PersonalityDBench: A Dataset for Personality Disorders - from Modeling to Controlled Generation

Federico Ravenda[1], **Seyed Ali Bahrainian**[2], **Daniele Montagnani**[1,3], **Antonietta Mira**[1,4], **Andrea Raballo**[1,5]
[1]Università della Svizzera italiana, [2]University of Tübingen,
[3]Pavia University, [4]Insubria University, [5]Cantonal Socio-Psychiatric Organization
Correspondence: federico.ravenda@usi.ch

Abstract

Personality disorders (PDs) are a complex class of mental health (MH) conditions characterized by persistent patterns of cognition, behavior, and emotional regulation that deviate from cultural norms. While social media has become a valuable resource for MH research, NLP has largely focused on more prevalent conditions (e.g., depression), leaving PDs underexplored. In this work, we introduce PersonalityDBench, a large-scale, clinically grounded dataset that supports multidimensional study of personality pathology, and standardized, reproducible evaluation of LLM steering toward clinically grounded behavioral targets. The dataset comprises two parts: **(1) PRISMA** (**P**e**R**sonality d**IS**order **MA**nifestations) is a clinically annotated collection of social media content spanning the full spectrum of PDs. It links clinically validated diagnostic criteria and dimensional trait frameworks with computational annotation and analysis methods to support fine-grained, multidimensional study of how PDs manifest in naturalistic, free-form language. Building on PRISMA, **(2) PersonaDSteering** is a benchmark for LLM steering evaluation that operationalizes clinically grounded PD profiles into structured behavioral elicitation tasks, enabling multidimensional steerability assessment, and supporting PD-consistent persona construction for simulated patient generation. This dataset may have application in the study and modeling of PD, adapting language models for clinical feature extraction, and powering personality-specific text generation for adaptive, personalized chat systems[1].

1 Introduction

Personality Disorders are chronic, rigid patterns of thinking, behavior, and emotions that deviate from cultural norms, emerge early, and persistently impair identity and relationships. They differ from Major Psychiatric Syndromes (depression, schizophrenia, bipolar disorder) which are episodic conditions with specific symptoms that fluctuate in severity and often respond to targeted biological treatments.

Once dismissed as clinically insignificant and diagnostically unreliable prior to the 1960s, PDs have undergone a profound transformation in medical perception. Today, these conditions stand recognized as significant MH challenges with far-reaching implications. Modern clinical understanding acknowledges that PDs carry substantial burden - not only in terms of increased morbidity and shortened lifespan - but also through the considerable personal suffering and societal impact they generate (Moran et al., 2016; Tyrer et al., 2010; Winsper et al., 2020). A recent narrative review reported high rates of PDs (4.4–21.5%) in populations across the Western world (Quirk et al., 2016).

PDs are grouped into three clusters: **Cluster A** (Odd/Eccentric), **Cluster B** (Dramatic/Emotional), and **Cluster C** (Anxious/Fearful) (APA, 2013).

Cluster A PDs - *schizotypal* (STPD), *schizoid* (SPD), and *paranoid* (PPD) - manifest as difficulties in forming social connections, characterized by detachment and mistrust; Cluster B PDs - *antisocial* (ASPD), *borderline* (BPD), *histrionic* (HPD), and *narcissistic* (NPD) - involve emotional dysregulation, impulsivity, and unstable relationships; and Cluster C disorders - *avoidant* (AvPD), *dependent* (DPD), and *obsessive-compulsive* (OCPD) - feature pervasive anxiety, fear, and maladaptive control-seeking behaviors. All three PD clusters demonstrate significant comorbidity (i.e., co-occurring MH conditions) with a wide range of MH disorders, including mood, anxiety, and substance use disorders (Lenzenweger et al., 2007). According to the National Institute of Mental Health, 84.5% of individuals with past-year PD also had one or more other MH disorder(s) (NIMH), including suicidal ideation (McClelland et al., 2023).

Given the severity and comorbidity of PDs, their study is crucial, yet increasingly constrained by

[1] Given the sensitive nature of the data, instructions for access are available at the following **Github Repository** https://github.com/Fede-stack/PersonalityDBench

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30239–30259
July 2-7, 2026

limited access to clinical data and privacy regulations. As a result, we turn to widely used Reddit platform, which provide linguistic data reflecting emotional expression, self-disclosure, and support-seeking behaviors (Bucci et al., 2019; Naslund et al., 2016), creating linguistic corpora that capture psychological phenomena in ecological contexts outside rigid clinical settings (Bahrainian et al., 2014). These characteristics have motivated extensive NLP-based MH research, including initiatives such as CLPsych (Tsakalidis et al., 2022) and eRisk (Losada et al., 2017). However, Reddit-based MH datasets have largely concentrated on syndromic conditions (e.g., depression) or transdiagnostic psychobehavioural conditions (e.g., suicidal ideation), leaving much of the broader clinical landscape, particularly more complex and enduring psychopathology, underrepresented (Gkotsis et al., 2017). To fill this gap, we introduce **PersonalityDBench**, the first clinically annotated dataset of social media content spanning the full spectrum of PDs. The *main contributions* of this work are: **(1)** We present a clinically annotated dataset with two complementary levels of analysis: at the post level, diagnostically relevant traits are identified using the Hierarchical Taxonomy of Psychopathology (HiTOP) (Kotov et al., 2017), a dimensional framework for psychopathology (see Appendix E); at the user level, structured diagnostic assessments are conducted using the SCID-5-PD criteria (First et al., 2016), enabling comprehensive characterization of PDs (see Appendix D). **(2)** To the best of our knowledge, PersonalityDBench is the first large-scale social media resource focused on PDs, accompanied by analyses of shared and disorder-specific psycholinguistic patterns. **(3)** We empirically show that HiTOP-informed traits are discriminative and interpretable features, capturing the dimensional structure of personality pathology and supporting the validity of the HiTOP framework in naturalistic language data: HiTOP's hierarchical structure emerges in freely generated social media language and can be leveraged for large-scale digital behavioral analysis. **(4)** Our findings indicate that Reddit content often provides sufficient information for structured diagnostic formulation, revealing that digital footprints can reflect clinically meaningful symptom constellations. **(5)** We introduce PersonaDSteering, a standardized benchmark for evaluating LLM steering on clinically grounded PD profiles via structured behavioral elicitation, and we empirically evaluate steerability across multiple LLMs to characterize robustness, model-specific strengths, and failure modes beyond simplistic single-behavior controls.

2 Related Work

Mental Health Datasets. The MH datasets available in the literature are diverse in nature. Some originate from patient interviews (Delgaram-Nejad et al., 2023) - relatively few, given the sensitive nature of the data - others represent synthetic datasets (Xu et al., 2025), and still others are derived from social media platforms.

In the context of depression, several datasets have been released. In Naseem et al. (2022), the authors present a dataset containing 3'553 Reddit posts relabeled from a binary classification into 4 depression severity levels. Pérez et al. (2023) released DepreSym, a comprehensive dataset of 21'580 sentences from Reddit labeled according to their relevance to 21 depression symptoms specified in the Beck Depression Inventory-II (BDI-II), a standardized clinical questionnaire.

Various initiatives have distinguished themselves through dataset creation within workshops, such as eRisk (Losada et al., 2017), focusing on conditions like Depression, Self-Harm, Anorexia, and Pathological Gambling; and CLPsych, which concentrates on suicide risk (MacAvaney et al., 2021) and mental well-being (Tseriotou et al., 2025).

Additionally, Gupta et al. (2022) introduce PRIMATE, a dataset designed to help training conversational systems to generate appropriate follow-up questions when triaging depression cases. Gratch et al. (2014); Yoon et al. (2022) presented two multimodal datasets for depression, DAIC-WOZ and D-Vlog respectively, consisting of YouTube videos.

Other datasets like (Yates et al., 2017; Cohan et al., 2018; Rani et al., 2024; Raihan et al., 2024; Low et al., 2020) focus on collecting scattered posts from Reddit across various MH subcommunities beyond depression. These are complemented by specific datasets related to hate speech (Piot et al., 2024), Suicide (Zhang et al., 2024), Panic and Anxiety (Mitrović et al., 2024), Schizophrenia (Bae et al., 2021), and Bipolar (Jagfeld et al., 2021).

The cited resources, while relevant, lack clinical annotations and mainly distinguish posts across subcommunities at the single-post level, overlooking users as *holistic entities* and their longitudinal patterns of expression.

Depression-focused social media modeling grew

by 150% between 2021 and 2024, likely supported by the availability of suitable datasets (Bucur et al., 2025); in contrast, comparable clinically grounded resources for other conditions, including PDs, remain scarce. Our dataset addresses this gap by providing a high-quality, clinically annotated dataset spanning the full PD spectrum, which is still underrepresented in computational research.

LLMs Steerability & Persona Generation in MH. LLMs are increasingly deployed in settings where *controllable generation* is not optional: models must reliably adapt their outputs to a desired style (Bahrainian et al., 2024), or behavioral profile while preserving coherence and task performance. This has catalyzed rapid progress on *steering* methods, including prompting-based control (Lester et al., 2021; Wei et al., 2022; Zhao et al., 2024; Chen et al., 2025), supervised tuning (Ouyang et al., 2022; Rafailov et al., 2023), and more recently *steering vectors* (Rimsky et al., 2024; Braun et al.; Im and Li, 2025) that modulate internal representations to induce targeted behaviors in an efficient way. Steering vectors are attractive because they are lightweight, composable, and often transferable across prompts and domains, making them a practical mechanism for generation control at inference time. Despite these efforts, steering-vector research still lacks standardized evaluation: studies rely on narrow, ad hoc datasets and task-specific criteria (e.g., refusal or sycophancy), limiting validation and cross-method comparability. This makes it hard to attribute gains to the steering method rather than data or prompt sensitivity (Tan et al., 2024), hindering reliable comparison, failure-mode analysis, and reproducibility across models and setups. In parallel, patient simulators have long been used for medical education and clinician training (Rao and Artino, 2025), and recent work has explored LLMs for depression screening (Ravenda et al., 2025b; Bucur et al., 2021). Yet LLM-based patient simulation in MH has largely been studied through prompt steering (Wang et al., 2024, 2025), and steering-vector methods have not been evaluated on clinically grounded MH targets. For this reason, a dedicated evaluation resource is needed to measure steerability in MH and to verify that steered simulated patients remain aligned with the intended MH profile, rather than drifting across prompts or contexts. Accordingly, to the best of our knowledge, this work is the first to link steering evaluation to a MH benchmark grounded in validated diagnostic constructs, enabling controllable generation toward psychologically meaningful behavioral targets rather than purely stylistic or single-attribute controls.

