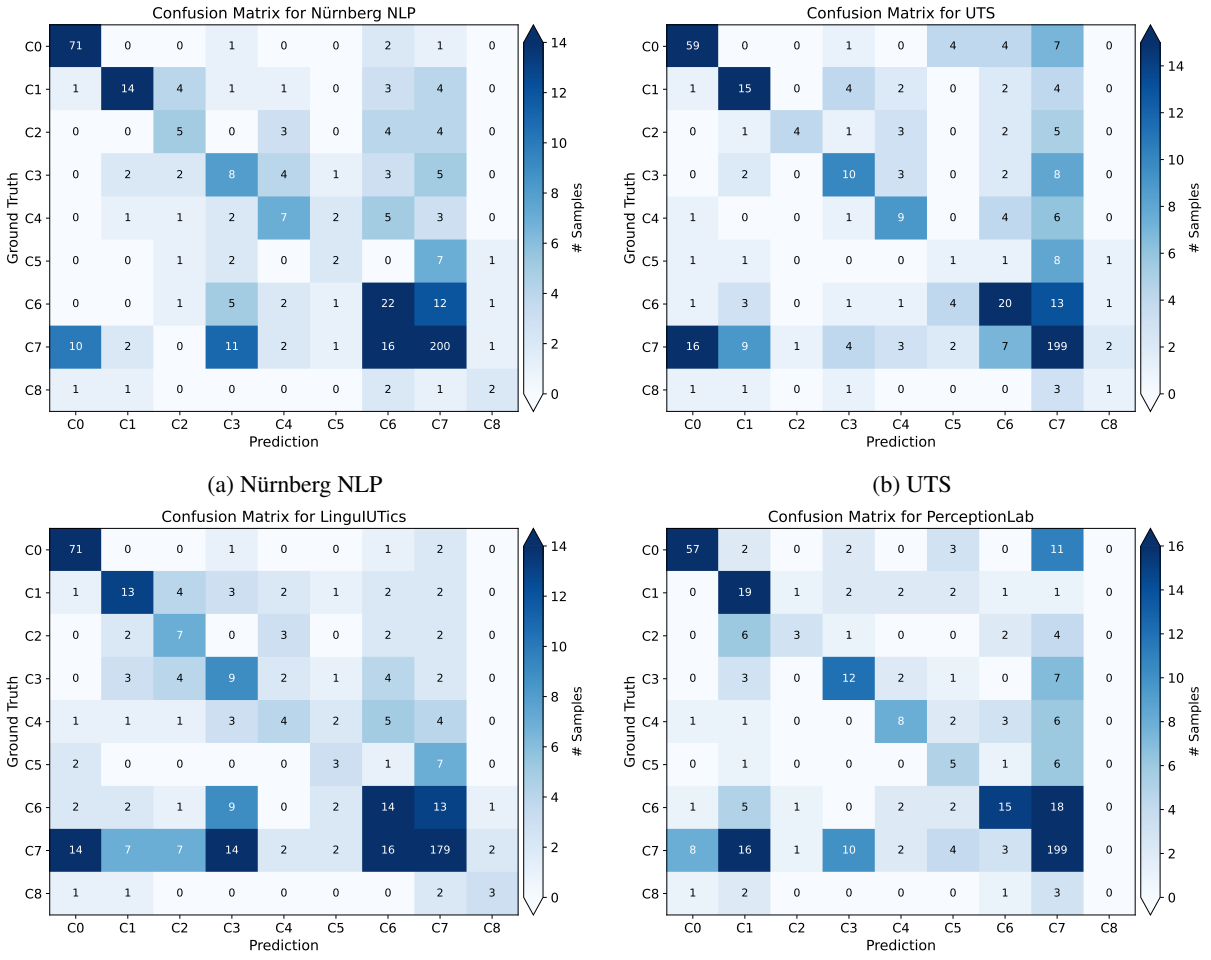
Table 3: Official PSYDEFDETECT leaderboard. Systems are ranked by macro-F1 over the *positive* classes 1–8 (the official metric of the shared task); for reference we additionally report macro-averaged precision, recall, and F1 computed over *all* nine classes (0–8). Accuracy is shared across both metric groups. The winning system's official macro-F1 of 0.4200 exceeds the strongest fine-tuned baseline reported in the dataset paper by $\approx$10.5 absolute points; its all-class F1 of 0.4732 exceeds that baseline by $\approx$9.9 points.

