TOKEN-WISE REWARD GUIDED REASONING ALONG DIAGNOSTIC CRITERIA FOR DEPRESSION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) offer broad-coverage early screening while providing natural-language descriptions of mental health risks. Fine-tuning on medical annotations is costly and risks eroding their general capabilities. Inference-time alignment offers a flexible alternative, steering frozen LLMs toward desired preferences through reward models and thereby injecting medical knowledge into depression analysis without fine-tuning the LLM. However, existing methods spread the reward signal uniformly across all tokens, diluting supervision on the few tokens that bear decisive diagnostic evidence. To address this, we propose the AlignDx framework, which recasts depression detection as a Reason-then-Assess task in which the model reasons along diagnostic dimensions grounded in established medical criteria, and a downstream classifier derives the final risk prediction from this structured analysis. During reward model training, gradient-based importance estimation quantifies each token's contribution and reweights the autoregressive reward objective accordingly, concentrating the optimization signal on decisive tokens; at inference, reward-guided decoding via logit-space fusion injects diagnostic criteria into generation. Experiments on three datasets show that AlignDx outperforms both traditional classifiers and prior LLM-based approaches in in-distribution and cross-platform settings, with gains driven by stronger high-risk detection. Our code and datasets are available at `https://anonymous.4open.science/r/AlignDx-3158`.

1 INTRODUCTION

Large Language Models are increasingly introduced into depression detection on social media, a task that requires inferring latent mental states from users' everyday behavioral traces. Situated at the intersection of machine learning, affective computing, and digital mental health, this emerging interdisciplinary field aims to offer a non-invasive and scalable complement to early clinical screening. The WHO estimates that approximately 5.7% of adults worldwide experience depression Lan et al. (2025), and depressive disorders are closely associated with elevated suicide risk. Against this backdrop, social media-based depression risk early detection offers an alternative screening approach with broad population coverage De Choudhury et al. (2013). Compared with structured self-reports in clinical settings, social media posts often reflect more spontaneous emotional disclosure and enable longitudinal monitoring. They therefore support large-scale early risk detection without imposing additional financial burdens on individuals Chancellor & De Choudhury (2020).

With their strong semantic understanding and reasoning capabilities, LLMs have shown great promise for depression detection on social media Kuzmin et al. (2025). Existing studies have leveraged LLMs to construct multi-source instruction datasets to jointly yield predictions and interpretable analyses, or to incorporate medical criteria such as the DSM-5 for structured symptom annotation Lan et al. (2025); American Psychiatric Association (2013). Nevertheless, they largely treat generated analyses or reasoning traces as auxiliary by-products and do not explicitly model which fragments are decisive for the final diagnosis Yang et al. (2024). More fundamentally, as illustrated in Figure 1, empirical evaluations indicate that when generating depression risk reasoning, LLMs tend to take shortcuts by relying on explicit keywords, while struggling with implicit expressions and complex contexts Shao et al. (2025); Ge et al. (2025). Consequently, predictions can hardly be consistently traced back to critical evidence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

While fine-tuning on medical annotations is expensive and may compromise the model's general capabilities, prompting alone tends to produce alignment that is shallow and unstable. Inference-time alignment has recently gained attention for its ability to steer frozen LLMs toward desired behaviors via reward models without fine-tuning the LLM. Since diagnostic criteria emphasize persistent symptoms and functional impairment rather than isolated emotional expressions, aligning LLMs with such medical preferences naturally counteracts the lexical shortcuts. However, directly adapting existing inference-time alignment methods to depression risk analysis faces two challenges. First, reward signals are distributed uniformly across tokens, failing to distinguish decisive fragments from redundant background information. Second, the reasoning traces are misaligned with medical diagnostic criteria. **How to precisely inject diagnostic criteria while selectively reinforcing critical evidence remains an open question.**

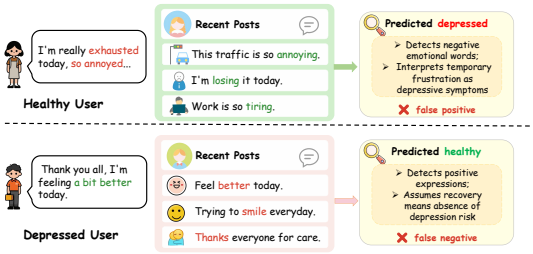


Figure 1: Illustration of lexical shortcut in depression detection. Models over-rely on explicit expressions, causing false predictions by ignoring diagnostic evidence.

To address these challenges, we reformulate depression risk analysis as a text generation task, in which the model identifies depression-related evidence from social media posts and derives its prediction from the reasoning trace. Guided by well-established diagnostic criteria, we further structure the risk analysis into multiple complementary dimensions, thereby eliciting high-quality explanations from LLMs. To operationalize this formulation, we propose **AlignDx**, a depression detection framework built upon the following insight: if a small perturbation of a token substantially changes the final prediction, that token exerts greater influence on the decision and should receive a larger optimization weight. To enable reward modeling, we introduce MDPD-10k, a multidimensional depression analysis preference dataset. Based on it, AlignDx employs gradient-based token importance estimation to quantify the sensitivity of the final prediction to each token and adaptively assigns optimization weights. These weights are incorporated into an autoregressive reward-modeling objective Xu et al. (2025), concentrating the training signal on critical tokens rather than distributing it uniformly across the sequence Yang et al. (2023). At inference time, AlignDx runs the base model and the trained reward-model LoRA adapter Hu et al. (2022) in parallel branches and performs reward-guided decoding through logit-space fusion, injecting diagnostic preferences whose strength can be flexibly tuned at inference time, without modifying the base model. A lightweight downstream classifier then assesses the generated multidimensional report to produce the final risk prediction, grounding the decision in traceable diagnostic evidence rather than raw, noisy posts.

Experiments are conducted on three social media datasets with two LLM backbones, comparing against traditional classification methods and state-of-the-art LLM-based approaches. Experimental results demonstrate that AlignDx outperforms all baselines under in-distribution and cross-platform settings. Specifically, compared to the advanced competitor GPT-5.4-mini, AlignDx achieves relative improvements of 9.3% in MCC and 4.3% in F1 score on the in-distribution benchmark. Under the cross-platform transfer from Weibo to Reddit, AlignDx further improves MCC and F1 by 4.8% and 5.0% relatively on eRisk, demonstrating robustness across platforms.

Our main contributions are summarized as follows: (1) We reformulate depression risk analysis as a Reason-then-Assess process, where the model first generates a multi-dimensional structured analysis and a downstream classifier derives the risk decision from this explicit reasoning, grounding predictions in traceable and interpretable evidence. (2) We propose AlignDx, which estimates each token's influence via gradients and incorporates these weights into an autoregressive reward-modeling objective, concentrating optimization on the most critical tokens. At inference time, AlignDx performs logit-space fusion to inject the learned diagnostic preference into the decoding process. (3) Experiments on the MDPD, WU3D, and the cross-platform eRisk dataset show that AlignDx consistently outperforms existing methods, with gains stemming primarily from improved identification of minority-class users at high depression risk, highlighting its potential to serve as a reference for medical diagnosis and early screening.

2 Related Work

Depression Detection on Social Media. Social media posts capture longitudinal trajectories of users' emotions, symptoms, behaviors, and life events, providing a unique data foundation for early depression detection De Choudhury et al. (2013); Chancellor & De Choudhury (2020). Early approaches Coppersmith et al. (2014); Yates et al. (2017) primarily relied on hand-crafted linguistic, affective, and behavioral features to extract depression signals from users' posting histories. With advances in representation learning, research gradually shifted toward neural encoding and the mining of clinically grounded signals to capture deeper semantic patterns Chen et al. (2025). However, depression-related posting behaviors differ across platforms. For example, the posting activity of depressed users exhibits opposite trends on Twitter and Weibo, causing severe performance degradation when models are transferred across platforms Shen et al. (2018). This reality has driven the field to move from learning surface-level statistical patterns toward constructing transferable clinical representations for improved cross-domain generalization.

LLMs for Depression Analysis. The emergence of LLMs has shifted depression detection from label prediction toward richer analysis grounded in symptoms, causes, and contextual evidence Garg et al. (2022). Early explorations leveraged semantic understanding and generation capabilities to construct instruction datasets for jointly producing predictions and interpretable analyses Yang et al. (2024); Kuzmin et al. (2025). More recent efforts Lan et al. (2025) have incorporated medical knowledge, such as structured symptom annotation according to diagnostic criteria and the integration of emotional trajectory summarization with downstream classification. Nevertheless, existing approaches Lan et al. (2025); Mahdavinejad et al. (2024) largely treat reasoning traces as auxiliary by-products, lacking explicit modeling of which fragments are decisive for judgment. Empirical investigations Shao et al. (2025); Ge et al. (2025) indicate that LLMs tend to take lexical shortcuts by over-relying on explicit keywords, while exhibiting insufficient targeted optimization for implicit expressions and complex contexts. Consequently, medical consistency, explanation reliability, and robustness to bias have emerged as central challenges in mental health detection.

Inference-Time Alignment. Beyond training-stage alignment methods Ouyang et al. (2022); Rafailov et al. (2023), inference-time alignment has emerged as a flexible paradigm that steers model behavior without retraining. Early approaches Nakano et al. (2021) performed preference selection after sequence generation through reranking or rejection sampling, incurring computational costs that grow linearly with the number of candidates. To enable fine-grained generation guidance, recent works have explored decoding-time interventions, including reward-guided search Khanov et al. (2024), distributed rejection sampling Li et al. (2024), and chunk-level value function guidance Liu et al. (2024). In particular, the introduction of autoregressive reward models Xu et al. (2025); Lin et al. (2025) decomposes sequence-level evaluation into token-level predictions, aligning reward signals with the autoregressive generation process. Such interventions enable preference control without costly retraining, providing a practical foundation for high-stakes LLM alignment.