3 Methods

Research Questions. The following Research Questions guided both the dataset construction and subsequent experiments:

(RQ1) Can the presence of specific diagnostic criteria for PDs be reliably identified from the longitudinal Reddit posting history of diagnosed users, and does this textual evidence suffice to meet the diagnostic thresholds defined by SCID-5-PD?

(RQ2) Can LLMs reliably identify and annotate specific diagnostic criteria, treated as dimensional symptom indicators, in user-generated texts, rather than being limited to a final categorical diagnosis?

(RQ3) Do empirically derived HiTOP traits, when observed in naturalistic language data, correspond to diagnostic profiles of categorical PDs, and can PDs be understood as quantitative configurations of traits across spectra rather than discrete categories?

(RQ4) Can steering vectors reliably control LLM generation toward clinically grounded behavioral patterns of PDs, thereby enabling standardized evaluation of steerability and the faithful generation of coherent, reusable PD-consistent personas?

3.1 Data Collection & Statistics

Reddit posts and comments were collected from users who participated in discussions on the following subreddits: **r/Schizotypal, r/ParanoidPersonality, r/Schizoid, r/BPD, r/BorderlinePDisorder, r/AvoidantAttachment, r/AvPD, r/narcissism, r/NPD, r/OCPD, r/DPD, r/hpd, r/aspd,** and **r/mentalillness**. After collection, posts were grouped by user and split by whether the author reported a formal diagnosis from a healthcare professional. This determination was made using regular expressions to identify diagnostic disclosures, with all positive matches manually verified to exclude self-diagnosed cases or cases where the diagnostic process was unclear. In total, **155** users were selected, a manageable cohort for detailed annotation.

For each user, all historical *posts* and *comments* were scraped based on the username. Each post was then classified as either related or unrelated to MH content. To perform this classification, we independently prompted three SOTA LLMs (listed in Appendix G.1) using an instruction that asked

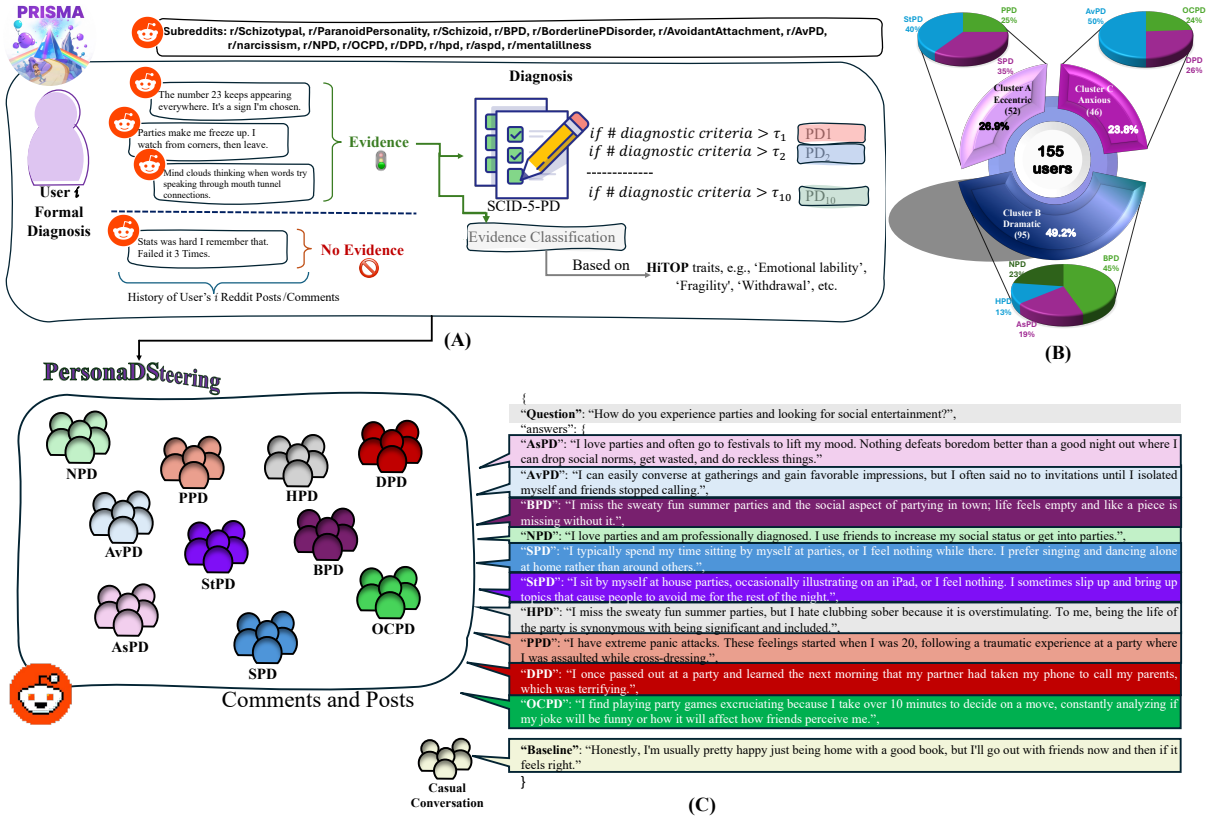

Figure 1: **(A)** Overview of the PRISMA dataset construction workflow, illustrating the post-level evidence classification, HiTOP trait mapping, and user-level SCID-5-PD diagnostic aggregation process. **(B)** Distribution of users (155 in total) across PD diagnoses and comorbidities within the PRISMA dataset. **(C)** PersonaDSteering overview: each question is paired with one reference answer per PD plus a **r/CasualConversation** baseline (example shown).

whether the post discussed MH topics in a broad sense - such as symptoms, treatments, therapy, medications, or any aspect of psychological well-being - regardless of the specific PD reported by the user. Subsequently, all content not directly related to the user was filtered out. Similarly to the annotation strategy described in MentalHelp dataset (Raihan et al., 2024) - where Flan T5, DisorBERT, and MentalBERT were combined to assign labels to 14M Reddit posts - we use an hard-voting approach for evidence classification: a post was counted as evidence if at least two LLMs concurred. The details of the prompting and models used are described in Section G.1 of the Appendix. This procedure allowed us to reduce the number of posts per user to consider during the annotation step. Table 1 summarizes the dataset statistics. Figure 1**(A)** illustrates the main steps of the PRISMA creation process, while Figure 1**(B)** shows the number of users per PD. Note that the total number of users and users per PD differ due to comorbidities across disorders. Once relevant evidence was identified, these posts were used to assess the diagnostic criteria for various PDs as operationalized in the SCID-5-PD structured interview. Each criterion is rated on a 0–2 scale (no evidence, subthreshold, threshold), then dichotomized to simplify annotation: a score of *0* indicates no or subthreshold evidence, whereas a score of *1* denotes clear and explicit evidence of the criterion being present in the user's Reddit posts. Importantly, we do not administer the SCID-5-PD interview on Reddit; rather, we use SCID-5-PD as a criterion-level evidence-extraction framework, leveraging its standardized operationalization of DSM-based criteria (and its allowance to code criteria from narrative materials when interviews are unavailable, see Appendix D). Annotation guidelines and validation protocols were designed by a licensed psychiatrist with decades of clinical experience. Annotation was then carried out by two MH researchers through collaborative case-by-case discussion, a deliberate methodological choice established in clinical psychology, where shared discussion during annotation produces more reliable ground truths than computing agreement post hoc (Oortwijn et al., 2021; Schmer-Galunder

et al., 2024; Zhu and Rzeszotarski, 2024). In addition, the identified evidences were mapped onto maladaptive traits defined within the HiTOP framework, allowing for a dimensional characterization of psychopathological features. This mapping enables a finer-grained, trait-level understanding of PD and supports transdiagnostic research, linking linguistic evidence to empirically derived dimensions of personality dysfunction. To achieve this, approximately ∼12'000 evidence posts were manually annotated for HiTOP traits by the same two annotators, following the same collaborative discussion protocol, allowing for multiple traits per post.

PersonaDSteering: Benchmark Construction. In addition to PRISMA, we collect a non-conditioned baseline from **r/CasualConversation**. PersonaDSteering is structured as a controlled QA benchmark in which each question is paired with one reference answer per PD plus a baseline answer, enabling direct comparisons between PD-conditioned and non-conditioned responses.

To evaluate steerability, we steer an LLM toward a target PD profile and assess whether, across this fixed question set, its answers consistently match the intended PD-specific behavioral pattern relative to the baseline condition.

Unlike existing steering evaluations, which are typically limited to isolated and coarse behaviors such as refusal or sycophancy, without standardized evaluation, our protocol enables a standardized, fine-grained, and multi-faceted assessment based on structured, clinically inspired elicitation. For each PD, we construct a bank of 100 closed-form questions inspired by SCID-5-PD interviews, designed to be short, unambiguous, answerable in constrained formats, and aligned as much as possible with a single behavioral dimension. A detailed explanation of the test question structure, together with the rationale behind the question design, is provided in the Appendix.

As illustrated in Figure 1 **(C)**, PersonaDSteering is formulated as a controlled QA benchmark in which each question is paired with one reference answer per PD and an additional *baseline* answer drawn from **r/CasualConversation**, providing a non-conditioned comparison target. Overall, it comprises 1'388 question-answering items; each question is paired with 11 aligned reference answers (10 PD-specific + 1 baseline), yielding a total of **15'268** reference answers.

To construct these aligned question-answer tuples, we first induce latent topics (Bahrainian et al., 2018) and generate candidate natural language questions. For each question, we retrieve PD-specific users' posts and comments useful to answer the query (see Section C in Appendix for details). The retrieved posts and comments are then used to generate a set of aligned PD-conditioned responses, one per PD, together with a baseline response from Casual Conversations subreddit; responses are subsequently normalized for consistency using *Gemini-3-pro*, yielding comparable answers for the same question across conditions. When no suitable response can be generated for a given condition, entries are manually completed by an annotator via targeted Reddit search aligned with the corresponding topic.

Finally, for each PD, we construct a bank of 100 closed-form test questions inspired by SCID-5-PD questions (see Figure 3 for an example). We use these test questions as the fixed assessment set for steerability: after steering an LLM, we evaluate whether, across the full question bank, the model's answers consistently match the intended PD-specific behavioral profile relative to the baseline condition.