8.2 Error Patterns

Figure 3 shows confusion matrices for the four top-ranked systems. A single error pattern dominates all four: systematic *over-prediction of the High-Adaptive level* (C7), the "L7 attractor" identified by the dataset paper as the central failure mode of language models on PSYDEFCONV (Na et al., 2026). More interesting is *how* the four systems trade this attractor against minority-class recall. LINGUIUTICS is the most conservative on C7 (diagonal of 179 vs. $\approx$200 for the others), a direct effect of its minority-class augmentation and logit-bias calibration; the cost is some C7 recall, but the benefit is improved recovery of Level 8 instances. PERCEPTIONLAB shows a different signature: its dynamic DMRS-Q retrieval produces markedly higher Level 1 recall (19/28), while also exhibiting a wider over-prediction of C7. No single design choice resolves the attractor; different strategies trade one form of error against another, and the choice determines *which* minority classes a system can recover.

8.3 What Worked, What Did Not

Three takeaways emerge across the 15 system papers. First, *ensembling is the strongest differentiator at the top*: the two highest-scoring systems both engineer error independence explicitly—NÜRNBERG NLP along three orthogonal axes, and UTS through a deliberative agent council—while TONI-NLP (Paul et al., 2026) reaches the same conclusion through portfolio comparison. Second, *task-specific supervision beats backbone scale*: mid-scale backbones (Qwen3-8B, Ministral-8B, Qwen3-14B) dominate the top half of the leaderboard, and ALIENANNOTATORS (Karip and Hossain, 2026) report directly that scale alone does not reliably help zero-shot performance whereas clinically grounded prompts consistently do. Third, *zero-shot prompting alone is insufficient*: no system that relied primarily on zero-shot prompting of a large LLM exceeded F1 = 0.30, and every described top-10 system combines prompting with fine-tuning, retrieval, or ensembling. Clinical theory, packed into a prompt, is necessary but must be coupled with task-specific supervision to translate into accurate classification.



(a) Nürnberg NLP

(b) UTS

(c) LinguIUTics

(d) PerceptionLab

Figure 3: Confusion matrices for the four top-ranked systems on the held-out test set. Rows denote ground-truth labels, columns denote predictions. Color scales are clipped for readability; cell annotations show raw counts.

9 Future Directions

Four directions stand out as natural next steps. (1) *Dialogue-level defensive trajectories*: moving from utterance-level classification to modeling how defenses evolve across a conversation, building on the emotion-level trajectory analysis already included in the dataset paper (Na et al., 2026). (2) *Multilingual* PSYDEFCONV: annotating comparable corpora in Chinese, Japanese, Spanish, and other languages, with DMRS CO-PILOT lowering the clinical-annotation cost that has historically blocked cross-lingual scaling. (3) *Coupling defenses and supporter strategies*: jointly modeling how supporters adapt their strategies to seeker defensive functioning, closing the loop back to the broader ESC research programme. (4) *Theory-aware generation*: using defense-level predictions to condition emotionally supportive responses, operationalizing the seeker-model → supporter-response direction outlined as future work in the dataset paper. The first and fourth directions, in particular, turn PSYDEFDETECT from a static classification benchmark into a building block for clinically informed dialogue systems.

10 Conclusion

We introduced PSYDEFDETECT, the first shared task on detecting DMRS-grounded psychological defense levels in emotional support dialogues. The task drew 172 CodaBench participants, 21 ranked teams, and 15 system description papers across 12 countries. The winning system reached macro-F1 = 0.420, a ≈10.5-point gain over the strongest fine-tuned baseline, yet leaves clear headroom on minority classes where the "L7 attractor" dominates. We release all task materials to support continued work at the intersection of clinical psychology and NLP.

Limitations

This shared task has several limitations. First, PSYDEFCONV is a relatively small corpus derived from English ESConv dialogues, so results may not generalize to other languages, cultures, clinical settings, or naturally occurring therapy conversations. Second, the task labels defenses at the DMRS level rather than at the individual-mechanism level. This choice improves annotation reliability, but it also collapses clinically meaningful distinctions within each level. Third, defense interpretation remains inherently contextual: even with trained annotators and adjudication, short seeker utterances can be ambiguous without broader personal or longitudinal information. Finally, our methodological analysis is based on the 15 teams that submitted system description papers; leaderboard entries without accompanying papers are included in the quantitative ranking but cannot be analyzed in the same level of detail.

Acknowledgments

We thank all 21 participating teams and the 15 teams that contributed system description papers for making this shared task a success. We thank the BioNLP workshop organizers for hosting PSYDEFDETECT. We gratefully acknowledge the *Label Studio Academic Program* for providing access to the annotation platform used to construct PSYDEFCONV.

Ethical Considerations

Data. PSYDEFCONV is derived from ESConv (Liu et al., 2021) under its stated terms. Only derived annotations, dialogue identifiers, and code are redistributed; downstream users must obtain ESConv separately and comply with its license.