3 Preliminary

We first introduce LLM preference alignment and its core concepts, which form the basis for optimizing response generation. Given an input prompt x, a model policy $\pi_\theta(y \mid x)$, and a carefully constructed preference dataset $\mathcal{D} = \{(x, y_w, y_l)\}$, y_w and y_l denote the preferred and non-preferred responses, respectively. The conventional RLHF paradigm first trains a reward model $r_\phi(x, y)$ and evaluates response quality using the Bradley–Terry Bradley & Terry (1952) model:

$$p(y_w \succ y_l \mid x) = \frac{\exp(r_\phi(x, y_w))}{\exp(r_\phi(x, y_w)) + \exp(r_\phi(x, y_l))}. \tag{1}$$

DPO Rafailov et al. (2023) further reparameterizes the reward function and expresses the implicit reward directly in terms of the policy π_θ and the reference policy π_{ref}:

$$r_\phi(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} - \log Z(x), \tag{2}$$

where $Z(x)$ is the partition function and β controls the divergence between the policy and reference distributions. The resulting DPO objective directly optimizes the log ratio difference:

$$\mathcal{L} = -E_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left([r_\phi(x, y_w) - r_\phi(x, y_l)]\right)\right]. \tag{3}$$

The optimal policy $\pi^\star$ can be explicitly represented by π_{base} and the reward Rafailov et al. (2023):

$$\log \pi^\star(y \mid x) = -\log Z(x) + \log \pi_{\text{base}}(y \mid x) + \frac{1}{\beta} r(x, y), \tag{4}$$

Eq. equation 4 shows that the optimal policy is essentially an exponential reweighting of the base distribution π_{base} according to the reward $r(x, y)$. This formulation establishes a direct correspondence between the policy and the reward, thereby enabling preference optimization without explicitly training a reward model, and provides a theoretical basis for subsequently modulating the base distribution through the reward term during the decoding stage.

4 METHODOLOGY

In this section, we propose AlignDx (Figure 2). Unlike previous discriminative settings, we formulate depression risk analysis as a report-based generation task Zeng et al. (2024); Yang et al. (2025). Given a user's social media posts, our goal is to generate an evidence-grounded depression risk analysis report. Generating such a report requires an LLM to identify diagnostically relevant evidence from the user's longitudinal behavior and organize this evidence into symptom-oriented interpretations. An LLM or a classifier determines the user's depression label from the risk report.

4.1 DIAGNOSTIC CRITERIA CONSTRUCTION

Multidimensional Criteria. Guided by widely accepted clinical criteria for diagnosing depression, spanning DSM-5, ICD-11, validated psychometric scales, psychological theories, and prior research American Psychiatric Association (2013); World Health Organization (2024), we distill multiple complementary diagnostic dimensions to inject medical knowledge and improve the reliability of social media–based depression detection. Specifically, these dimensions cover: (1) **Core Affect**, including depressed mood, hopelessness, and loss of interest as major depressive symptoms Kroenke et al. (2001); (2) **Cognitive Distortion**, capturing negative views of the self, world, and future as well as rumination patterns Beck et al. (1979); (3) **Somatic–Executive**, modeling sleep problems, fatigue, attention difficulties, decision-making issues, and appetite changes Hamilton (1960); (4) **Suicide Risk**, identifying high-risk signals such as death-related thoughts, self-harm tendencies, and suicidal plans; (5) **Linguistic Style**, capturing social media language patterns including absolutist expressions and self-focused language Al-Mosaiwi & Johnstone (2018); (6) **Behavioral Function**, reflecting social withdrawal, functional impairment, and maladaptive coping behaviors; and (7) **External Stress**, incorporating negative life events, environmental stressors, and anxiety-related manifestations.

Together, these dimensions offer comprehensive coverage of emotional, cognitive, physiological, risk-related, linguistic, behavioral, and contextual evidence. Building on them, a Judge LLM assesses each dimension along five criteria, namely coverage, accuracy, grounding, implicit-signal recognition, and objectivity, and the aggregated score serves as the overall quality measure. The complete prompts are provided in Appendix D.

Preference Set Generation. We construct the initial dataset upon SWDD Cai et al. (2023), a Weibo corpus offering substantial demographic and behavioral diversity. We retain all depression risk users to preserve behavioral patterns, and apply BGE-M3 embeddings Chen et al. (2024) with K-means clustering MacQueen (1967) to select representative non-risk users and reduce redundancy. After cleaning and sampling, the final dataset comprises 10k Weibo users. Conventional preference-set construction accepts or rejects each candidate holistically, overlooking that different models excel along different dimensions. Since depression-related evidence is inherently multifaceted, integrating their complementary strengths yields more comprehensive analyses. Motivated by this, we evaluate each dimension independently to capture fine-grained quality differences. Specifically, Qwen3.5-9B and Qwen3.5-27B each generate dimension-specific reasoning (three rollouts per model), and Qwen3.5-Plus serves as the judge to assign a quality score to every dimension.

Within each dimension, we take the highest- and lowest-scoring responses as the preferred and rejected samples, so that a single preference pair consolidates the strongest dimensions from different models. The preference pairs are organized into the Multidimensional Depression Assessment

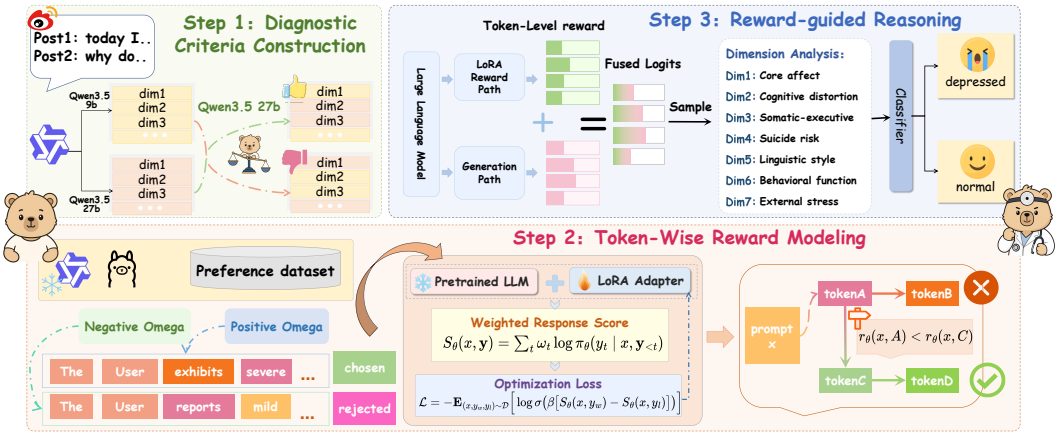


Figure 2: Overview of the AlignDx framework. AlignDx distill diagnostic criteria, learns token-wise rewards weighted by gradient-based importance, and performs reward-guided decoding via logit-space fusion to produce structured risk analyses for downstream classification.

Preference Dataset (MDPD-10k). The dataset provides comparative signals for preference learning, valued for relative rankings rather than absolute medical perfection. The detailed preference construction procedure is described in Appendix A.

4.2 Token-Wise Reward Modeling

Autoregressive Reward Reparameterization. High-stakes medical tasks such as depression detection require models to carefully align with fine-grained preferences, including accurate evidence extraction and precise symptom attribution, at inference time rather than relying solely on static training-time alignment or prompting. Therefore, we introduce a reward-guided alignment approach that steers the frozen model toward medical criteria at inference time, injecting fine-grained diagnostic guidance without any fine-tuning.

Existing sequence-wise reward models cannot directly assess the quality of the next generated token, preventing their reward signals from providing fine-grained guidance during autoregressive decoding. To bridge this gap, and inspired by GenARM Xu et al. (2025), we reparameterize the reward as a weighted autoregressive likelihood. This formulation decomposes a sequence-wise reward into an aggregation of token-wise log-probabilities, thereby aligning the reward with the autoregressive generation process. Specifically, for a generated depression analysis report y, we define

$$r_\theta(x, y) = \sum_t \omega_t \log \pi_\theta(y_t \mid x, y_{<t}), \tag{5}$$

where ω_t denotes the importance weight of the t-th response token. Uniform weights ($\omega_t = 1$) recover the standard autoregressive reward, whereas token-specific weights adaptively concentrate the optimization signal on decisive tokens rather than spreading it uniformly.

Gradient-Guided Token Importance Estimation. The key challenge therefore lies in estimating ω_t. Prior work Ballout et al. (2024) shows that gradient attribution offers a precise, data-driven measure of token influence. Building on this insight, we propose an offline gradient-based estimation strategy grounded in the following principle: if a small perturbation to a token representation substantially alters the model's prediction, that token exerts a stronger influence on the decision and should receive a larger optimization weight.

For a preference sample (x, y), let $h_i^{(\ell)}$ be the hidden representation at position i in layer ℓ after a forward pass, and let $o_{N,j}$ be the output logit for the j-th vocabulary token at the final position N. We adopt the maximum logit as scalar proxy for the model's final prediction tendency:

$$s = \max_j o_{N,j}. \tag{6}$$

We then compute the gradient of this scalar target with respect to the hidden representation at each layer and aggregate the resulting ℓ_1 norms to obtain a token-wise importance score:

$$g_i = \sum_{\ell=0}^{L} \left\| \frac{\partial s}{\partial h_i^{(\ell)}} \right\|_1 . \tag{7}$$

A larger gradient magnitude indicates that perturbing the token representation has a stronger effect on the final prediction. Since preference optimization updates only the response-generation process, we retain scores for response tokens alone. To prevent samples with substantially different gradient scales from dominating training, we normalize the scores within each response:

$$\omega_t = \frac{g_t}{\frac{1}{T}\sum_{k=1}^{T} g_k}. \tag{8}$$

Weighted Autoregressive Reward Objective. For compactness, we denote the weighted autoregressive score of a response by

$$S_\theta(x, y) = \sum_t \omega_t \log \pi_\theta(y_t \mid x, y_{<t}). \tag{9}$$

For each preference sample (x, y_w, y_l), the loss is:

$$\mathcal{L} = -E_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \left[S_\theta(x, y_w) - S_\theta(x, y_l) \right] \right) \right], \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid function and $\beta > 0$ is a scaling coefficient that controls the strength of preference optimization. By maximizing the weighted log-likelihood advantage of the preferred response over the non-preferred one, this objective compels the model to discriminate differences in analysis quality at the token level. Additional theoretical analysis is provided in Appendix B.

4.3 Reward-Guided Reasoning

Owing to the autoregressive design of the reward model, generating each token requires only a single forward pass through the base model and one through the reward model. Conventional reward-guided generation methods Nakano et al. (2021); Li et al. (2024) evaluate candidates only after complete sequences have been generated, causing the computational cost to grow linearly with the number of candidates. In contrast, our method fuses the two distributions directly at every decoding step:

$$\pi(y_t \mid x, y_{<t}) \propto \ \pi_{\text{base}}(y_t \mid x, y_{<t})^{\frac{1}{1+\alpha}} \cdot \pi_{\text{rm}}(y_t \mid x, y_{<t})^{\frac{\alpha}{1+\alpha}}. \tag{11}$$

where $\alpha \geq 0$ governs the strength of preference guidance. As α increases, the preferences encoded by the reward model exert a progressively stronger correction on the sampling distribution. This fusion mechanism allows the alignment strength to be tuned at inference time, enabling adaptation to application scenarios with varying levels of risk sensitivity without any model fine-tuning.