Statistics	mean	std	max	min
# of posts	567	664	2242	3
# of sentences	730	878	7454	12
# of words	19'673	21'833	130'830	441
# of Users: 155				

Table 1: PRISMA's dataset statistics. Values are reported as averages, standard deviation, minimum and maximum per user.

4 Dataset Analysis

The following experiments follow the RQs' order.

4.1 Are digital footprints sufficient for structured PD diagnosis?

Structured diagnosis from digital footprints. For **(RQ1)**, we design the annotation task following SCID-5-PD guidelines, focusing on explicit identification of individual diagnostic criteria in users' Reddit posts. In contrast to resources such as eRisk (Losada et al., 2019, 2020; Parapar et al., 2021), which link posting histories to self-reported questionnaire scores (e.g., BDI-II), our approach grounds diagnostic evidence in the *textual content* produced on social media. While eRisk is an important step toward mapping posts to depressive-

Model	Cohen's κ	ACC	F1	P	R	MAE
`llama-3.3-70b`	0.237	0.670	0.770	0.663	0.919	2.518
`llama-3.3-70b` + *aRAG*	0.310	0.698	0.785	0.686	0.918	1.948
`deepseek-v3.1`	0.313	0.686	0.760	0.702	0.827	2.223
`deepseek-v3.1` + *aRAG*	0.377	0.706	0.762	0.741	0.785	1.616
`gemini-2.5-flash-lite`	0.332	0.671	0.710	0.754	0.670	2.057
`gemini-2.5-flash-lite` + *aRAG*	0.369	0.677	0.691	0.813	0.601	1.756
`mistral-medium-3.1`	0.243	0.680	0.783	0.660	0.961	2.445
`mistral-medium-3.1` + *aRAG*	0.335	0.693	0.763	0.713	0.820	1.782

Table 2: Performance comparison of different LLMs and classification strategies on diagnostic predictions.

symptom severity, it is limited by the static nature of questionnaire scores, which capture only a snapshot, and by the fact that symptoms may be inferred indirectly rather than explicitly self-reported (Ravenda et al., 2025c). In this work, we address a different question: whether Reddit posts contain sufficient textual evidence for specific diagnostic criteria, grounding annotations in observable language rather than questionnaire scores. We do not treat *present* criteria as an exhaustive clinical profile; instead, we test whether evidence meets the SCID-5-PD diagnostic threshold τ, i.e., whether at least τ criteria show strong evidence in a user's posts, where τ is the SCID-5-PD minimum required for diagnosis (see Table 3 in Appendix). Our results indicate that in almost $\sim$70% of cases, users who self-report a professional diagnosis provide sufficient textual evidence to meet the corresponding SCID-5-PD threshold. This proportion varies across disorders: it exceeds 75% for Schizoid, Schizotypal, Borderline, Narcissistic, and Antisocial PD; reaches 96% for Avoidant PD; and falls below 30% for Paranoid and Dependent PD. For Histrionic PD, only one out of twelve users shows sufficient evidence. These findings suggest that, for many PDs, social media posts contain enough cues to support a preliminary diagnosis, while also highlighting few PDs that are more difficult to capture through text alone.

Automatic assessment of diagnostic criteria with LLMs. For **(RQ2)**, we evaluate whether state-of-the-art LLMs can automatically identify individual diagnostic criteria from users' Reddit histories. While prior work shows that LLMs are effective in clinical reasoning tasks, they tend to overestimate symptom presence when applied to MH data (Ravenda et al., 2025a). To mitigate this issue, we compare two modeling strategies. The first approach conditions the model on the full posting history of each user to predict the presence of each diagnostic criterion. The second employs an adaptive Retrieval-Augmented Generation (aRAG) framework (Ravenda et al., 2025b), which retrieves only the most diagnostically relevant posts for each criterion, reducing noise and improving focus (see Appendix F for details). Model performance is evaluated using Cohen's κ to measure agreement with manual annotations, alongside Accuracy, macro-averaged Precision, Recall, and F1-score to account for class imbalance. ($\sim$60% of instances labeled as strong evidence). We additionally report the Mean Absolute Error (MAE), capturing discrepancies between LLM-predicted and manually annotated counts of strong-evidence criteria. Results in Table 2 show that aRAG-enhanced models often outperform their non-retrieval counterparts, particularly in terms of Cohen's κ, Accuracy, F1-score, and MAE. Although agreement remains in the fair range (typically $\kappa \in [0.21, 0.40]$), aRAG substantially reduces overestimation: across models, 57% of posts are labeled as strong evidence with aRAG, compared to 68% without retrieval, closely matching the percentage observed in manual annotations. The full prompt used for criterion classification is reported in Appendix G.2. Additional analyses are reported in the Appendix A, including ablation experiments in which LLMs are evaluated on clinically inspired vignettes generated as concise summaries of users' posts and comments.

4.2 HiTOP is cool! Do HiTOP traits relate to PDs diagnosis in Social Media data?

For **(RQ3)**, we focus on all HiTOP traits, i.e., stable, dimensional psychopathological characteristics capturing maladaptive personality functioning at the lowest level of the HiTOP hierarchy, below syndromes, subfactors, and spectra. HiTOP traits are empirically derived from large-scale factor analyses, making them data-driven rather than arbitrary. Focusing on traits, its most fine-grained constructs, allows us to test the framework using indicators directly observable in users' naturalistic language, rather than broader syndromes or spectra.

For each PD, we compute the relative frequency of each trait across diagnosed individuals and apply min-max scaling to enable comparability across PDs: the most frequent trait for a PD is mapped to 1, and the remaining traits are scaled to $[0, 1]$. In this way, we do not show absolute prevalence, but the *relative profile* of traits within each PD.

Figure 2 shows the distributions of HiTOP traits for each PD, with traits color-coded according to their corresponding spectrum: **Internalizing/Negative Affectivity**, **Thought Disorder/Psychoticism**, **Detachment**, **Disinhibited**

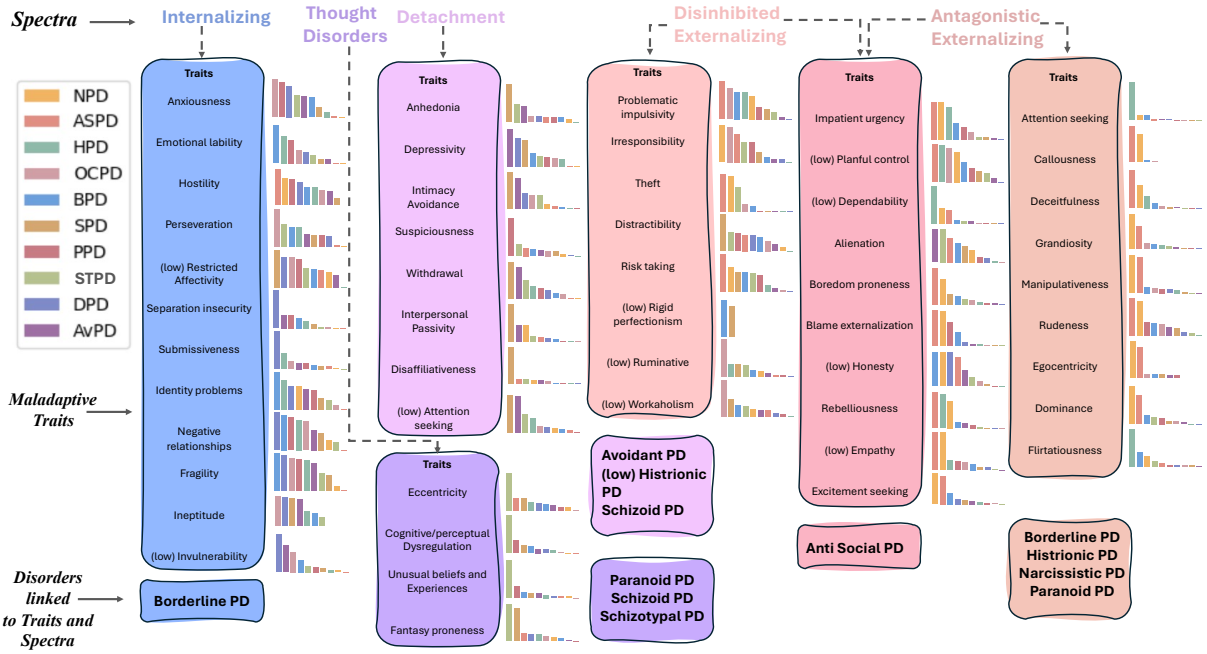

Figure 2: Visualization of HiTOP traits associated with each PD, organized across the major psychopathological spectra. Spectra, trait clusters, and their corresponding PDs are color-matched to highlight their hierarchical relationships. Next to each trait, bar plots show the degree to which each PD express the specific trait.

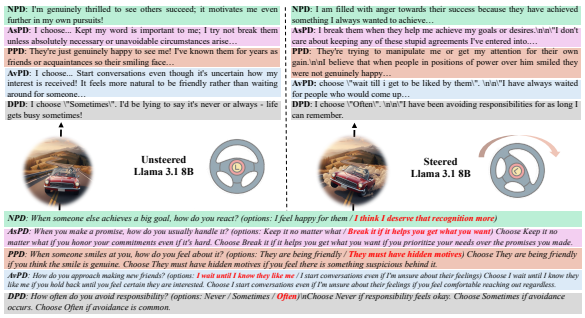

Figure 3: Examples of five test-set questions, each targeting a different PD (NPD, AsPD, PPD, AvPD, DPD), with ground-truth labels highlighted in red, and the corresponding responses generated by the unsteered and steered settings (conditioned on the target PD).

Externalizing, and Antagonistic Externalizing. Based on the HiTOP framework, we expected distinct trait-PD associations: Internalizing in BPD, reflecting anxiety, emotional lability, identity problems, and fragility; Detachment in SPD and AvPD, marked by anhedonia, intimacy avoidance, and withdrawal; Thought Disorder/Psychoticism in STPD, PPD, and SPD, with eccentricity and unusual beliefs; Disinhibited Externalizing in ASPD and partly in BPD, with impulsivity and irresponsibility; Antagonistic Externalizing in NPD, HPD, ASPD, and BPD, with callousness, manipulativeness, and grandiosity. Our empirical findings confirmed these expectations: Cluster B disorders (BPD, NPD, HPD) peaked on Antagonism and Disinhibition; Cluster C (AvPD, DPD, and OCPD) showed stronger Internalizing and Detachment; Cluster A (STPD, PPD, SPD) aligned with Psychoticism and Detachment; and ASPD clearly combined Disinhibition and Antagonism (additional analyses stratified by cluster and individual PD are reported in Appendix Section A.1). Overall, these patterns align with HiTOP's theoretical expectations and support a dimensional view of PDs as distinct quantitative configurations of traits across spectra rather than discrete categories. We also find that PDs not explicitly represented in the original HiTOP framework (DPD and OCPD) can be naturally accommodated within this trait–spectrum structure, supporting the HiTOP perspective that each condition can be understood as a mixture of traits distributed across multiple spectra.