Intended use. The dataset and resulting models are intended *solely for research* on language and defensive functioning. They are **not diagnostic tools** and must not be used to make clinical, legal, or employment decisions about individuals. The corpus is in English and reflects the domains covered by ESConv, so results may not generalize across cultures or clinical settings.

References

Shefayat E Shams Adib, Ahmed Alfey Sani, Md Hasibur Rahman Alif, and Ajwad Abrar. 2026. LinguIUTics at PsyDefDetect: Iterative imbalance-aware finetuning of Qwen3-8B for psychological defense mechanism classification. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Shubham Agarwal, Thomas Searle, and Richard Dobson. 2026. KCL-Cogstack at PsyDefDetect: A hierarchical approach to detecting defense mechanisms in supportive dialogue. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Liudmila Babakova, Christopher Luongo-Vázquez, and Ilia Stepin. 2026. Explainators at PsyDefDetect: Hierarchical prompting and representation-based classification for psychological defenses. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Xin Bai, Guanyi Chen, Tingting He, Chenlian Zhou, and Yu Liu. 2025. Emotional supporters often use multiple strategies in a single turn. *Preprint*, arXiv:2505.15316.

Subhrajyoti Basu. 2026. Neural nexus at PsyDefDetect: Fine-tuning RoBERTa with focal loss and role-tagged dialogue history for defense level detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024. EmpCRL: Controllable empathetic response generation via in-context commonsense reasoning and reinforcement learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5734–5746, Torino, Italia. ELRA and ICCL.

Tong Chen, Zimu Wang, Yiyi Miao, Haoran Luo, Yuanfei Sun, Wei Wang, Zhengyong Jiang, Procheta Sen, and Jionglong Su. 2025. MedFact: A large-scale Chinese dataset for evidence-based medical fact-checking of LLM responses. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32340–32353, Suzhou, China. Association for Computational Linguistics.

Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the CLPsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 177–190, St. Julians, Malta. Association for Computational Linguistics.

Anh Phuong Chu, Luong Duc Tran, Dat Hoang Do, Phuong Tu Mai, Quynh Hoang Le, and Cat Duy

Can. 2026. DAL team at PsyDefDetect: From supervised encoders to hierarchical LLM-RAG for psychological defense detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Brandon Colelough, Davis Bartels, and Dina Demner-Fushman. 2025. Overview of the ClinIQLink 2025 shared task on medical question-answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 378–387, Vienna, Austria. Association for Computational Linguistics.

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, Denver, Colorado. Association for Computational Linguistics.

Mariagrazia Di Giuseppe, Katie Aafjes-van Doorn, Vera Békés, Bernard S Gorman, Karl Stukenberg, and Sherwood Waldron. 2024. Therapists' defense use impacts their patients' defensive functioning: a systematic case study. *Research in Psychotherapy: Psychopathology, Process, and Outcome*, 27(2):797.

Tamjid Hasan Fahim, Syed Asif Johan, and Saad Bin Maksud. 2026. PerceptionLab at PsyDefDetect: Overcoming LLM polarization via rubric-grounded retrieval and supervised clinical reasoning distillation for fine-grained ordinal classification. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Sigmund Freud. 1936. Inhibitions, symptoms and anxiety. *The Psychoanalytic Quarterly*, 5(1):1–28.

Dima Galat and Marian-Andrei Rizoiu. 2026. UTS at PsyDefDetect: Multi-agent councils and absence-based reasoning for defense mechanism classification. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Muhammad Abu Horaira, Mehreen Rahman, and Nahian Chowdhury. 2026. Eraserhead at PsyDefDetect: Prompt design and class rebalancing for psychological defense mechanism detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Yining Hua, Hongbin Na, Zehan Li, Fenglin Liu, Xiao Fang, David Clifton, and John Torous. 2025a. A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, 8(1):230.

Yining Hua, Steve Siddals, Zilin Ma, Isaac Galatzer-Levy, Winna Xia, Christine Hau, Hongbin Na, Matthew Flathers, Jake Linardon, Cyrus Ayubcha, and John Torous. 2025b. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World Psychiatry*, 24(3):383–394.

Bin Huang, Liuyuan Su, Kaixuan Yuan, Guanghui Zhao, Shixin Zhang, and Kunli Zhang. 2026. zzucs at PsyDefDetect: Bridging long-tail imbalance and clinical rubrics for DMRS defense-level detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261, Bangkok, Thailand. Association for Computational Linguistics.

Siam Rahman Karip and Nahid Hossain. 2026. Alien-Annotators at PsyDefDetect: What lies between the lines: Probing lightweight open-source LLMs for psychological defense mechanism detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Veronica Lynn, Alissa Goodman, Kate Niederhoffer, Kate Loveys, Philip Resnik, and H. Andrew Schwartz. 2018. CLPsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46, New Orleans, LA. Association for Computational Linguistics.

Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025. Detecting conversational mental manipulation with intent-aware prompting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9176–9183, Abu Dhabi, UAE. Association for Computational Linguistics.

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 118–127, San Diego, CA, USA. Association for Computational Linguistics.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.

Durjoy C. Paul, Arshitha Basavaraj, Callum Chan, Veronica Perez-Rosas, Diana Inkpen, Francisco Pereira, and Juan Antonio Lossio-Ventura. 2026. TONI-NLP at PsyDefDetect: Defense mechanism detection via LLM-based ensemble methods. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

J Christopher Perry and Melissa Henry. 2004. Studying defense mechanisms in psychotherapy using the defense mechanism rating scales. *Advances in psychology*, 136:165–192.

Oarisa Rebayet, Radiul Walee, Symom Hossain Shohan, Kawsar Ahmed, and Mohammed Moshiul Hoque. 2026. CS_Metro at PsyDefDetect: Detecting psychological defense mechanisms in mental health dialogues with summarization-enhanced transformer ensembles. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Pritha Saha, Shuvodwip Saha, and Anik Mahmud Shanto. 2026. transformer_1376 at PsyDefDetect: A QLoRA-based generative framework for context-aware psychological defense mechanism detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Sarvesh Soni, Soumya Gayen, and Dina Demner-Fushman. 2025. Overview of the ArchEHR-QA 2025 shared task on grounded question answering from electronic health records. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 396–405, Vienna, Austria. Association for Computational Linguistics.

Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large language models and empathy: Systematic review. *J Med Internet Res*, 26:e52597.

Philipp Steigerwald, Eric Rudolph, and Jens Albrecht. 2026. Nürnberg NLP at PsyDefDetect: Multi-axis voter ensembles for psychological defence mechanism classification. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, Seattle, USA. Association for Computational Linguistics.

Talia Tseriotou, Jenny Chim, Ayal Klein, Aya Shamir, Guy Dvir, Iqra Ali, Cian Kennedy, Guneet Singh Kohli, Anthony Hills, Ayah Zirikly, Dana Atzil-Slonim, and Maria Liakata. 2025. Overview of the CLPsych 2025 shared task: Capturing mental health dynamics from social media timelines. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 193–217, Albuquerque, New Mexico. Association for Computational Linguistics.

George E Vaillant. 2012. *Adaptation to life*. Harvard University Press.

Hoang-Thuy-Duong Vu, Quoc-Cuong Pham, and Huy-Hieu Pham. 2026. VISHC at PsyDefDetect: Mitigating data scarcity in psychological defense classification with context-aware synthetic augmentation. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024a. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 470–475, Bangkok, Thailand. Association for Computational Linguistics.

Zimu Wang, Hongbin Na, Rena Gao, Jiayuan Ma, Yining Hua, Ling Chen, and Wei Wang. 2025. From posts to timelines: Modeling mental health dynamics from social media timelines with hybrid LLMs. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 249–255, Albuquerque, New Mexico. Association for Computational Linguistics.

Zimu Wang, Lei Xia, Wei Wang, and Xinya Du. 2024b. Document-level causal relation extraction with knowledge-guided binary question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955, Miami, Florida, USA. Association for Computational Linguistics.

Donald W Winnicott. 2018. Ego distortion in terms of true and false self. In *The person who is me*, pages 7–22. Routledge.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William K. Cheung, and Chenghua Lin. 2025. Overview of the BioLaySumm 2025 shared task on lay summarization of biomedical research articles and radiology reports. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 365–377, Vienna, Austria. Association for Computational Linguistics.

Qingyang Xu, Yaling Shen, Stephanie Fong, Zimu Wang, Yiwen Jiang, Xiangyu Zhao, Jiahe Liu, Zhongxing Xu, Vincent Lee, and Zongyuan Ge. 2026. Do no harm: Exposing hidden vulnerabilities of LLMs via persona-based client simulation attack in psychological counseling. *Preprint*, arXiv:2604.04842.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. ESCoT: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.

Xiangyu Zhao, Yaling Shen, Yiwen Jiang, Zimu Wang, Jiahe Liu, Maxmartwell H Cheng, Guilherme C Oliveira, Robert Desimone, Dominic Dwyer, and Zongyuan Ge. 2025. It hears, it sees too: Multimodal LLM for depression detection by integrating visual understanding into audio language models. *Preprint*, arXiv:2511.19877.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

Ayah Zirikly, Philip Resnik, Özlem 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, Minneapolis, Minnesota. Association for Computational Linguistics.