LLMs are poorly suited to producing calibrated confidence scores, which risk-threshold-based screening critically relies on. We therefore delegate the final risk decision to a lightweight downstream classifier ϕ, rather than soliciting a direct verdict from the LLM. Trained on gold-standard report–label pairs $(\mathbf{r}^*, c)$ from MDPD, ϕ is optimized with the binary cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\mathbb{E}_{(\mathbf{r}^*,c)} \left[c \log P_\phi(c = 1 \mid \mathbf{r}^*) + (1 - c) \log(1 - P_\phi(c = 1 \mid \mathbf{r}^*)) \right], \tag{12}$$

At test time, ϕ takes as input the report $\mathbf{r}$ generated by the aligned model:

$$p_{\text{cls}} = P_\phi(c = 1 \mid \mathbf{r}), \hat{c} = \mathbf{1}\{p_{\text{cls}} > \tau\}, \tag{13}$$

where τ is the depression risk threshold. Unlike prediction directly from raw social media text, this classifier takes explicit, structured depression reasoning as input. Consequently, the risk decision is grounded in multidimensional evidence attribution and symptom-level interpretations rather than noisy raw user-generated content.

Table 1: Performance comparison on the MDPD, eRisk, and WU3D datasets. The best and second-best results are highlighted in bold and underlined, respectively. MDPD is split into training and test sets, while eRisk and WU3D are used only for testing.

Model	MDPD (train-and-test)			eRisk (test-only)			WU3D (test-only)		
	ACC	F1	MCC	ACC	F1	MCC	ACC	F1	MCC
Classifier-based Methods									
TF-IDF	0.8715	0.8226	0.7252	0.8856	0.0000	-0.0167	0.6932	0.6660	0.5057
HAN	0.7990	0.7133	0.5591	0.5336	0.1908	0.0185	0.5661	0.5506	0.2708
BERT	0.8420	0.7638	0.6453	0.8878	0.0000	0.0000	0.7442	0.5160	0.3646
MentalRoBERTa	0.8360	0.7706	0.6446	0.8878	0.0000	0.0000	0.6729	0.5890	0.3553
LLM-based Methods: Direct Assess									
Gemini-3.1-Flash-Lite	0.8565	0.8078	0.7008	0.7822	0.4530	0.4074	0.8116	0.7689	0.6682
DeepSeek-V3.2	0.8590	0.8112	0.7064	0.8328	0.4899	0.4323	0.8685	0.8236	0.7432
GPT-5.4-mini	0.8645	0.8204	0.7212	0.8735	<u>0.5788</u>	<u>0.5334</u>	<u>0.9014</u>	<u>0.8615</u>	<u>0.7980</u>
LLaMA-3.1-8B-Instruct	0.7010	0.6865	0.5113	0.6667	0.3311	0.2550	0.7744	0.7323	0.6106
Qwen3-4B-Instruct-2507	0.6475	0.6521	0.4526	0.5413	0.2968	0.2301	0.6570	0.6477	0.4847
LLM-based Methods: Reason-then-Assess									
Gemini-3.1-Flash-Lite	0.8035	0.7571	0.6322	0.7680	0.4509	0.3850	0.7584	0.6964	0.5818
DeepSeek-V3.2	0.8353	0.7807	0.6674	<u>0.8885</u>	0.4516	0.4322	0.8625	0.8052	0.7185
GPT-5.4-mini	0.6570	0.6552	0.4556	0.6428	0.3388	0.2777	0.6597	0.6446	0.4736
LLaMA-3.1-8B-Instruct	0.7699	0.7356	0.5892	0.4488	0.2511	0.1466	0.5608	0.5842	0.3644
–Best-of-N (N=5)	0.6896	0.6214	0.4508	0.6302	0.2707	0.1957	0.7122	0.6776	0.5258
–SFT	0.7790	0.7424	0.5991	0.7503	0.3606	0.2005	0.8144	0.7673	0.6257
–DPO	0.7940	0.7553	0.6198	0.7712	0.3810	0.2134	0.8251	0.7776	0.6396
–GenARM	0.6153	0.6293	0.4069	0.5243	0.2785	0.1915	0.6357	0.6315	0.4548
–AlignDx	0.8593	0.7686	0.6799	0.8562	0.5328	0.4760	0.8880	0.8336	0.7524
Qwen3-4B-Instruct-2507	0.7700	0.7359	0.5895	0.7624	0.4066	0.3432	0.7441	0.7056	0.5696
–Best-of-N (N=5)	<u>0.8940</u>	<u>0.8392</u>	<u>0.7607</u>	0.8704	0.5520	0.4932	0.8982	0.8510	0.7791
–SFT	0.7465	0.7200	0.5665	0.6513	0.3517	0.2050	0.7351	0.7034	0.5431
–DPO	0.7495	0.7218	0.5688	0.6546	0.3539	0.2064	0.7363	0.7043	0.5443
–GenARM	0.7549	0.7206	0.5627	0.7899	0.4399	0.3815	0.7638	0.7221	0.5950
–AlignDx	**0.9060**	**0.8553**	**0.7882**	**0.8891**	**0.6080**	**0.5588**	**0.9081**	**0.8656**	**0.8015**

5 EXPERIMENT

In this section, we empirically evaluate AlignDx against existing depression detection methods and investigate its underlying mechanisms through four research questions: **RQ1:** Does AlignDx outperform existing baselines while maintaining stable generalization under cross-platform settings? **RQ2:** How does the Reason-then-Assess paradigm affect performance? **RQ3:** Which components of AlignDx are indispensable to its performance, and what trade-off between alignment strength and content quality does the framework exhibit?

5.1 EXPERIMENT PROTOCOL

Baseline Methods. (1) Traditional classification methods: TF-IDF combined with Logistic Regression Ramos et al. (2003), HAN Yang et al. (2016), BERT Devlin et al. (2019), and MentalRoBERTa Ji et al. (2022); (2) zero-shot direct prompting based on general LLMs, including the base model, SFT Ouyang et al. (2022), DPO Rafailov et al. (2023), Best-of-N Nakano et al. (2021) and GenARM Xu et al. (2025). We trained Qwen3-0.6B as a classifier. All methods requiring training or fine-tuning are trained on MDPD. Additional experimental details are provided in Appendix C.

Dataset. To comprehensively validate the model's performance, we evaluate our method on the MDPD, WU3D, and eRisk datasets. The MDPD dataset is divided into training, validation, and test sets in a ratio of 7:1:2. WU3D is a public Weibo depression detection dataset Wang et al. (2020); given the large scale of its original data, we randomly sample 10% of the users, totaling 3,256 individuals, to serve as the test subset. The eRisk dataset Parapar et al. (2025); Crestani et al. (2022) originates from the CLEF eRisk evaluation task and contains longitudinal Reddit posts with binary depression labels, we utilize all 909 users for testing.

Evaluation Metrics. To comprehensively evaluate the performance of depression risk classification, this study employs Accuracy, F1 score, and the Matthews Correlation Coefficient (MCC) Chicco & Jurman (2020) as evaluation metrics. Specifically, Accuracy reflects the overall classification correctness of the model; the F1 score evaluates the balance between precision and recall when identifying users at risk of depression; and the MCC, a more balanced and robust performance evaluation in scenarios with imbalanced class distributions.

5.2 Main Results

In-Distribution Performance (RQ1). AlignDx with Qwen3-4B-Instruct-2507 as the base model achieves the strongest overall performance on MDPD. Notably, the unadapted Qwen3-4B-Instruct-2507 is the weakest among all LLM-based baselines on MDPD, with an MCC of only 0.4526. This substantial contrast indicates that the performance gains do not stem merely from the inherent capacity of the base model. Instead, the proposed token-wise reward model training and reward-guided decoding effectively enhance the model's ability to transform longitudinal user posts into discriminative and structured representations for depression risk identification.

Out-of-Distribution Generalization (RQ1). The cross-dataset results further demonstrate the generalization capability of AlignDx. On eRisk, AlignDx achieves more pronounced improvements in F1 and MCC than in Accuracy, indicating that its gains primarily arise from improved identification of minority-class users at high risk of depression, rather than simply favoring the majority non-depressed class. Under the substantial distribution shift from Chinese Weibo to English Reddit, BERT and MentalRoBERTa obtain relatively high Accuracy on eRisk, yet both yield F1 and MCC scores of zero. This result suggests that these models collapse into predicting a single class under severe class imbalance, thereby losing meaningful discriminative ability in practical depression risk identification.

Table 2: Ablation results on MDPD, WU3D, and eRisk. The best results are highlighted in bold. "w/o" denotes the removal of the corresponding component.

Dataset	Model Variant	ACC	F1	MCC
MDPD	Full Model	**0.9060**	**0.8553**	**0.7882**
	w/o Classifier	0.7765	0.7357	0.5852
	w/o Weight	0.8891	0.8338	0.7512
	w/o Reward Model	0.8887	0.8322	0.7492
WU3D	Full Model	**0.9081**	**0.8656**	**0.8015**
	w/o Classifier	0.7638	0.7221	0.5950
	w/o Weight	0.8938	0.8475	0.7743
	w/o Reward Model	0.9024	0.8552	0.7853
eRisk	Full Model	**0.8891**	**0.6080**	**0.5588**
	w/o Classifier	0.7899	0.4399	0.3815
	w/o Weight	0.8850	0.6008	0.5506
	w/o Reward Model	0.8884	0.5944	0.5441

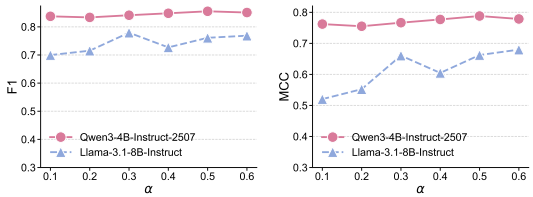
Effectiveness of Alignment (RQ2). To separate the effect of the Reason-then-Assess paradigm from AlignDx's own contribution, we compare the Direct and Reasoning groups in Table 1, where the latter routes predictions through a generated depression report rather than raw posts. The paradigm is not universally beneficial: its value hinges on the quality of the intermediate reasoning rather than the two-stage structure itself. GPT-5.4-mini, the strongest direct baseline, degrades sharply once its predictions pass through generated reports, indicating that unaligned reasoning may introduce loosely relevant or redundant content. Within the same paradigm and base model, however, AlignDx consistently surpasses the SFT and DPO variants that share the identical pipeline. The paradigm exposes an interpretable evidence bottleneck, but only alignment guarantees a clinically grounded report. This also explains why AlignDx on small base model matches or exceeds far larger models under the same paradigm—the gain comes from how reasoning is aligned, not from a paradigm-level advantage.