4.3 Assessing LLM Steerability for Personality-Disorder Expression

To address **(RQ4)**, we use Contrastive Activation Addition (CAA) steering (Rimsky et al., 2024) on four instruction-tuned LLMs: `Llama-3.1-8B-Instruct`, `Llama-3.2-3B-Instruct`, `Mistral-7B-Instruct`, and `Qwen3-4B-Instruct`. We evaluate each steered model against its unsteered version on the PersonaDSteering test set (see Figure

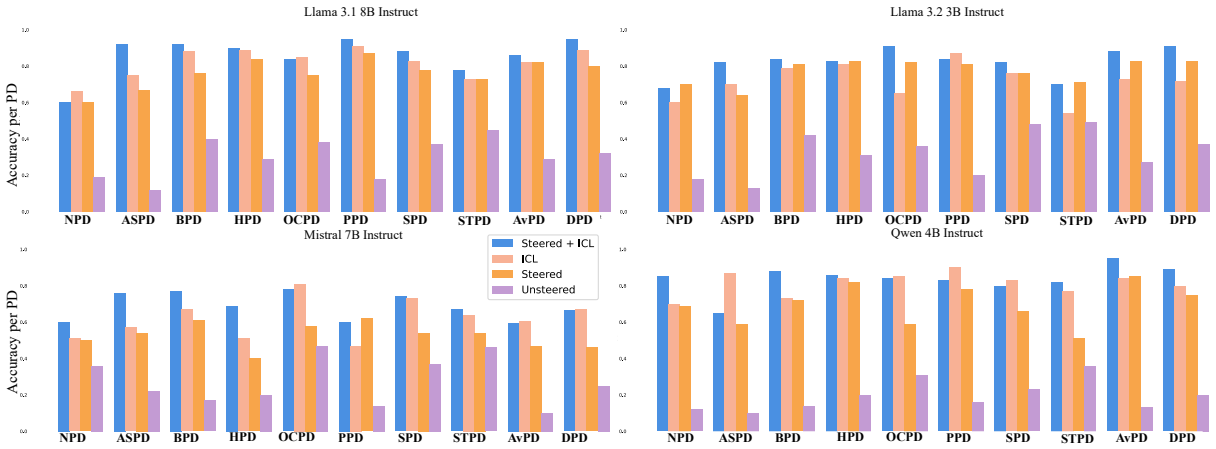

Figure 4: Steerability results on PersonaDSteering for four instruction-tuned LLMs. Bars report performance by PD under four conditions: Unsteered, Steering only, ICL only, and ICL+Steering

3 for examples), using the same questions across conditions. CAA defines a steering direction by contrasting two sets of PersonaDSteering answers to the same questions. $\mathcal{D}^+$ is the collection of reference answers written to reflect a target PD across many questions, while $\mathcal{D}^-$ is the corresponding baseline set of answers drawn from Casual Conversation (i.e., non-conditioned, "neutral" answers). For a chosen layer ℓ, we compute the steering vector as the average hidden representation elicited by the PD answers and subtract the average hidden representation elicited by the baseline answers:

$$s_\ell = \frac{1}{|\mathcal{D}^+|} \sum_{x \in \mathcal{D}^+} h_\ell(p) - \frac{1}{|\mathcal{D}^-|} \sum_{x \in \mathcal{D}^-} h_\ell(p)$$

Intuitively, s_ℓ captures "what changes in the model's internal activations when moving from a neutral (Casual Conversation) response style to a PD-consistent response style", aggregated across many questions. At inference time, given an input prompt p, we apply activation addition at a set of intermediate transformer layers:

$$h'_\ell(p) = h_\ell(p) + \lambda\, s_\ell \qquad \ell \in \mathcal{L},$$

where $\mathcal{L}$ denotes the set of steered layers (e.g., $\mathcal{L} = \{10, \ldots, 13\}$), which may consist of a single layer or multiple consecutive layers, where λ is the steering strength and $h'_\ell(x)$ is used for generation.

To select the layers to steer, *although* our goal was not to optimize steering performance, we sampled 20 held-out questions from the training set, not used for steering vector computation, and qualitatively assessed whether the steered LLM's responses were consistent with the intended PD-specific behavioral profile. Based on this assessment, we tested combinations of four consecutive layers: early ($\{3, \ldots, 6\}$), intermediate ($\{10, \ldots, 13\}$ or $\{8, \ldots, 11\}$), and final ($\{28, \ldots, 31\}$). We selected intermediate layers $\{10, \ldots, 13\}$ for Llama-3.1-8B-Instruct, Mistral-7B-Instruct, and Qwen3-4B, and $\{8, \ldots, 11\}$ for Llama-3.2-3B-Instruct.

In all experiments, we use $\lambda = 0.5$. *Importantly*, our goal is not to optimize steering performance but to establish a standardized evaluation signal; steering quality can be improved with model-specific tuning of the layer set and λ (and potentially per-PD calibration). This is out of scope for this work and left for future research.

Across all four models, steering at the designated mid layers yields the strongest and most consistent gains, while ablating to early layers (3–6) or to the final four layers markedly reduces the effect. This supports the interpretation that the control signal is most effectively injected at intermediate layers (Rimsky et al., 2024). For example, Qwen 4B shows severe hallucinations and poor coherence when steered in early layers (3–6) or in the last four layers, so we do not report those ablation results. This failure mode often manifests as token-level corruption and nonsensical continuations (e.g., the model starts emitting garbled strings instead of selecting one of the provided options), as in the following example:

> *"When someone else achieves a big goal, how do you react? . . . "* → `own#define, Position to beFsublicniizeile colon statelassasted sacrifice-spiFormat to!.`

Such instability is consistent with prior obser-

vations that aggressive or poorly placed activation interventions can push hidden states out-of-distribution and lead to non-plausible text (Braun et al., 2026).

To disentangle implicit control from explicit activation steering, we consider four conditions: **(1) Unsteered** ($\lambda = 0$), where no steering is applied; **(2) ICL**, where we additionally condition the model via in-context demonstrations specifying the target PD; **(3) steering only**, where we apply CAA steering without in-context examples; and **(4) ICL+steering**, combining in-context demonstrations with CAA steering. Figure 4 indicates that CAA can outperform ICL in some cases (notably for `Llama-3.2-3B-Instruct`), and that combining CAA with ICL often yields additional gains over either component alone. Figure 4 also highlights PD-specific differences in steerability: OCPD, PPD, and DPD are generally easier to elicit, whereas NPD remains more challenging. Moreover, among the evaluated models, `Mistral-7B-Instruct` appears the hardest to steer, with the clearest difficulties for ASPD and NPD (a pattern shared by the other models as well), plausibly reflecting the impact of safety alignment that suppresses harmful or socially unsafe generations (Zhang et al., 2025). Finally, in the steering vectors setting, the strongest overall performance is observed for the two Llama variants, in particular with `Llama-3.2-3B-Instruct` yielding the most consistent gains across disorders. In addition to the QA benchmark, Section 7 in the Appendix includes an ablation study testing whether symptoms expressed by the steered LLM transfer to an open-ended conversational setting, assessing PD-consistent behavior beyond the controlled QA format.

5 Conclusion

In this paper, we introduce PersonalityDBench, the first clinician-annotated dataset that combines PD modeling/detection with PD-specific text generation, enabling both **(1)** clinical understanding/modeling and **(2)** controllable persona generation. For **(RQ1–RQ2)**, we show that SCID-5-PD criteria can be identified in Reddit posts and that a RAG-based setup improves performance and reduces LLM overestimation toward clinically aligned predictions. For **(RQ3)**, we find that HiTOP-informed traits are discriminative, interpretable signals in natural language, consistent with the mapping between traits, spectra, and diagnostic categories. For **(RQ4)**, we present PersonaD-Steering and benchmark activation steering across LLMs, showing that steering (often further helped by in-context conditioning) can shift generations toward disorder-consistent patterns, with PD-specific controllability differences that motivate more robust clinically grounded control and evaluation.

Beyond the methodology, this work offers an epistemological contribution, demonstrating that digital traces can meaningfully reflect psychopathological structure and serve as a principled bridge between NLP and clinical psychology. More broadly, PersonalityDBench provides a foundation for adapting language models to the extraction of clinically grounded psychological traits and features, supporting the development of computationally informed tools for MH research.

6 Limitations

This work has some limitations that should be acknowledged. First, the dataset relies on reported diagnoses by a MH professional from Reddit users, which cannot be independently verified and may be subject to reporting bias. Moreover, language expressed in online contexts may not fully reflect the complexity of real-world clinical symptomatology, as it is influenced by factors such as anonymity, community norms, and self-presentation bias.

Additionally, while frameworks such as SCID-5-PD and HiTOP ensure a clinically grounded interpretation, the identification of linguistic and emotional markers remains probabilistic and cannot replace professional diagnosis or clinical evaluation. The choice of SCID-5-PD was motivated by its rigor and comprehensive coverage of all PDs, which makes it uniquely suitable for structured diagnostic assessment. Nonetheless, other standardized psychological instruments - such as the Schizotypal Personality Questionnaire (SPQ) for schizotypy or the McLean Screening Instrument for BPD (MSI-BPD) for borderline personality disorder - can complement this framework by providing disorder-specific screening and self-report perspectives.

Future work should expand to cross-platform validation, inclusion of clinician–patient dialogues, and integration of multimodal signals to strengthen ecological validity and diagnostic robustness.

7 Ethical Considerations

Use of publicly available data and privacy protection. All data are drawn from publicly accessible Reddit content. Reddit users post content under pseudonyms, and our data collection excludes personally identifiable information. User identifiers (e.g., usernames) are removed or anonymized, and all illustrative excerpts presented in the paper are paraphrased to mitigate re-identification risks. We do not attempt to deanonymize users or link posts across communities.

Risk of bias amplification and harmful stereotyping. Some personality disorders featured in PersonalityDBench (e.g., ASPD, NPD, BPD) are subject to societal stigma and negative stereotyping. Models trained or steered using this dataset may inadvertently learn, reinforce, or amplify biases present in the data, including over-pathologizing certain behaviors, conflating diagnostic labels with moral judgments, or producing outputs that perpetuate harmful stereotypes. We urge caution in downstream use and recommend further auditing, fairness evaluation, and bias mitigation.

Intended use and misuse prevention. PersonalityDBench and its associated models are released strictly for research and educational purposes. The dataset and models **must not** be used for:

- Systems that impersonate real individuals, simulate communities without transparency, or manipulate public opinion;
- De-anonymization attempts or linking of users across platforms or datasets;
- Commercial applications in hiring, law enforcement, insurance, or other high-stakes decision-making contexts where misuse could result in harm or discrimination.

Any downstream application must adhere to stringent ethical guidelines, including respect for the dignity and well-being of individuals whose posts contribute to the dataset, transparency about model limitations, and careful auditing for potential harms.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback which helped to improve the manuscript. This research has partly been founded by the SNSF (Swiss National Science Foundation) grant 200557.

References

APA. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association.

Yi Ji Bae, Midan Shim, and Won Hee Lee. 2021. Schizophrenia detection using machine learning approach from social media content. *Sensors*, 21(17):5924.

Seyed Ali Bahrainian, Jonathan Dou, and Carsten Eickhoff. 2024. Text simplification via adaptive teaching. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6574–6584, Bangkok, Thailand. Association for Computational Linguistics.

Seyed Ali Bahrainian, Marcus Liwicki, and Andreas Dengel. 2014. Fuzzy subjective sentiment phrases: A context sensitive and self-maintaining sentiment lexicon. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 361–368.

Seyed Ali Bahrainian, Ida Mele, and Fabio Crestani. 2018. Predicting topics in scholarly papers. In *Advances in Information Retrieval*, pages 16–28, Cham. Springer International Publishing.

Joschka Braun, Carsten Eickhoff, and Seyed Ali Bahrainian. Beyond multiple choice: Evaluating steering vectors for adaptive free-form summarization. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Joschka Braun, Carsten Eickhoff, and Seyed Ali Bahrainian. 2026. Beyond multiple choice: Evaluating steering vectors for summarization. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 3849–3884, Rabat, Morocco. Association for Computational Linguistics.

Sandra Bucci, Matthias Schwannauer, and Natalie Berry. 2019. The digital revolution and its impact on mental health care. *Psychology and Psychotherapy: Theory, Research and Practice*, 92(2):277–297.

Ana-Maria Bucur, Adrian Cosma, and Liviu P Dinu. 2021. Early risk detection of pathological gambling, self-harm and depression using bert. *arXiv preprint arXiv:2106.16175*.

Ana-Maria Bucur, Andreea-Codrina Moldovan, Krutika Parvatikar, Marcos Zampieri, Ashiqur R KhudaBukhsh, and Liviu P Dinu. 2025. Datasets for depression modeling in social media: An overview. *arXiv preprint arXiv:2503.21513*.

Kai Chen, Zihao He, Taiwei Shi, and Kristina Lerman. 2025. Steer-bench: A benchmark for evaluating the steerability of large language models. *arXiv preprint arXiv:2505.20645*.

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. *arXiv preprint arXiv:1806.05258*.

Alessandro De Grandi, Federico Ravenda, Andrea Raballo, and Fabio Crestani. 2024. The emotional spectrum of llms: Leveraging empathy and emotion-based markers for mental health support. *arXiv preprint arXiv:2412.20068*.

Oliver Delgaram-Nejad, Dawn Archer, Gerasimos Chatzidamianos, Louise Robinson, and Alex Bartha. 2023. The dais-c: A small, specialised, spoken, schizophrenia corpus. *Applied Corpus Linguistics*, 3(3):100069.

Antonio Di Noia, Iuri Macocco, Aldo Glielmo, Alessandro Laio, and Antonietta Mira. 2024. Beyond the noise: intrinsic dimension estimation with optimal neighbourhood identification. *arXiv preprint arXiv:2405.15132*.

Zohar Elyoseph, Inbar Levkovich, and 1 others. 2024. Comparing the perspectives of generative ai, mental health experts, and the general public on schizophrenia recovery: case vignette study. *JMIR Mental Health*, 11(1):e53043.

MB First, JB Williams, LS Benjamin, and RL Spitzer. 2016. Structured clinical interview for dsm-5 personality disorders: Scid-5-pd. *Testkatalog*, page 30.

George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, 7(1):1–11.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, and 1 others. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, volume 14, pages 3123–3128. Reykjavik.

Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. 2022. Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts. *arXiv preprint arXiv:2205.13884*.

Shawn Im and Yixuan Li. 2025. A unified understanding and evaluation of steering methods. *arXiv preprint arXiv:2502.02716*.

Glorianna Jagfeld, Fiona Lobban, Paul Rayson, and Steven H Jones. 2021. Understanding who uses reddit: Profiling individuals with a self-reported bipolar disorder diagnosis. *arXiv preprint arXiv:2104.11612*.

Roman Kotov, Robert F Krueger, David Watson, Thomas M Achenbach, Robert R Althoff, R Michael Bagby, Timothy A Brown, William T Carpenter, Avshalom Caspi, Lee Anna Clark, and 1 others. 2017. The hierarchical taxonomy of psychopathology (hitop): A dimensional alternative to traditional nosologies. *Journal of abnormal psychology*, 126(4):454.

Mark F Lenzenweger, Michael C Lane, Armand W Loranger, and Ronald C Kessler. 2007. Dsm-iv personality disorders in the national comorbidity survey replication. *Biological psychiatry*, 62(6):553–564.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Inbar Levkovich, Eyal Rabin, Michal Brann, and Zohar Elyoseph. 2024. Large language models outperform general practitioners in identifying complex cases of childhood anxiety. *Digital Health*, 10:20552076241294182.

David E Losada, Fabio Crestani, and Javier Parapar. 2017. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, pages 346–360. Springer.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357. Springer.

David E Losada, Fabio Crestani, and Javier Parapar. 2020. erisk 2020: Self-harm and depression challenges. In *European conference on information retrieval*, pages 557–563. Springer.

Daniel M Low, Laurie Rumker, John Torous, Guillermo Cecchi, Satrajit S Ghosh, and Tanya Talkar. 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, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. 2021. Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 70–80, Online. Association for Computational Linguistics.

Heather McClelland, Seonaid Cleare, and Rory C O'Connor. 2023. Suicide risk in personality disorders: A systematic review. *Current psychiatry reports*, 25(9):405–417.

Sandra Mitrović, Oscar William Lithgow-Serrano, and Carlo Schillaci. 2024. Comparing panic and anxiety on a dataset collected from social media. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 153–165.

Paul Moran, Helena Romaniuk, Carolyn Coffey, Andrew Chanen, Louisa Degenhardt, Rohan Borschmann, and George C Patton. 2016. The influence of personality disorder on the future mental health and social adjustment of young adults: a population-based, longitudinal cohort study. *The Lancet Psychiatry*, 3(7):636–645.

Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. 2022. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM web conference 2022*, pages 2563–2572.

John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences*, 25(2):113–122.

NIMH. Personality disorders. https://www.nimh.nih.gov/health/statistics/personality-disorders. Accessed: 2025-10-06.

Yvette Oortwijn, Thijs Ossenkoppele, and Arianna Betti. 2021. Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks. In *Proceedings of the workshop on human evaluation of NLP systems (HumEval)*, pages 131–141.

World Health Organization and 1 others. 1992. Icd-11. *(No Title)*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2021. erisk 2021: pathological gambling, self-harm and depression challenges. In *European Conference on Information Retrieval*, pages 650–656. Springer.

Anxo Pérez, Marcos Fernández-Pichel, Javier Parapar, and David E Losada. 2023. Depresym: A depression symptom annotated corpus and the role of llms as assessors of psychological markers. *arXiv preprint arXiv:2308.10758*.

Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2024. Metahate: A dataset for unifying efforts on hate speech detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 2025–2039.

Shae E Quirk, Michael Berk, Andrew M Chanen, Heli Koivumaa-Honkanen, Sharon L Brennan-Olsen, Julie A Pasco, and Lana J Williams. 2016. Population prevalence of personality disorder and associations with physical health comorbidities and health care service utilization: A review. *Personality Disorders: Theory, Research, and Treatment*, 7(2):136.

Andrea Raballo, Federico Ravenda, and Antonietta Mira. Diagnosing schizophrenia spectrum disorders: Large language models (llms) vs. leading international psychiatrists (lips). *Psychiatry and Clinical Neurosciences*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Shafkat Farabi, Ana-Maria Bucur, Tharindu Ranasinghe, and Marcos Zampieri. 2024. Mentalhelp: A multi-task dataset for mental health in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11196–11203.

Saima Rani, Khandakar Ahmed, and Sudha Subramani. 2024. From posts to knowledge: Annotating a pandemic-era reddit dataset to navigate mental health narratives. *Applied Sciences*, 14(4):1547.

Arya S Rao and Anthony R Artino. 2025. Ai-driven osce preparation in medical education: Promise, pitfalls, and practical implications.

Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, and Antonietta Mira. 2025a. Navigating through the hidden embedding space: steering llms to improve mental health assessment. *arXiv preprint arXiv:2510.16373*.

Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, Antonietta Mira, and Noriko Kando. 2025b. Are LLMs effective psychological assessors? leveraging adaptive RAG for interpretable mental health screening through psychometric practice. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8975–8991, Vienna, Austria. Association for Computational Linguistics.

Federico Ravenda, Antonio Preti, Michele Poletti, , Antonietta Mira, and Andrea Raballo. 2025c. Rethinking psychometrics though llms: How item semantics shape measurement and prediction in psychological questionnaires. *Nature Scientific Reports*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Sonja Schmer-Galunder, Ruta Wheelock, Zaria Jalan, Alyssa Chvasta, Scott Friedman, and Emily Saltz. 2024. Annotator in the loop: A case study of in-depth rater engagement to create a prosocial benchmark dataset. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1319–1328.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. *Advances in Neural Information Processing Systems*, 37:139179–139212.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, and 1 others. 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.

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*. Association for Computational Linguistics.

Peter Tyrer, Roger Mulder, Mike Crawford, Giles Newton-Howes, Erik Simonsen, David Ndetei, Nestor Koldobsky, Andrea Fossati, Joseph Mbatia, and Barbara Barrett. 2010. Personality disorder: a new global perspective. *World Psychiatry*, 9(1):56.