5.3 Ablation And Sensitivity Analysis (RQ3)

To disentangle the individual contributions of AlignDx's three key components, we ablate them individually, with results in Table 2. Removing the classifier causes the sharpest degradation across all three datasets, confirming that unconstrained report generation alone cannot yield a reliable decision signal. Replacing importance weighting with uniform weighting leads to a smaller but stable drop across all datasets, and notably its effect does not overlap with that of the reward model. This

supports our central hypothesis that supervisory signal in a clinical report is unevenly distributed: concentrating optimization on decisive tokens yields more informative supervision than treating all tokens equally. The full model achieves the best scores on all evaluation metrics and datasets, and removing any single component affects minority-class identification (F1/MCC) more than overall accuracy, demonstrating that the performance gains stem from a substantial improvement in high-risk detection capability rather than from exploiting class imbalance.

We investigate the inference-time reward guidance coefficient α, which controls the contribution of the reward model to the fused next-token distribution. Figure 3 reports the F1 and MCC of both base models for $\alpha \in [0.1, 0.6]$. For Qwen3-4B, both metrics generally increase with α, peak at $\alpha = 0.5$, and decline slightly at $\alpha = 0.6$. This consistent trend indicates that Qwen benefits from relatively strong reward guidance, whereas excessive guidance begins to constrain the base generation distribution. LLaMA-3.1-8B is substantially more sensitive to α, reaching a local optimum at $\alpha = 0.3$, while its highest MCC is obtained at $\alpha = 0.6$. These results show that the guidance coefficient should be calibrated for each backbone rather than treated as a model-independent constant.



(a) F1 (b) MCC

Figure 3: Sensitivity of Qwen3-4B and LLaMA-3.1-8B to the reward-guided decoding coefficient α in terms of (a) F1 and (b) MCC.

5.4 Case Study

We further compare the depression risk analyses of AlignDx and GPT-5.4-mini on each dataset; a representative case from eRisk is presented in Figure 4. AlignDx more consistently selects clinically relevant evidence and faithfully preserves the role of each piece of evidence in its original context. In particular, it better distinguishes the user's own disclosures from quoted news, casual expressions, artistic themes, and community-specific language, whereas GPT-5.4-mini sometimes mistakes such content for clinical evidence. Detailed examples and complete outputs are provided in Appendix E.

Diagnostic Dimension: Core Affect
"evidence": None,
"reasoning": "No posts express sustained sadness, emptiness, despair, or anhedonia. **The content is primarily news-based and factual**, with no first-person emotional reporting of low mood or loss of interest in activities."
Diagnostic Dimension: Cognitive Distortion
"evidence": None,
"reasoning": "There are no negative self-views (worthlessness), hostile views of the world, or hopelessness about the future. While some posts reference trauma (e.g., deaths, abuse), **these are reported as external events, not internal cognitive distortions**. No rumination or metacognitive loss of control is present."

Diagnostic Dimension: Core Affect
"evidence": ["'I feel like I have failed everyone' says Nigerian nurse accused of cheating after arriving in UK", "'I despair as debts mount': teachers' pension delays put lives on hold", "Simon Cowell 'feels empty' after Liam Payne's *****"],
"reasoning": "The posts contain **clear negative affect, including failure, despair, and emptiness.** While some items are news headlines rather than personal status updates, the language still reflects emotionally heavy content and sadness-related themes."
Diagnostic Dimension: Cognitive Distortion
"evidence": ["'I feel like I have failed everyone' says Nigerian nurse accused of cheating after arriving in UK", "'I despair as debts mount': teachers' pension delays put lives on hold"],
"reasoning": "**There are signs of self-blame and hopeless framing.** The text suggests negative self-evaluation and pessimistic appraisal of circumstances, though the evidence is limited and often embedded in quoted news material."

Figure 4: Sample case on eRisk. GPT-5.4-mini (right) relies on surface emotional cues, whereas AlignDx (Qwen3-4B, left) grounds its reasoning in the posts' factual nature.

In addition, we provide the further analysis of token-wise weights in Appendix F, the inference speed analysis in Appendix G, and the performance analysis of downstream classifiers in Appendix H.

6 Conclusions and Discussions

In this paper, we proposed AlignDx, a framework for depression risk detection on social media. AlignDx introduces a gradient-weighted, token-wise reward-modeling objective that concentrates preference optimization on decisive tokens rather than distributing the signal uniformly across the sequence. At inference time, a logit-fusion strategy performs reward-guided decoding via a tunable guidance strength. Extensive experiments demonstrate that AlignDx improves predictive accuracy, with gains stemming primarily from better identification of high-risk minority-class users rather than from favoring the majority class. We note several limitations: our data cover only a limited range of languages and platforms, and our study focuses solely on depression. Future work will extend this framework to broader mental health computing tasks, such as sentiment analysis, and so on.

AI USE STATEMENT

Generative AI tools were used to assist in constructing the MDPD-10k preference dataset and generating multidimensional instruction data for depression risk analysis. In addition, these tools were used to assist in drafting and language polishing of portions of the paper, with all content subject to the authors' review and revision. All AI-generated content was reviewed and verified by the authors to ensure its quality and consistency with established diagnostic criteria. Generative AI was not used to fabricate data, generate experimental results, or determine research conclusions. All experimental results reported in this work were obtained by training and evaluating our models on publicly available benchmark datasets. The authors take full responsibility for the content of this work.

ETHICS STATEMENT

This study analyzes publicly available, anonymized social media datasets (MDPD, WU3D, eRisk) originally released by their curators; we did not collect new data from human participants. We will not release any user-level raw posts or personally identifiable information. Our method is designed as a supportive early-risk screening tool, not a clinical diagnosis, and should not be used for automated decision-making about individuals; decisions with safety implications must involve qualified mental health professionals. Since depression-related expression varies across languages, platforms, and demographics, our findings may not generalize beyond the Chinese Weibo and English Reddit data evaluated here, and performance should be audited across demographic groups before real-world deployment. All authors have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have made extensive efforts to ensure the reproducibility of our work. All datasets used in our experiments (MDPD, WU3D, eRisk) are publicly available, and our complete data processing and preference dataset construction pipeline is described in Appendix A, including user filtering, clustering-based downsampling, and the dimension-wise preference pair generation procedure. Full experimental details—backbone models, LoRA configurations, hyperparameters, hardware, and software versions—are reported in Appendix C. The prompts used for multidimensional report generation and judge-based evaluation are provided verbatim in Appendix D. We report complete quantitative results in Tables 1–2 and 6–8, ablations in Table 2, and qualitative analyses with full model outputs in Appendix E. The source code and datasets used in this work have been released in an anonymous GitHub repository to facilitate reproducibility.

REFERENCES

Mohammed Al-Mosaiwi and Tom Johnstone. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542, 2018. doi: 10.1177/2167702617747074.

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, Arlington, VA, 5 edition, 2013.

Mohamad Ballout, Ulf Krumnack, Gunther Heidemann, and Kai-Uwe Kühnberger. Efficient knowledge distillation: Empowering small language models with teacher model insights. In *International Conference on Applications of Natural Language to Information Systems*, pp. 32–46. Springer, 2024.

Aaron T. Beck, A. John Rush, Brian F. Shaw, and Gary Emery. *Cognitive Therapy of Depression*. Guilford Press, New York, 1979.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324. URL https://doi.org/10.1093/biomet/39.3-4.324.

Yicheng Cai, Haizhou Wang, Huali Ye, Yanwen Jin, and Wei Gao. Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*, 217:119538, 2023.

Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: A critical review. *npj Digital Medicine*, 3(1):43, 2020. doi: 10.1038/s41746-020-0233-7. URL https://www.nature.com/articles/s41746-020-0233-7.

Chen Chen, Fenghuan Li, Haopeng Chen, and Yuankun Lin. Heterogeneous subgraph network with prompt learning for interpretable depression detection on social media. *Knowledge-Based Systems*, 315:113215, 2025.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.

Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pp. 51–60, 2014.

Fabio Crestani, David E. Losada, and Javier Parapar. Early detection of depression from social media: The erisk 2022 task. In *CLEF 2022 Working Notes*, 2022.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference*, pp. 47–56, 2013.

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

Muskan Garg, Chandni Saxena, Veena Krishnan, Ruchi Joshi, Sriparna Saha, Vijay Mago, and Bonnie J. Dorr. Cams: An annotated corpus for causal analysis of mental health issues in social media posts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6387–6396, Marseille, France, 2022. European Language Resources Association.

Zhuohan Ge, Nicole Hu, Darian Li, Yubo Wang, Shihao Qi, Yuming Xu, Han Shi, and Jason Zhang. A survey of large language models in mental health disorder detection on social media. In *2025 IEEE 41st International Conference on Data Engineering Workshops (ICDEW)*, pp. 164–176, 2025. doi: 10.1109/ICDEW67478.2025.00027.

Max Hamilton. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, 23(1):56–62, 1960. doi: 10.1136/jnnp.23.1.56.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 7184–7190, Marseille, France, 2022. European Language Resources Association.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. Args: Alignment as reward-guided search. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shgx0eqdw6.

Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613, 2001. doi: 10.1046/j.1525-1497.2001.016009606.x.

Gleb Kuzmin, Petr Strepetov, Maksim Stankevich, Natalia Chudova, Artem Shelmanov, and Ivan Smirnov. Exploring large language models for detecting mental disorders. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34523–34547, 2025.

Xiaochong Lan, Zhiguang Han, Yiming Cheng, Li Sheng, Jie Feng, Chen Gao, and Yong Li. Depression detection on social media with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 2155–2171, 2025.

Bolian Li, Yifan Wang, Ananth Grama, and Ruqi Zhang. Cascade reward sampling for efficient decoding-time alignment. In *ICML 2024 Next Generation of AI Safety Workshop*, 2024. URL https://openreview.net/forum?id=9d6PhLN4FT.

Baijiong Lin, Weisen Jiang, Yuancheng Xu, Hao Chen, and Ying-Cong Chen. Parm: Multi-objective test-time alignment via preference-aware autoregressive reward model. In *International Conference on Machine Learning*, 2025.

Zhixuan Liu, Zhanhui Zhou, Yuanfu Wang, Chao Yang, and Yu Qiao. Inference-time language model alignment via integrated value guidance. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297. University of California Press, 1967.

Mohammad Saeid Mahdavinejad, Peyman Adibi, Amirhassan Monadjemi, and Pascal Hitzler. They look like each other: Case-based reasoning for explainable depression detection on twitter using large language models, 2024.

Rei Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Nelson Elhage, David Krueger, Katherine Button, Michael Knight, Brian Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2021.

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

Javier Parapar, Anxo Perez, Xi Wang, and Fabio Crestani. erisk 2025: contextual and conversational approaches for depression challenges. In *European Conference on Information Retrieval*, pp. 416–424. Springer, 2025.

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

Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pp. 29–48. New Jersey, USA, 2003.

Zongru Shao, Xin Wang, Zhanyang Liu, Chenhan Wang, and K. P. Subbalakshmi. Systematic evaluation of machine-generated reasoning and phq-9 labeling for depression detection using large language models, 2025.

Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, Xiangnan He, Huanbo Luan, Jie Tang, Thanassis Tiropanis, Tat-Seng Chua, and Wendy Hall. Cross-domain depression detection via harvesting social media. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 1611–1617, 2018. doi: 10.24963/ijcai.2018/223.

Yiding Wang, Zhenyi Wang, Chenghao Li, Yilin Zhang, and Haizhou Wang. A multimodal feature fusion-based method for individual depression detection on sina weibo. In *2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC)*, pp. 1–8. IEEE, 2020. doi: 10.1109/IPCCC50635.2020.9391501. URL https://doi.org/10.1109/IPCCC50635.2020.9391501.

World Health Organization. *Clinical Descriptions and Diagnostic Requirements for ICD-11 Mental, Behavioural and Neurodevelopmental Disorders*. World Health Organization, Geneva, 2024. ISBN 978-92-4-007726-3. URL https://www.who.int/publications/i/item/9789240077263.

Yuancheng Xu, Udari Madhushani Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumitra Ganesh. Genarm: Reward guided generation with autoregressive reward model for test-time alignment. In *International Conference on Learning Representations*, 2025.

Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Erxue Min, and Sophia Ananiadou. Selective preference optimization via token-level reward function estimation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27876–27889. Association for Computational Linguistics, 2025.

Rui Yang, Zhe Wang, Zhi Chen, et al. Mentalllama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference*, pp. 4489–4500, 2024.

Shentao Yang, Shujian Zhang, Congying Xia, Yihao Feng, Caiming Xiong, and Mingyuan Zhou. Preference-grounded token-level guidance for language model fine-tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480–1489, 2016.

Andrew Yates, Arman Cohan, and Nazli Goharian. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2968–2978, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1322. URL https://aclanthology.org/D17-1322/.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 58348–58365. PMLR, 2024.

A Data Processing

A.1 Dataset Overview

We construct the preference corpus from three user-level social-media datasets: the Chinese Weibo Suicide Detection Dataset (SWDD), WU3D, and the English Reddit eRisk dataset. Although their source formats differ, we organize all records as user histories and retain text that can support user-level depression-risk assessment. Table 3 reports the user counts of the processed files used in this study.

Table 3: User-level statistics of the three datasets.

Dataset	Control	Depression	Total
MDPD	6,579	3,421	10,000
WU3D	10,848	21,722	32,570
eRisk	807	102	909

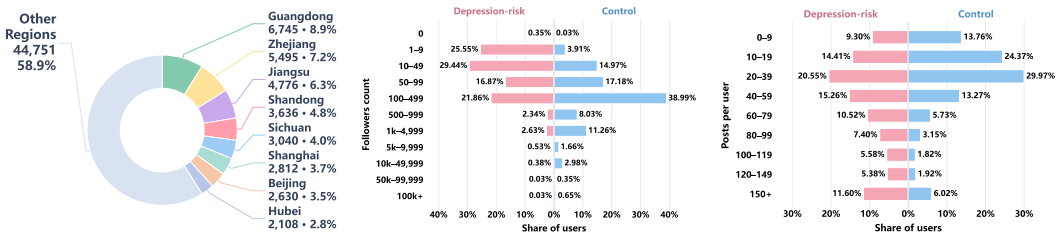


Figure 5: MDPD characteristics: geographic diversity, follower counts, and user post counts (left to right).

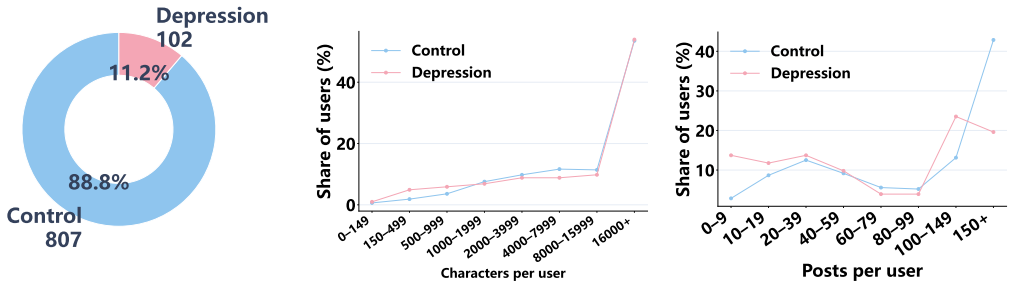


Figure 6: eRisk characteristics: label composition, character counts, and user post counts (left to right).

A.2 MDPD

We retain the 12-month posting history ending at each user's latest valid post to capture recent and stable behavioral patterns while avoiding excessive noise from outdated content. Control users with fewer than 150 characters are removed because such limited text provides insufficient evidence for reliable clinical analysis; this length constraint is not applied to depression-risk users to avoid further reducing the already limited positive samples.

For control down-sampling, we concatenate each user's posts and encode them with BGE-M3 to obtain user-level semantic representations. Long posting histories are processed using token-level chunking to preserve complete semantic information. We then cluster the user representations with KMeans and retain the users nearest the cluster centers, thereby preserving the semantic diversity of the control population while reducing its size. All depression-risk users are retained.

representative non-risk users and reduce redundancy. After cleaning and sampling, the final dataset comprises 10k Weibo users, whose statistics are summarized in Table 4.

Table 4: Statistics of the MDPD dataset.

Group	Users	Posts	Posts/User
Depression-risk users	3,421	234,094	68.43
Non-depression-risk users	6,579	296,600	45.08
Total	10,000	530,694	53.07

Figure 5 summarizes MDPD's geographic composition, follower-count distribution, and posts per user.

A.3 eRisk

We retain the 12-month history ending at each user's latest valid post and cap each history at 32K tokens. Figure 6 reports label composition, non-whitespace characters per user, and posts per user.

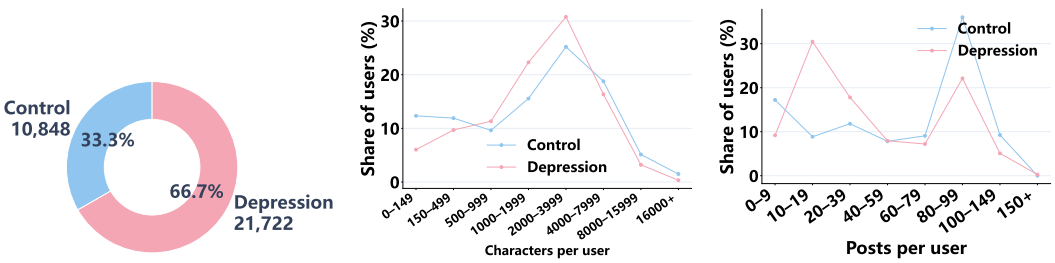


Figure 7: WU3D characteristics: label composition, character counts, and user post counts (left to right).

A.4 WU3D

We retain the 12-month history ending at each user's latest valid post and cap each history at 32K tokens. Figure 7 reports label composition, non-whitespace characters per user, and posts per user.

A.5 Multi-Dimensional Clinical Framework

Clinical practice provides mature diagnostic frameworks for depressive disorders, but their structured use in social-media depression detection remains underexplored. To ground model decisions in medical knowledge and improve the accuracy and credibility of user-level analysis, we organize observable depression-related evidence into multiple complementary dimensions derived from DSM-5 and ICD-11 diagnostic principles. Core Affect, Cognitive Distortion, and Somatic–Executive cover the affective, cognitive, and physical manifestations central to clinical assessment. Suicide Risk is separated because death- and self-harm-related expressions require explicit assessment, while Behavioral Function captures real-world impairment. Linguistic Style and External Stress complement clinical symptoms with textual markers and life-event context. Table 5 provides the detailed scope and motivation of each dimension.

For every dimension, we construct dimension-level pairwise preferences to capture quality differences across distinct aspects of the clinical analysis. The judge LLM evaluates each dimension from five perspectives: *coverage*, *accuracy*, *grounding*, *implicit-signal capture*, and *objectivity*. Each criterion is scored from 1 to 5, and their mean is used as the dimension-level quality score. This design permits two candidate reports to receive different preference directions across dimensions instead of reducing the entire clinical analysis to a single report-level judgment.

The dimensions are complementary rather than independently diagnostic. Core Affect and Cognitive Distortion characterize emotional and cognitive manifestations; Somatic–Executive and Behavioral Function reflect physical and real-world impairment; Suicide Risk isolates urgent safety signals; Linguistic Style captures indirect textual markers; and External Stress situates the evidence within the user's life context. This organization supports dimension-specific evidence grounding while reducing reliance on isolated keywords.

A.6 Preference Set Generation

Existing preference-data construction methods typically treat each candidate response as an indivisible unit and rank it using a holistic score. However, models with different capacities and inductive biases often exhibit complementary strengths across clinical dimensions . One model may better capture implicit affective signals, while another may provide more accurate evidence extraction, clinical reasoning, or risk identification. Holistic selection therefore preserves weak dimensions of a strong response while discarding valuable local analyses from other candidates. Since depression-related evidence is often multifaceted and temporally distributed, we construct preference pairs through fine-grained dimension-wise evaluation and fusion..

For each user, Qwen3.5-9B and Qwen3.5-27B independently generate structured depression-risk reports from the same posting history. Each report contains evidence and reasoning for the following multiple dimensions:

Table 5: Clinical and theoretical grounding of the multi-dimensional depression-risk analysis framework.

Dimension	Captured Characteristics	Primary Grounding	Inclusion Rationale
Core Affect	Persistent sadness, emptiness, hopelessness, irritability, and diminished interest or pleasure in previously valued activities.	DSM-5 criteria A1–A2; PHQ-9 depressed-mood and anhedonia items.	Depressed mood and anhedonia are cardinal manifestations of a major depressive episode and are often directly observable in social-media posts.
Cognitive Distortion	Negative beliefs about the self, world, and future; worthlessness or excessive guilt; hopelessness; and persistent rumination.	Beck's cognitive theory; response-styles theory; DSM-5 criterion A7.	Recurring thoughts and self-evaluations can reveal depression-related cognitive patterns even when explicit emotion words are absent.
Somatic–Executive	Sleep and circadian disruption, appetite or weight change, fatigue, reduced energy, impaired concentration, indecisiveness, and difficulty completing daily tasks.	DSM-5 criteria A3, A4, A6, and A8; PHQ-9; HAM-D.	Physical and executive symptoms provide complementary evidence of persistence and functional burden across a user's posting history.
Suicide Risk	Recurrent thoughts of death, self-harm expressions, suicidal ideation, intent, planning, or preparation, together with severe worthlessness when relevant.	DSM-5 criterion A9; established clinical suicide-risk assessment principles.	These signals have immediate safety implications and should not be diluted within a general affective score.
Linguistic Style	Frequent absolutist terms such as always and never, elevated first-person singular pronouns, and sustained self-focused language.	Psycholinguistic findings on absolutist language and self-focus in depression.	Linguistic form can reveal rigid thinking and excessive self-focus when users communicate distress indirectly or avoid diagnostic vocabulary.
Behavioral Function	Social withdrawal, inactivity, impaired study, work, self-care, or social roles, and maladaptive coping through alcohol, substances, or avoidance.	Behavioral activation theory; WHODAS 2.0 functional domains.	Functional impairment helps distinguish sustained clinical risk from transient negative mood and connects distress to real-world consequences.
External Stress	Bereavement, relationship breakdown, unemployment, financial strain, interpersonal conflict, and co-occurring anxiety, agitation, panic, or somatic tension.	DSM-5 psychosocial problem codes and anxious-distress specifier; HAM-D anxiety factors.	Context clarifies why distress emerges, whether it persists across events, and whether clinically relevant co-occurring symptoms are present.