Bar Urkin, Josef Parnas, Andrea Raballo, and Danny Koren. 2024. Schizophrenia spectrum disorders: an empirical benchmark study of real-world diagnostic accuracy and reliability among leading international psychiatrists. *Schizophrenia Bulletin Open*, 5(1):sgae012.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, and 1 others. 2024. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*.

Xi Wang, Anxo Perez, Javier Parapar, and Fabio Crestani. 2025. Talkdep: clinically grounded llm personas for conversation-centric depression screening. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6554–6558.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Catherine Winsper, Ayten Bilgin, Andrew Thompson, Steven Marwaha, Andrew M Chanen, Swaran P Singh, Ariel Wang, and Vivek Furtado. 2020. The prevalence of personality disorders in the community: a global systematic review and meta-analysis. *The British Journal of Psychiatry*, 216(2):69–78.

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. Mentalchat16k: A benchmark dataset for conversational mental health assistance. *Preprint*, arXiv:2503.13509.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12226–12234.

Junbo Zhang, Ran Chen, Qianli Zhou, Xinyang Deng, and Wen Jiang. 2025. Understanding and mitigating over-refusal for large language models via safety representation. *arXiv preprint arXiv:2511.19009*.

Tianlin Zhang, Kailai Yang, Shaoxiong Ji, Boyang Liu, Qianqian Xie, and Sophia Ananiadou. 2024. Suicidemoji: Derived emoji dataset and tasks for suicide-related social content. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1136–1141.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. Is in-context learning sufficient for instruction following in llms? *arXiv preprint arXiv:2405.19874*.

Shengqi Zhu and Jeffrey Rzeszotarski. 2024. "get their hands dirty, not mine": On researcher-annotator collaboration and the agency of annotators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8773–8782.

A Exploratory Ablations

A.1 Emotional profiles as PDs' Markers

Following the work of De Grandi et al. (2024), we aim to evaluate whether it is possible to identify specific *emotional profiles* within PDs that may serve as markers for distinguishing among different PDs to answer. To this end, we employed the model introduced in De Grandi et al. (2024), constructing the emotional profile of each user on the basis of the emotions expressed across their posting history.

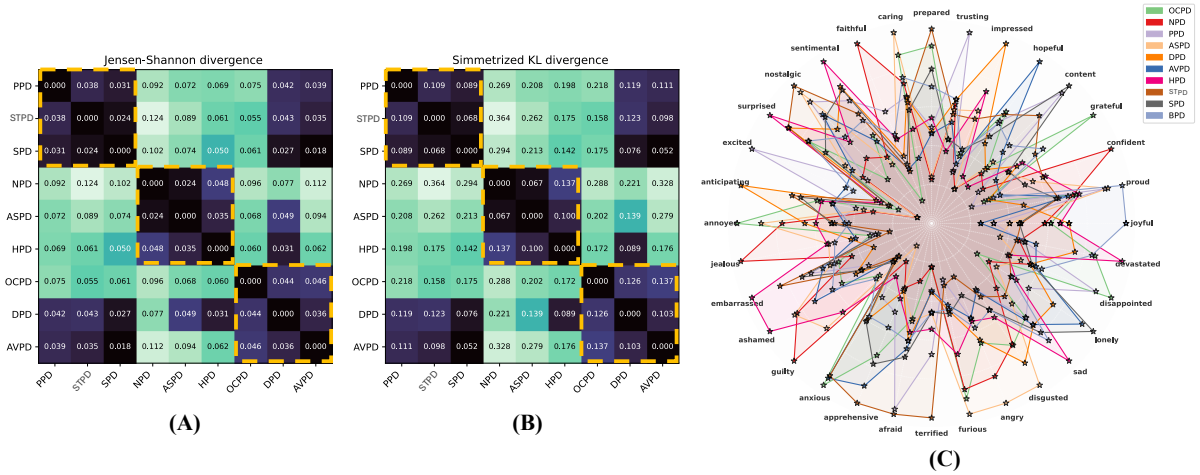

Figure 5: Emotional profile distributions across PD spectra. The matrices show the pairwise divergence of emotion distributions between PDs using **(A)** JS divergence and **(B)** Symmetrized KL divergence. **(C)** shows a radar plot of emotion frequencies conditioned on each PD.

The model from De Grandi et al. (2024) is a fine-tuned version of Mistral 7B [1], designed to infer the underlying emotion conveyed in a text. We follow the model's original prompting scheme based on its special tokens: each Reddit post is provided as input using the following structure, where the model generates the corresponding emotional label after the token <|emotion|>:

<|prompter|> p <|endoftext|> <|emotion|> $\rightarrow e$

where p is the input post and e is the emotion predicted by the model.

For each post, we extracted a probability distribution over emotions using a top-K sampling procedure with $K = 10$. The resulting distributions were then aggregated across all posts of a user with a given PD diagnosis and normalized so that the resulting vector lies on the probability simplex. Formally, let RP denote the set of Reddit posts for a given user, and $e_{i,k}$ the k-th sampled emotion from post i. The emotional profile of a user is defined as:

$$EmoProfile = \frac{1}{|RP|} \sum_{i \in RP} \frac{1}{K} \sum_{k=1}^{K} e_{i,k}$$

In cases of comorbidity, the emotional distribution was computed separately for each PD to which the user was assigned.

Figure 5 shows divergence matrices between emotional profiles, computed using *Jensen–Shannon* (JS) **(A)** and symmetrized *Kullback–Leibler* (KL) **(B)** divergence, obtained by averaging the KL divergence between pairs of emotional distributions in both directions to ensure symmetry, formally

$$\frac{1}{2}[KL(P\|Q) + KL(Q\|P)]$$

Lower values indicate greater similarity between PDs. Disorders within the same cluster show smaller divergence, confirming shared affective patterns. Notably, Schizoid PD is closer to Avoidant and Dependent PDs, consistent with literature describing overlapping traits of withdrawal and interpersonal dependency. Note that Borderline PD is excluded from this analysis due to its high rate of comorbidities, which would confound the interpretation of its emotional profile. These findings suggest that emotional profiles may serve as reliable affective markers for differentiating PDs while capturing shared emotional dynamics across related disorders. Finally, Figure 5**(C)** displays radar plots of scaled emotional profiles, highlighting discriminant emotions such as *apprehensive* and *afraid* for Schizotypal PD, *furious* and *disgusted* for Antisocial PD, *ashamed* for Histrionic PD, *confident* for Narcissistic PD, and *lonely* for both Schizoid and Avoidant PDs.

A.2 Vignette-based assessment of LLMs diagnostic performance

We leveraged the collected users' post histories to construct clinical vignettes - concise summaries of users' symptom trajectories - increasingly em-

[1] DeGra/RACLETTE-v0.2

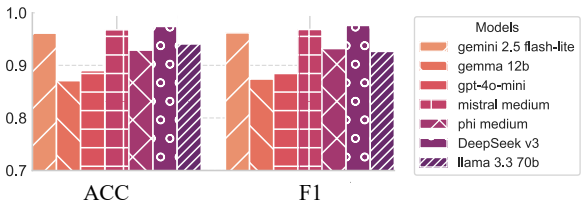

Figure 6: Performance of different LLMs on the vignette-based diagnostic prediction task based on accuracy (ACC) and macro F1 (F1).

ployed in recent literature to evaluate clinicians' diagnostic accuracy and to compare human performance with that of LLMs (Elyoseph et al., 2024; Urkin et al., 2024; Levkovich et al., 2024; Raballo et al.). In approximately $\sim 95\%$ of the users, users explicitly referred to adolescent experiences when describing past or formative life events, often reflecting the developmental onset typical of personality disorders. Each vignette was generated using `grok-3-mini` model, based on users' prior Reddit narratives, and designed to capture clinically relevant PD features expressed across time. In cases where multiple PDs were comorbid, one specific disorder was selected, and the LLM was explicitly instructed to focus on the diagnostic criteria of that disorder only. A total of 155 vignettes were produced and subsequently reviewed and refined by two licensed MH clinicians to ensure diagnostic coherence and avoid hybrid cases that ambiguously overlap multiple PD profiles. In only three cases, the vignettes were considered consistent with two possible PDs, as the users exhibited strong and balanced trait constellations across both diagnostic categories. We evaluated LLMs in a didactic diagnostic setting, asking each model to predict the most likely PD from clinically validated vignettes. This task simulates a training scenario, assessing whether models can integrate symptom narratives into coherent diagnostic hypotheses. As shown in Figure 6, many LLMs also used for **(RQ2)** perform strongly overall, demonstrating the ability to reason over structured clinical information. `deepseek-v3.1` , `gemini-2.5-flash-lite`, and `mistral-medium-3.1` achieved the highest accuracy and F1-macro, showing close alignment with clinician-assigned diagnoses. In contrast, `gpt-4o-mini` and `gemma-3-12b` performed slightly worse, particularly in distinguishing schizotypal from schizoid (misclassified 5/20 and 6/20 times, respectively), a confusion consistent with their overlapping interpersonal and affective features. Overall, these results show that LLMs can reach clinically meaningful diagnostic accuracy with structured case material, highlighting their potential for clinical education and benchmarking clinician–AI diagnostic agreement.

B Beyond Question Answering: Evaluating PD Consistency in Conversation

To assess whether steering transfers beyond controlled QA, we run a conversational diagnostic study where a questioning model interacts with a steered LLM and attempts to elicit diagnostically informative evidence over multiple turns. The interviewer is an LLM configured to ask discriminative clinical-style questions, while the respondent is a steered `Llama-3.2-3B-Instruct` model, selected because it showed the strongest steerability under steering vectors alone (i.e., without in-context learning) in our benchmark results.

Concretely, for each target personality disorder (PD), we generate 10 independent conversations. In each conversation, the interviewer (Gemini-3-Flash) is tasked with identifying the interlocutor's PD by asking questions over a fixed budget of 25 dialogue turns (see prompt in Section G.6). To promote varied questioning strategies while keeping the steered model's behavior deterministic, we set the interviewer's decoding temperature to 0.9 and keep the rest of the setup fixed across runs. The steered `Llama-3.2-3B-Instruct` model answers each question under the target PD steering configuration, with the goal of producing responses that remain behaviorally consistent with the intended PD profile throughout the interaction.

After the conversation ends, the full transcript is evaluated by a MH expert annotator, who assigns a PD diagnosis to the simulated patient. Because the dialogue is limited to 25 turns (and thus may not expose sufficient evidence for a single confident diagnosis), the annotator may optionally provide up to three candidate PDs when uncertain. A conversation is counted as a correct simulation if the target PD is included in the annotator's set of up to three predicted diagnoses.

The radar plot of the conversational diagnostic study in Figure 7 shows strong PD-specific differences in recognizability from dialogue: conditions such as NPD, BPD, and DPD (and, in many runs, ASPD and PPD) are classified correctly almost always, whereas SPD is markedly harder and

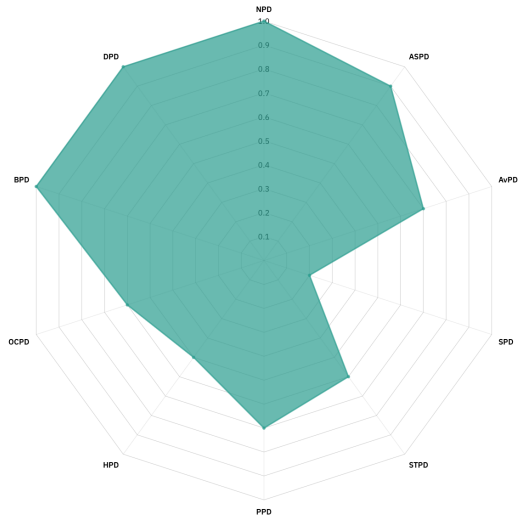

Figure 7: Conversational diagnostic accuracy of the steered simulated patients across PD targets (radar plot), computed over 10 conversations per disorder under a 25-turn interview budget.

is frequently confused with nearby Cluster A/C presentations (especially STPD, PPD, and AvPD). This pattern suggests that some PD profiles yield more salient, high-signal behavioral cues in dialogue, while others manifest through subtler absences (e.g., flattened affect, limited desire for intimacy) that are difficult to disambiguate in short interactions.

NPD/BPD/DPD (and often ASPD/HPD) tend to be easier because their characteristic interpersonal and affective dynamics create explicit, discriminative cues within few turns: entitlement/admiration-seeking and status framing (NPD), emotional volatility and relationship instability (BPD), reassurance-seeking and submissive dependency (DPD), rule-breaking/callousness (ASPD), and attention-seeking/dramatic expressivity (HPD). A second contributing factor is that these disorders (especially Cluster B presentations) often produce higher-intensity social signals that are more distinguishable from a neutral baseline persona, increasing separability in a diagnostic setting.

By contrast, SPD appears difficult and is often mistaken for STPD/PPD/AvPD because these conditions overlap in "distance from others" and reduced social engagement, but differ in *why* that distance occurs, differences that may require richer context than a short dialogue provides. In conversation, schizoid-like responses can look similar to avoidant (withdrawal driven by fear of criticism), paranoid (withdrawal driven by mistrust), or schizotypal (withdrawal plus odd beliefs/perceptual style), especially when the interaction does not elicit enough concrete evidence about motivation, internal experience, and long-term relational patterns.

Overall, this experiment shows that steering can produce personas that are not only QA-consistent but also diagnostically recognizable in open-ended conversation. Yet recognizability varies substantially by PD, with Cluster B-style high-salience interpersonal dynamics being easier to elicit and identify than low-expressivity detachment profiles such as SPD.

C Steering Benchmark

In this Section, we describe the construction of PersonaDSteering as a stepwise pipeline (visually described in Figure 8) that turns PRISMA-based user histories into aligned question–answer tuples across conditions (10 PDs plus a non-conditioned baseline), enabling controlled and reproducible evaluation of LLM steering, conceptually aligned with (Chen et al., 2025). We denote by $\mathcal{D}$ the set of conditions and by $\mathcal{C}_d$ the corpus associated with condition $d \in \mathcal{D}$ (PD-specific corpora drawn from PRISMA users' posts/comments, and a baseline corpus drawn from r/CasualConversation). The process starts by inducing a set of latent topics $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$ and generating candidate natural-language questions q. For each question q and condition d, we retrieve PD-specific supporting evidence by selecting a small set of posts/comments $E_{q,d} \subset \mathcal{C}_d$ that are maximally relevant to answering q under d. Retrieval is performed via a three-stage ranking pipeline: first, a lexical candidate-generation step based on BM25 assigns each document $x \in \mathcal{C}_d$ a score $s_{\text{bm25}}(q, x)$ and yields an initial candidate set $A_{q,d}$ (top-N by s_{bm25}). Second, we apply dense retrieval with `Qwen/Qwen3-Embedding-8B` by computing embeddings $e(\cdot)$ and semantic similarity $s_{\text{dense}}(q, x) = \cos(e(q), e(x))$, then selecting the top-20 most similar items $B_{q,d} = \text{TopK}_{x \in A_{q,d}}(s_{\text{dense}}(q, x), 20)$. Third, we rerank $B_{q,d}$ using `Gemini-3.0-Flash` with a learned relevance signal $s_{\text{rerank}}(q, x)$ and retain the top-5 final evidence items $E_{q,d} = \text{TopK}_{x \in B_{q,d}}(s_{\text{rerank}}(q, x), 5)$, which prioritize both topical relevance and answer usefulness. Conditioned on the question and its retrieved evidence, we generate one response per condition, producing

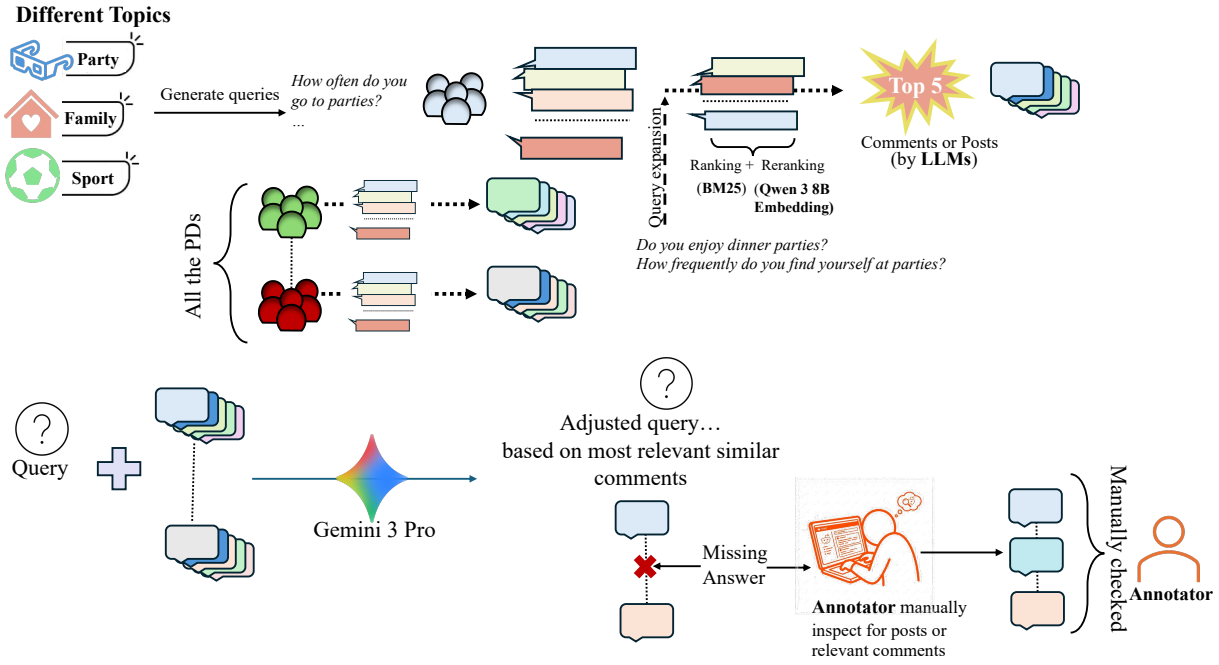

Figure 8: PersonaDSteering construction workflow: topic-guided question generation, evidence retrieval, and aligned PD-conditioned QA generation.

an aligned tuple $\left(q, \{a_{q,d}\}_{d \in \mathcal{D}}\right)$, where each $a_{q,d}$ is generated as $a_{q,d} = G(q, E_{q,d}, d)$ for a generator G that uses the evidence to synthesize a condition-consistent answer grounded in user text; for the baseline $d = b$, evidence is retrieved from the non-conditioned corpus so that the resulting answer serves as a neutral comparison target. To reduce superficial variation (e.g., length, formatting, and stylistic idiosyncrasies) that could confound cross-condition comparisons, all generated answers are subsequently normalized with *Gemini-3-pro*, yielding $\tilde{a}_{q,d} = N(a_{q,d})$ for a normalization operator $N(\cdot)$ that enforces consistent surface form while preserving the underlying content and behavioral signal. Finally, when no suitable response can be produced for a given condition due to insufficient or mismatched evidence, the corresponding entry is completed by an annotator through targeted Reddit search aligned with the question's topic, ensuring that each question maintains full coverage across $\mathcal{D}$ and preserving the aligned structure required for standardized steering evaluation.

D SCID-5-PD

The *Structured Clinical Interview for DSM-5 Personality Disorders* (*SCID-5-PD* First et al. (2016)) is a semi-structured diagnostic interview developed by the American Psychiatric Association to assess the ten DSM-5 personality disorders, grouped into Clusters A, B, and C. It provides categorical (presence/absence) assessments, supporting flexible use in clinical and research contexts.

Each disorder is evaluated through a standardized set of questions corresponding to DSM-5 diagnostic criteria, rated on a 0–2 scale, where 0 indicates absence, 1 subthreshold presence, and 2 that the criterion is met. A diagnosis is assigned when the number of "2" ratings meets the DSM-5-defined threshold (see Table 3). The SCID-5-PD is designed to build rapport, facilitate clinical judgment, and ensure diagnostic reliability across evaluators.

Importantly, although SCID-5-PD is traditionally administered as a clinician–patient interview, SCID items are not arbitrary conversational prompts: they are standardized probes that directly instantiate the underlying DSM-5-based (APA, 2013) criteria. For example, for Avoidant Personality Disorder, the question "Have you avoided jobs or tasks that involve dealing with many people? Can you give me examples? What was the reason - was it simply dislike of being around people, or fear of criticism or rejection?" maps directly onto Criterion 1 ("Avoids occupational activities that involve significant interpersonal contact because of fears of criticism, disapproval, or rejection."). This tight coupling between item text and diagnostic construct makes SCID-5-PD a high-fidelity frame-

work for identifying specific, clinically meaningful psychopathological features.