1. core affective symptoms and anhedonia;
2. cognitive patterns and distortions;
3. somatic symptoms and executive function;
4. severe risk;
5. social-media linguistic features;
6. behavioral coping and social functioning;
7. external stressors and comorbidity.

Each dimension contains relevant evidence extracted from the user's posting history and clinical reasoning derived from that evidence. This structured representation allows the analysis blocks generated by different models for the same clinical dimension to be evaluated independently.

We employ Qwen3.5-Plus as the judge LLM. The judge receives both the original posts and the complete candidate report, allowing it to assess the report against the source text rather than relying

on surface fluency alone. For each clinical dimension, the judge assigns five subscores on a five-point scale:

- **Coverage**: whether the report captures the relevant evidence that is actually present in the user's posts;
- **Accuracy**: whether the cited content is exact and relevant, without fabricated quotations, misattribution, or unsupported overstatement;
- **Grounding**: whether the clinical reasoning is coherent and faithfully supported by the original posts;
- **Implicit capture**: whether the report recognizes valid signals that are distributed across posts or only become apparent through cross-post integration;
- **Objectivity**: whether the report avoids both overreacting to literal wording and being excessively conservative toward weak but valid evidence.

For each model-generated report, we compute the mean of five subscores as the quality score of each clinical dimension, and average the multi-dimension scores to obtain an overall report quality score. The overall score measures report-level quality and is used for source-model tie-breaking when necessary. The primary preference construction is performed independently at the dimension level. For each user and clinical dimension, we compare the corresponding analysis blocks generated by Qwen3.5-9B and Qwen3.5-27B. The block with the higher dimension score is selected as the preferred block, while the lower-scoring block is selected as the rejected block. Each block retains its source model and dimension score for traceability. When two blocks receive identical scores, the Qwen3.5-27B block is selected as preferred and the Qwen3.5-9B block as rejected to ensure deterministic and reproducible construction.

After dimension-wise selection, we concatenate the preferred blocks and rejected blocks in the predefined clinical order to form the user-level preferred and rejected reports, respectively. Therefore, the final reports do not necessarily originate from a single model response but may combine complementary strengths from both models. For report-level predictions and clinical summaries, we apply a majority-source rule: each report adopts these fields from the model contributing more corresponding blocks. If both models contribute the same number of blocks, we compare their overall quality scores and select the higher-quality model. If the tie remains, Qwen3.5-27B is used as the source for the preferred report and Qwen3.5-9B for the rejected report. This rule only determines report-level fields and does not modify the selected dimension blocks.

The final training instance for each user consists of the original posting history, the preferred report, and the rejected report, forming the preference dataset for subsequent optimization. Compared with holistic candidate selection, this strategy avoids discarding useful information due to weaknesses in only a subset of dimensions and prevents locally inferior analyses from being retained solely because of a high overall score. The resulting preference pairs provide fine-grained, source-grounded, and clinically structured supervision across multiple facets of depression-risk analysis.

B Theoretical Analysis

Let the context at position t be $h_t = (x, y_{<t})$, and let the vocabulary logits be $\mathbf{z}_t \in \mathbf{R}^{|V|}$. The token probability and log-likelihood are defined as

$$\begin{aligned} \pi_\theta(v \mid h_t) &= \frac{e^{z_{t,v}}}{\sum_u e^{z_{t,u}}}, \\ \ell_t(\theta) = \log \pi_\theta(y_t \mid h_t) &= z_{t,y_t} - \log \sum_u e^{z_{t,u}}. \end{aligned} \tag{14}$$

Taking the derivative with respect to any vocabulary position v gives the standard softmax gradient:

$$\frac{\partial \ell_t}{\partial z_{t,v}} = \mathbf{1}[v = y_t] - \pi_\theta(v \mid h_t). \tag{15}$$

For the reward model loss

$$\mathcal{L} = -\mathbf{E}\left[\log \sigma(\beta\Delta)\right], \tag{16}$$

where

$$\Delta = S_\theta(x, y_w) - S_\theta(x, y_l). \tag{17}$$

The sensitivity of the loss with respect to the weighted scores is

$$\frac{\partial \mathcal{L}}{\partial S_w} = -\beta\sigma(-\Delta), \frac{\partial \mathcal{L}}{\partial S_l} = +\beta\sigma(-\Delta). \tag{18}$$

Since

$$S_w = \sum_t \omega_t^w \ell_t^w, S_l = \sum_t \omega_t^l \ell_t^l, \tag{19}$$

the gradient of the loss with respect to an individual log-probability becomes

$$\frac{\partial \mathcal{L}}{\partial \ell_t^w} = -\beta\sigma(-\Delta)\,\omega_t^w, \frac{\partial \mathcal{L}}{\partial \ell_t^l} = +\beta\sigma(-\Delta)\,\omega_t^l. \tag{20}$$

Applying the chain rule to the logits gives

$$\frac{\partial \mathcal{L}}{\partial z_{t,v}^w} = -\beta\sigma(-\Delta)\,\omega_t^w\left(\mathbf{1}[v = y_t^w] - \pi_\theta(v \mid h_t)\right), \tag{21}$$

$$\frac{\partial \mathcal{L}}{\partial z_{t,v}^l} = +\beta\sigma(-\Delta)\,\omega_t^l\left(\mathbf{1}[v = y_t^l] - \pi_\theta(v \mid h_t)\right). \tag{22}$$

Taking the chosen response as an example, substituting $v = y_t^w$ and $v \neq y_t^w$, respectively, yields

$$\frac{\partial \mathcal{L}}{\partial z_{t,v}^w} = \begin{cases} -\beta\sigma(-\Delta)\omega_t^w(1 - \pi_\theta(y_t^w \mid h_t)), & v = y_t^w, \\ +\beta\sigma(-\Delta)\omega_t^w\pi_\theta(v \mid h_t), & v \neq y_t^w. \end{cases} \tag{23}$$

The gradient of the target token y_t^l in the rejected response is positive, while the gradient of non-target tokens is negative. Therefore, the optimization direction for the rejected response is opposite: the target token probability is suppressed, whereas non-target token probabilities are increased.

All gradient terms contain the common factor $\beta\sigma(-\Delta)\omega_t$. Therefore, within the same response, the magnitude of the logit update at position t is strictly proportional to ω_t. A high ω_t makes each optimization step sharpen the softmax peak of the target token and further suppress competing tokens; a low ω_t produces only a negligible update, leaving the distribution at that position almost unchanged from the base model initialization.

For a fixed response, the ℓ_2 norm of the logit gradient at position t satisfies

$$\left\|\nabla_{z_t}\mathcal{L}\right\|_2 = \beta\sigma(-\Delta)\,\omega_t \cdot \underbrace{\left\|e_{y_t} - \pi_\theta(\cdot \mid h_t)\right\|_2}_{\leq\sqrt{2}}, \tag{24}$$

where e_{y_t} denotes the one-hot vector of the target token. Since $\beta\sigma(-\Delta)$ is identical for all positions t within the same sample, we have

$$\frac{\|\nabla_{z_t}\mathcal{L}\|}{\|\nabla_{z_{t'}}\mathcal{L}\|} = \frac{\omega_t}{\omega_{t'}} \cdot \frac{\|e_{y_t} - \pi_\theta(\cdot \mid h_t)\|}{\|e_{y_{t'}} - \pi_\theta(\cdot \mid h_{t'})\|} \tag{25}$$

Thus, when the softmax residuals are comparable, the update magnitude is strictly proportional to ω_t.

According to Eq. 8,

$$g_i = \sum_{\ell=0}^{L} \left\| \frac{\partial s}{\partial h_i^{(\ell)}} \right\|_1, s = \max_j o_{N,j}. \tag{26}$$

Let a perturbation δh_i be applied to the hidden representation of the i-th token. Using the first-order Taylor expansion, the change of the final prediction surrogate is

$$\Delta s \approx \sum_{\ell} \left\langle \frac{\partial s}{\partial h_i^{(\ell)}}, \delta h_i^{(\ell)} \right\rangle, \tag{27}$$

and therefore

$$|\Delta s| \leq \sum_{\ell} \left\| \frac{\partial s}{\partial h_i^{(\ell)}} \right\|_1 \cdot \|\delta h_i^{(\ell)}\|_\infty. \tag{28}$$

Under perturbations satisfying $\|\delta h_i^{(\ell)}\|_\infty \leq \epsilon$, the upper bound of the maximum possible prediction shift is exactly $\epsilon\, g_i$. Therefore, g_i serves as the tight upper-bound coefficient of the maximum prediction change induced by a unit perturbation, i.e., a influence measure of token i.

Comparing the reward model π_{rm} with the frozen base model π_{base}, we can write:

$$\log \pi_{\mathrm{rm}}(y_t \mid x, y_{<t}) = \log \pi_{\text{base}}(y_t \mid x, y_{<t}) + \delta_t(x, y_{\leq t}). \tag{29}$$

Due to the gradient structure described above, after training, the magnitude of δ_t is positively correlated with ω_t (tokens with high ω_t receive larger parameter updates), i.e.,

$$|\delta_t| \propto \omega_t. \tag{30}$$

In other words, π_{rm} does not deviate from π_{base} uniformly at all positions, but only produces corrections at those tokens that influence the final diagnostic prediction.

The logit-space fusion at inference time is:

$$\begin{aligned} \log \pi(y_t \mid x, y_{<t}) = & \frac{1}{1+\alpha} \log \pi_{\text{base}}(y_t \mid x, y_{<t}) \\ & + \frac{\alpha}{1+\alpha} \log \pi_{\mathrm{rm}}(y_t \mid x, y_{<t}) + \text{const}. \end{aligned} \tag{31}$$

Substituting the previous approximation, we obtain:

$$\begin{aligned} \log \pi(y_t \mid x, y_{<t}) \approx & \log \pi_{\text{base}}(y_t \mid x, y_{<t}) \\ & + \frac{\alpha}{1+\alpha} \cdot \omega_t \cdot \phi(x, y_{\leq t}) + \text{const}, \end{aligned} \tag{32}$$

where $\phi(\cdot)$ is a potential function related to the preference direction (learned from preference data).