In our setting, we do not transplant the SCID interview format onto social media. Instead, we repurpose SCID-5-PD as a criterion-level evidence-extraction framework: we code whether a user's naturalistic text contains explicit evidence corresponding to each DSM-based criterion, rather than attempting a full diagnostic determination. This use is aligned with the SCID-5-PD manual, which permits coding criteria from narrative materials (e.g., case notes, letters, collateral reports) when a full interview is unavailable (First et al., 2016). Conceptually, this is also consistent with how clinical constructs are frequently measured outside a live interview via validated self-report instruments that capture self-expressed experiences and behaviors without clinician interaction.

Accordingly, the goal is not to "administer SCID on Reddit", but to leverage SCID's operationalized criteria as a structured, clinically grounded taxonomy for coding the presence of specific features in user-generated narratives. At the same time, we explicitly acknowledge that social media posts differ from interviews in depth, structure, and context; for this reason, we treat SCID coding as evidence extraction rather than diagnosis. The central question is whether users' narratives contain factual elements that correspond to SCID criteria, enabling systematic analysis while preserving the conceptual integrity of the SCID framework.

E The Hierarchical Taxonomy of Psychopathology (HiTOP)

The *Hierarchical Taxonomy of Psychopathology (HiTOP)* model (Kotov et al., 2017) was developed to overcome key limitations of traditional categorical systems such as DSM (APA, 2013) and ICD (Organization et al., 1992), including comorbidity, diagnostic instability, and within-disorder heterogeneity. Unlike categorical nosologies, HiTOP organizes psychopathology as a hierarchy of continuous dimensions derived empirically from symptom covariation.

At the top of the hierarchy lies a General Factor of Psychopathology (p-factor), reflecting the shared liability across all mental disorders. Below it, HiTOP defines several superspectra and spectra, which represent broad domains of dysfunction: *Internalizing*, *Thought Disorder*, *Detachment*, *Disinhibited Externalizing*, *Antagonistic Externalizing*, and *Somatoform*.

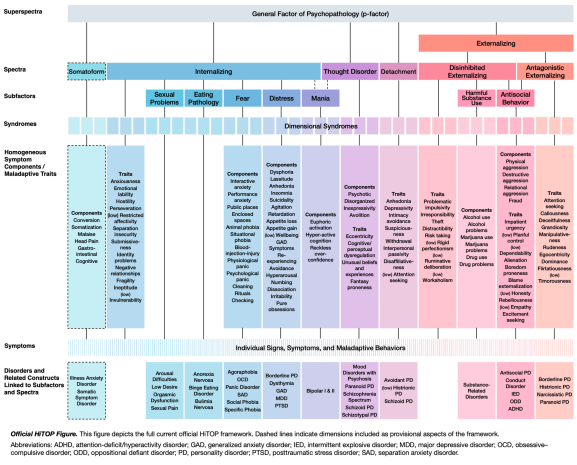

Figure 9: Official *Hierarchical Taxonomy of Psychopathology (HiTOP)* framework. The figure illustrates the hierarchical organization of psychopathology, from the general factor (*p*-factor) down to spectra, subfactors, syndromes, and maladaptive traits. Image is taken from the HiTOP Consortium, available at https://www.hitop-system.org

Each spectrum encompasses narrower subfactors (e.g., *Distress* and *Fear* within Internalizing), syndromes (e.g., *Major Depression*, *PTSD*, *Social Phobia*), and finally symptom components or maladaptive traits (e.g., *anxiousness*, *hostility*, *impulsivity*, *detachment*).

This hierarchical organization allows flexible granularity - from broad liability dimensions to specific traits - and facilitates the integration of personality and clinical disorders within a unified framework. Importantly, HiTOP dimensions are empirically derived from large-scale factor analytic studies, enhancing reliability and minimizing arbitrary diagnostic boundaries.

Figure 9 (taken from the HiTOP Consortium) illustrates the official structure of HiTOP, showing the hierarchical relations between superspectra, spectra, subfactors, syndromes, and traits. Consistent color coding highlights the correspondence between spectra, their lower-level traits, and related disorders.

F Adaptive Retrieval-Augmented Generation (aRAG)

The *adaptive Retrieval-Augmented Generation* (aRAG) framework generalizes traditional RAG by introducing an adaptive mechanism in the retrieval phase, allowing the number of retrieved doc-

	PPD	SPD	STPD	ASPD	BPD	HPD	NPD	AvPD	DPD	OCPD
Total Criteria	7	7	9	7	9	8	9	7	8	8
Threshold Required	≥ 4	≥ 4	≥ 5	≥ 3	≥ 5	≥ 5	≥ 5	≥ 4	≥ 5	≥ 4

Table 3: Diagnostic criteria and thresholds for Personality Disorders (SCID-5-PD).

uments to vary according to the semantic structure of the query. Given a query q and a corpus $D = \{d_1, \ldots, d_n\}$, the goal is to generate an informed output y through a generative model g_θ as:

$$y = g_\theta(q, \mathcal{R}^*(q)),$$

where $\mathcal{R}^*(q) \subset D$ is the adaptively retrieved subset of documents most relevant to q.

Unlike standard RAG approaches that fix the number k of retrieved documents, aRAG determines the optimal neighborhood size k^* dynamically, based on the local semantic density around q. Both queries and documents are embedded in a shared semantic space R^D using a dense retrieval models, specifically, in the case of this work, `msmarco-MiniLM-L-12-v3`[2], and their similarity is computed as:

$$\text{sim}(q, d_i) = \frac{\langle q, d_i \rangle}{\|q\| \, \|d_i\|},$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product. The adaptive retrieval set is then defined as:

$$\mathcal{R}^*(q) = \{d_i \mid \text{sim}(q, d_i) \geq \tau^*(q)\},$$

with $\tau^*(q)$ (or equivalently k^*) determined through intrinsic dimensionality estimation methods such as ABIDE (Di Noia et al., 2024). The optimal k^* corresponds to the largest neighborhood in which semantic meaning remains stable, identified by the condition:

$$\frac{\rho_{k^*+1}}{\rho_{k^*}} \approx 1,$$

where ρ_k denotes the distance to the k-th nearest neighbor. This ensures that only semantically coherent documents are retrieved, avoiding irrelevant or noisy context.

Finally, the generative process is expressed as:

$$y = g_\theta\left(q, \{d_i \in D : \text{sim}(q, d_i) \geq \tau^*(q)\}\right),$$

producing an output grounded in adaptively filtered knowledge. The adaptivity of aRAG thus enables the framework to generalize across domains with heterogeneous data density, mitigating redundancy and improving semantic alignment between retrieved context and generative inference. In essence, aRAG extends classical RAG into a context-sensitive retrieval–generation pipeline where the amount of evidence is not fixed but dynamically optimized based on the density of the embedding representations.

G Prompts Used

G.1 Evidence Classification

For this task, we employed the following LLMs: `gpt-4o-mini`, `gemini-2.5-flash-lite`, and `mistral medium 3.1`, with `temperature = 0` to force deterministic outputs.

Prompt:

Task: Classify whether the following post contains mental health related content.

Instructions:

1. **YES:** The post discusses

   mental health

   topics, symptoms, treatments, experiences, conditions, therapy, medications, or any aspect related to psychological wellbeing.

2. **NO:** The post contains content completely unrelated to mental health (e.g., sports, cooking, technology, general news, entertainment, etc.).

Post: {POST}

Question: Does the post contain mental health

related content?

Response format: YES/NO

[2] https://huggingface.co/sentence-transformers/msmarco-MiniLM-L-12-v3

G.2 Diagnostic Criteria Classification

All the models discussed in Table 2 were run with `temperature = 0` to force deterministic outputs.

Prompt without aRAG:

Analyze the provided Reddit posts against the SCID-5-PD diagnostic criteria listed below.

 # Diagnostic Criteria for {DIAGNOSIS}

{CRITERIAS}

 # Scoring

For each criterion, assign:

- **0** = No evidence OR subthreshold evidence (the behavior/pattern is absent, unclear, or does not meet the threshold for clinical significance)
- **1** = Clear evidence (the posts demonstrate a consistent, unambiguous pattern that meets the criterion)

 # Rules

- Base assessment only on explicit post content
- Look for consistent patterns, not isolated incidents
- Consider context (jokes vs. genuine expressions)

 # Reddit Posts

{POSTS}

 # Output

Return ONLY a list of scores, one per criterion, in order:

[score1, score2, score3, etc.]

G.3 Prompt for Steering

The prompt used to reproduce the experiments in Figure 4 is identical across all evaluated models. Decoding is performed with temperature set to 0.

You are Dave, a 35-year-old man. Answer the question concisely and clearly (in first person), choosing only one option.

{QUESTION}

G.4 Vignettes Generation

To generate the vignettes, we used the `grok-3-mini` model, prompting it to follow the style and structure of two reference cases ("David" and "Michael") from the vignettes presented in Raballo et al..

Prompt:

Instruction: Based on the provided Reddit posts and the user's mentioned adolescent memories, create a clinical vignette following the structure of the two given examples. The vignette must include all the relevant information for the diagnosis in a concise way.

Example 1: {DAVID}

Example 2: {MICHAEL}

Details: Focus only on information explicitly stated in the posts and in the memories the user describes from their teenage years. Construct the vignette knowing that the diagnosis is {DIAGNOSIS}, **without citing the diagnosis** in the vignette. Stay strictly within the factual content provided, avoiding interpretations or assumptions, and highlight the evidences relevant for the diagnosis.

You are also provided with the SCID-5-PD criteria to guide evidence selection: {CRITERIAS}

Inputs:

- **Posts:** {POSTS}
- **Patient name:** {NAME}

Note: The vignette must not exceed 400 words.

G.5 Vignettes Classification

All the models discussed in Figure 6 were run with `temperature = 0` to force deterministic outputs.

Prompt:

Analyze the following psychiatric vignette and provide a diagnosis of a specific Personality Disorder.

Vignette:

{VIGNETTE}

Based on the clinical presentation, symptoms, timeline, and other relevant factors described, identify the most likely primary diagnosis.

Output format:

diagnosis

G.6 Multi-turn Conversations

Prompt:

You are a clinical interviewer assessing different Personality Disorders (Avoidant, Dependent, Obsessive-Compulsive, Paranoid, Schizotypal, Schizoid, Histrionic, Narcissistic, Borderline, Antisocial) in a simulated patient named Dave.

Ask exactly 25 concise, open-ended questions, one at a time, and wait for Dave's reply after each question.

Your goal is to understand whether Dave shows a specific personality disorder or not

Do NOT mention DSM, criteria letters/numbers, or the SCID. Keep questions concise, factual, and non-accusatory. Ask only one simple and concise question at a time.