Summing over the entire response y, we obtain the sequence-level distribution ratio:

$$\log \frac{\pi(y \mid x)}{\pi_{\text{base}}(y \mid x)} \approx \frac{\alpha}{1+\alpha} \sum_t \omega_t \phi(x, y_{\leq t}). \tag{33}$$

This shows that AlignDx's final distribution is equivalent to a selective exponential reweighting of the base model distribution along the direction of gradient-importance weights.

C EXPERIMENTAL SETTINGS

All experiments are conducted on a single NVIDIA RTX PRO 6000 GPU using PyTorch 2.10.0+cu128 and OpenRLHF[1]. We evaluate SFT, DPO, and AlignDx using both LLaMA-3.1-8B-Instruct and Qwen3-4B-Instruct-2507. For all experiments, the maximum context length is uniformly set to 16,384. All models are trained for one epoch using bfloat16 precision, DeepSpeed ZeRO Stage 2.

For SFT, both models use a batch size of 32, a learning rate of 1×10^{-5}, and a LoRA scaling factor of 64. The LoRA ranks are set to 8 and 64 for LLaMA-3.1-8B-Instruct and Qwen3-4B-Instruct-2507, respectively.

For DPO, the two backbone models use the same hyperparameter settings. Specifically, the batch size is 32, the learning rate is 1×10^{-5}, and both the LoRA rank and scaling factor are set to 64. The DPO coefficient is set to $\beta = 0.1$.

For AlignDx, we adopt the same hyperparameter settings for LLaMA-3.1-8B-Instruct and Qwen3-4B-Instruct-2507. The LoRA rank, scaling factor, and dropout rate are set to 8, 16, and 0.05, respectively. We use a batch size of 16 and a learning rate of 5×10^{-5}. The coefficient β in the reward model objective is set to 0.05.

In the inference phase, reward-guided decoding is implemented with Sglang by loading the backbone model and the AlignDx reward adapter as the base and LoRA branches, respectively. The inference server uses the Triton attention backend and the PyTorch sampling backend. The context length is set to 32,768. We allocate 90% of the GPU memory to static memory management, allow at most 10 concurrent requests. Generation is performed using 10 parallel HTTP workers, with a temperature of 0.6, a repetition penalty of 1.15, and a maximum generation length of 4,096 tokens. The reward-guidance coefficient α is set to 0.5.

D PROMPT INSTRUCTION

D.1 MULTI-DIMENSIONAL ANALYSIS PROMPT

The following prompt is used to instruct the LLM to analyze a user's historical social-media posts under the multi-dimensional clinical framework and generate a structured depression-risk report:

10

 # Role and Objective
You are an expert clinical psychologist and NLP computational linguist. Analyze a user's historical social−media posts and produce one structured depression−risk screening report based on a multi−dimensional clinical framework.

 # Evaluation Dimensions
1. Core Affective and Anhedonic: sadness, emptiness, despair, irritability, or loss of interest or pleasure.
2. Cognitive Pattern Distortions: negative views of the self, world, and future; rumination; and loss of cognitive control.
3. Somatic Symptoms and Executive Function: sleep disturbance, fatigue, impaired concentration or decision−making, and marked appetite or weight changes.
4. Severe Risk: death−related thoughts, profound worthlessness, self−harm tendencies, or suicidal ideation or plans.
5. Social−Media Linguistic Features: absolutist language and elevated first−person singular pronouns.
6. Behavioral Coping and Social Function: destructive coping, substance or alcohol misuse, or social withdrawal.
7. External Stressors and Comorbidity: bereavement, breakup, job loss, severe anxiety, or panic.

 # Analysis Rules
1. Evidence−based: a present dimension must be supported by one or two exact quotes from the user's posts.
2. Do not invent evidence, alter quotations, or force unsupported clinical interpretations.

[1] https://github.com/OpenRLHF/OpenRLHF

3. Recover implicit signals only when they are clearly supported by the complete post history.
4. Analyze all seven dimensions first and derive is_depressed only from the combined dimension−level findings.
5. The label indicates depression risk rather than a clinical diagnosis. Prioritize recall when concrete, quotable risk evidence is present, but reject jokes, quotations, generic venting, or isolated temporary stress without self−applicable evidence.
6. Keep exact quotations only in each evidence array. Use brief integrative English prose in reasoning and clinical_summary.
7. Return one valid JSON object with no surrounding text.

 # Input Posts
<USER_POSTS>

 # Output Format

```
{
  "dimensions_analysis": {
    "dim_1_core_affective": {
      "evidence": [], "reasoning": ""
    },
    "dim_2_cognitive_pattern": {
      "evidence": [], "reasoning": ""
    },
    "dim_3_somatic_executive": {
      "evidence": [], "reasoning": ""
    },
    "dim_4_severe_risk": {
      "evidence": [], "reasoning": ""
    },
    "dim_5_linguistic_features": {
      "evidence": [], "reasoning": ""
    },
    "dim_6_behavioral_coping": {
      "evidence": [], "reasoning": ""
    },
    "dim_7_stressors_comorbidity": {
      "evidence": [], "reasoning": ""
    }
  },
  "overall_assessment": {
    "is_depressed": true,
    "clinical_summary": ""
  }
}
```

The overall assessment item is reserved for directly prompting the LLM to produce diagnostic results when no downstream classifier is employed; when a classifier is used for evaluation, these tokens are excluded from it.

D.2 Report Judge Prompt

The following prompt is used to instruct the judge model to evaluate a candidate report against the user's original posts:

10

 # Role and Objective
You are a senior clinical psychology judge. Score one candidate depression−risk report against the original user's posts.

 # Evaluation Dimensions
Score the report on the same seven dimensions:
1. Core Affective and Anhedonic

2. Cognitive Pattern Distortions
3. Somatic Symptoms and Executive Function
4. Severe Risk
5. Social−Media Linguistic Features
6. Behavioral Coping and Social Function
7. External Stressors and Comorbidity

 # Scoring Goal
For each dimension, assign a score from 1 to 5 for:
− coverage: captures supported evidence in the posts;
− accuracy: uses exact, relevant, non−fabricated quotes;
− grounding: provides clinically coherent and faithful reasoning;
− implicit_capture: recognizes supported distributed or cross−post evidence;
− objectivity: avoids both literal triggering and excessive conservatism.

Compute each dimension score as the rounded average of its five subscores. A score of 5 is excellent and a score of 1 is poor.

 # Judge Requirements
1. Judge against the original posts, not the report alone.
2. Reward supported explicit, implicit, cumulative, and low−intensity repeated signals.
3. Penalize shallow or overly literal analysis that misses indirect depression evidence.
4. Penalize invented evidence, inaccurate quotes, and unsupported overstatement.
5. A correct final label can still receive a low score when the dimension analysis is incomplete or biased.
6. Do not require an explicit statement such as 'I am depressed.' Keep scores moderate or low when evidence is sparse, isolated, joking−only, or better explained by temporary stress.

 # Input Posts
<USER_POSTS>

 # Candidate Report JSON
<CANDIDATE_REPORT_JSON>

 # Output Format
Return only valid JSON:

```
{
  "dimension_scores": {
    "dim_1_score": 0,
    "dim_2_score": 0,
    "dim_3_score": 0,
    "dim_4_score": 0,
    "dim_5_score": 0,
    "dim_6_score": 0,
    "dim_7_score": 0
  },
  "dimension_subscores": {
    "dim_1": {"coverage": 0, "accuracy": 0, "grounding": 0, "implicit_capture": 0, "objectivity": 0},
    "dim_2": {"coverage": 0, "accuracy": 0, "grounding": 0, "implicit_capture": 0, "objectivity": 0},
    "dim_3": {"coverage": 0, "accuracy": 0, "grounding": 0, "implicit_capture": 0, "objectivity": 0},
    "dim_4": {"coverage": 0, "accuracy": 0, "grounding": 0, "implicit_capture": 0, "objectivity": 0},
    "dim_5": {"coverage": 0, "accuracy": 0, "grounding": 0, "implicit_capture": 0, "objectivity": 0},
    "dim_6": {"coverage": 0, "accuracy": 0, "grounding": 0, "implicit_capture": 0, "objectivity": 0},
    "dim_7": {"coverage": 0, "accuracy": 0, "grounding": 0, "implicit_capture": 0, "objectivity": 0}
  },
  "total_score": 0,
  "bias_assessment": {
    "relatively_unbiased": false,
    "target_biased": false,
    "explicit_expression_bias": false,
    "implicit_signal_miss": false,
    "surface_trigger_reasoning": false,
    "evidence_coverage_sufficient": false,
    "evidence_accuracy_sufficient": false,
```

```
    "reasoning_correctness_sufficient": false
  },
  "judge_reasoning": ""
}
```


E CASE STUDY

We compare AlignDx and GPT-5.4-mini on three representative multi-dimensional analyses, focusing on the generated evidence and reasoning.

E.1 MDPD: CAPTURING IMPLICIT FATIGUE ACROSS POSTS

Representative posts include:

> "Sometimes fatigue arrives silently; perhaps even the courage to open up is barely there."
>
> "Fatigue and decay erode me from within, little by little, and I cannot resist them."
>
> "It really seems to be getting harder and harder to feel happy."
>
> "When you lose interest in life, the only thing left is the responsibility that comes with being in this world."

AlignDx identifies the repeated fatigue expressions as Somatic–Executive evidence and links them to the user's declining psychological energy across posts. GPT-5.4-mini leaves this dimension unsupported because the history contains no explicit sleep, appetite, weight, concentration, or decision-making complaint. The comparison shows AlignDx's ability to recover a distributed, low-intensity fatigue signal that is conveyed through repeated descriptions of inner depletion rather than a standardized symptom statement.

E.2 ERISK: HANDLING QUOTED EVIDENCE

The user primarily shares news headlines containing emotionally salient first-person clauses:

> "I feel like I have failed everyone' says Nigerian nurse accused of cheating after arriving in UK"
>
> "I despair as debts mount: teachers' pension delays put lives on hold"
>
> "Simon Cowell feels empty after Liam Payne's death"

For Core Affect, GPT-5.4-mini cites "I feel like I have failed everyone," "I despair as debts mount," and "feels empty" as negative-affect evidence. It reuses the first two headlines for Cognitive Pattern, Linguistic Features, and External Stress. In each instance, however, the first-person clause belongs to a nurse, teachers awaiting pensions, or Simon Cowell rather than to the account owner. AlignDx preserves this speaker attribution across dimensions: it excludes third-party emotions and stressors from the user's clinical evidence and distinguishes quoted journalistic language from the user's own linguistic style. This prevents emotionally salient headline wording from being misinterpreted as personal depressive expression.

E.3 WU3D: ARTISTIC THEMES VERSUS AUTOBIOGRAPHICAL SYMPTOMS

The history contains negative vocabulary in an artwork description and a later post about its completion:

> "Meaningless Reallife Numb Booooom Ots Sins72 Oasis"
>
> "When we are young, we easily fall into anxiety about emptiness."
>
> "I hope that, while experiencing this work, those tired of meaning itself can set aside its pressure and indulge themselves for once."
>
> "The animated music video I spent two years drawing is finally complete. Doing it alone was difficult, but I am also very happy."

For Core Affect, GPT-5.4-mini cites “Meaningless,” “Numb,” and the sentence about anxiety over emptiness, interpreting their repetition as a negative affective pattern. AlignDx instead preserves the artwork context and avoids converting thematic vocabulary into autobiographical symptoms. For Cognitive Pattern, it retains the longer existential passage but correctly characterizes its tone as reflective rather than hopeless or self-critical. Under External Stress, AlignDx further integrates the later report of completion, happiness, and continued creative participation, preventing the difficulty of a two-year creative process from being overstated as a personal clinical stressor.

E.4 Model Responses

Complete evidence, reasoning, and overall assessments are reported for all three cases.

E.4.1 MDPD

AlignDx complete response.


```
{
  "dimensions_analysis": {
    "dim_1_core_affective": {
      "evidence": [
        "At this point, what room is left to undo anything? Things remain but people change; fickleness is human nature.",
        "When things are bad, they only get worse.",
        "Then I discovered that perhaps the worst feeling in this world is being powerless and having nowhere to vent. I seem to be lying low and waiting, feeling an instant of despair after all the ups and downs. Then I will personally end all of this, like something withering before it ever bloomed.",
        "This sudden sadness is tragic beyond words.",
        "It really seems to be getting harder and harder to feel happy.",
        "Every second falls silent, losing the warmth it should have had, until it is no longer alive.",
        "When you lose interest in life, the only thing left is the responsibility that comes with being in this world."
      ],
      "reasoning": "The user repeatedly expresses profound emotional depletion, emptiness, and a sense of despair. Phrases like 'no longer vibrant', 'despair', 'no interest in life', and 'a sense of nothingness' indicate sustained low mood and anhedonia. These are consistent with core affective symptoms of depression, particularly the feeling of inner emptiness and loss of pleasure."
    },
    "dim_2_cognitive_pattern": {
      "evidence": [
        "Whether one is a fool or fools oneself, it is all a joke. I always have to do something to prove my existence and struggle over things that are hardly necessary. I may look indestructible, yet I always break at the slightest touch.",
        "I need to be understood, cared for, and loved. Because I am too reserved, I do not know how to express or ask for it.",
        "Even I cannot understand myself. With this inborn sorrow, I am ready at any moment for the end and to move forward without fear.",
        "I never thought I feared pain; it is the humiliating predicament that can deprive me of everything.",
        "The source of longing for all beautiful things is evil.",
        "The isolated individual develops a suffocating silence and gradually becomes numb.",
        "I know you are in my heart, and I know we are not so well suited. Yet I do not know when every word and action of yours began to stir my feelings. I will not tell you, and you may never guess. We remain at the edge and may never reach the other shore; intruding into this inner world has only trapped me in a cocoon of my own making."
      ],
      "reasoning": "The user exhibits pervasive negative self−view ('self−deception', 'inability to understand oneself'), hostile views of relationships ('disgust at emotional intimacy'), and hopelessness about connection. There is a clear pattern of rumination on emotional fragility, self−deception, and existential isolation, indicating cognitive distortions consistent with depressive thinking."
    },
    "dim_3_somatic_executive": {
      "evidence": [
        "Sometimes fatigue arrives silently; perhaps even the courage to open up is barely there.",
        "I keep hoping and imagining, but I understand reality and reason even more.",
```

```
      "Fatigue and decay erode me from within, little by little; I cannot resist them, yet I still cannot bring myself to an end.",
      "I try to live life for myself and minimize sadness."
    ],
    "reasoning": "The user describes persistent fatigue, emotional exhaustion, and a decline in executive functioning such as decision-making and emotional regulation. The phrase 'fatigue from within' and 'unable to self-terminate' suggest a marked loss of energy and cognitive control, which aligns with executive dysfunction commonly seen in depressive states."
  },
  "dim_4_severe_risk": {
    "evidence": [
      "Go die. Get as far away from me as possible and never appear before me again in this life, never.",
      "Then I will personally end all of this, like something withering before it ever bloomed.",
      "If I really am terrible, do not let me remain and harm the world. If I am not too terrible, give me enough luck to repay everything.",
      "The end happens in an instant, but it took far too long to make the decision."
    ],
    "reasoning": "The user explicitly expresses suicidal ideation, including a direct command to 'go to death' and a desire to end life due to perceived personal worthlessness. Phrases like 'end everything' and 'never appear again' indicate serious self-harm tendencies and death-related thoughts, meeting the threshold for severe risk."
  },
  "dim_5_linguistic_features": {
    "evidence": [
      "Always, always being, always struggling, always getting lost, always hoping, always imagining.",
      "Never, it has always been, it has never been, I can never do it.",
      "Completely, thoroughly, absolutely, indestructible, yet always breaking at the slightest touch.",
      "Forever, never again, never exists.",
      "Only, only what matters deserves understanding.",
      "If, if only, if I could choose, if I really am terrible."
    ],
    "reasoning": "The posts contain a high frequency of absolutist language ('always', 'never', 'only', 'if only'), first-person singular pronouns ('I', 'my', 'mine'), and emotionally charged, self-referential statements. This linguistic pattern indicates excessive self-focus and cognitive rigidity, typical of depressive rumination and self-criticism."
  },
  "dim_6_behavioral_coping": {
    "evidence": [
      "I need to be understood, cared for, and loved.",
      "Loneliness is the norm.",
      "I try to live life for myself and minimize sadness.",
      "I am only walking the road beneath my feet and experiencing the road I must travel. The harder I work, the stronger and more resilient I become.",
      "I give up. This is really terrible."
    ],
    "reasoning": "The user demonstrates significant social withdrawal and emotional avoidance, expressing a need for connection but being unable to initiate or sustain it. The repeated use of 'isolation', 'no understanding', and 'trying to reduce sadness' suggests maladaptive coping strategies. There is no evidence of active engagement or recovery behaviors, indicating functional stagnation."
  },
  "dim_7_stressors_comorbidity": {
    "evidence": [
      "Things remain but people change; fickleness is human nature.",
      "Social interaction is my talent and ability is my confidence, yet emotionally I am a complete failure.",
      "If you are not strong enough, what right do you have to blame others for not choosing you?",
      "If it could be healed easily, why would there be so many problems?",
      "If it is something for which there is no if."
    ],
    "reasoning": "The user references relational failures (emotional breakdown in relationships), personal inadequacy, and perceived emotional betrayal. These are concrete stressors that contribute to deep distress. However, the distress is not framed as acute panic or a single event but as a chronic, internalized burden, suggesting comorbid anxiety and depression rather than isolated trauma."
```

```
    }
  },
  "overall_assessment": {
    "is_depressed": true,
    "clinical_summary": "The user exhibits sustained core affective distress, cognitive distortions, and
executive dysfunction across multiple dimensions. Key evidence includes explicit suicidal ideation,
pervasive self−negation, and a pattern of emotional exhaustion. Linguistic features show high absolutism
 and self−focus, while behavioral patterns reflect social withdrawal and maladaptive coping. These
elements collectively form a valid early depression−risk chain, meeting the criteria for subclinical to
moderate depressive tendency with severe risk indicators."
  }
}
```


GPT-5.4-mini complete response.


```
{
  "dimensions_analysis": {
    "dim_1_core_affective": {
      "evidence": [
        "When things are bad, they only get worse.",
        "It really seems to be getting harder and harder to feel happy.",
        "This sudden sadness is tragic beyond words.",
        "When you lose interest in life, the only thing left is the responsibility that comes with being in this
world.",
        "I am not trapped by anyone else; I have simply never let myself go, unwilling to give up yet always
forcing myself to choose."
      ],
      "reasoning": "The posts show repeated low mood, sadness, loss of enjoyment, and an ongoing sense
of emotional depletion rather than a single passing complaint."
    },
    "dim_2_cognitive_pattern": {
      "evidence": [
        "At this point, what room is left to undo anything? Things remain but people change; fickleness is
human nature.",
        "Then I discovered that perhaps the worst feeling in this world is being powerless and having
nowhere to vent.",
        "Even I cannot understand myself. With this inborn sorrow, I am ready at any moment for the end
and to move forward without fear.",
        "An internal struggle with myself.",
        "I despise my recent behavior.",
        "Always searching for the destination while lost on the road, yet still unable to reach it."
      ],
      "reasoning": "There is a clear pattern of hopelessness, pessimism about relationships and the world,
 self−criticism, and cognitive entrapment with little sense of resolution."
    },
    "dim_3_somatic_executive": {
      "evidence": [],
      "reasoning": "The posts do not provide direct, quotable evidence of sleep disturbance, appetite or
weight change, or specific concentration and decision−making impairment."
    },
    "dim_4_severe_risk": {
      "evidence": [
        "Go die. Get as far away from me as possible and never appear before me again in this life, never.",
        "Then I will personally end all of this, like something withering before it ever bloomed.",
        "I am ready at any moment for the end and to move forward without fear.",
        "Why was I brought into this world?",
        "If I really am terrible, do not let me remain and harm the world. If I am not too terrible, give me
enough luck to repay everything.",
        "The end happens in an instant, but it took far too long to make the decision."
      ],
      "reasoning": "The text contains explicit death−related language, 'ending' references, and self−
invalidating statements that raise concern for severe risk even though no detailed plan is stated."
    },
```

```
    "dim_5_linguistic_features": {
      "evidence": [
        "It is getting harder and harder to feel happy.",
        "I have never let myself go.",
        "Always searching for the destination while lost on the road, yet still unable to reach it.",
        "Never appear again."
      ],
      "reasoning": "The language includes absolutist and totalizing phrasing, with repeated self-focused
narration and broad negative generalizations that are consistent with depressive linguistic style."
    },
    "dim_6_behavioral_coping": {
      "evidence": [
        "The isolated individual develops a suffocating silence and gradually becomes numb.",
        "On countless nights I approach collapse and suffocation, then awaken each morning and repeat
the cycle.",
        "It really seems to be getting harder and harder to feel happy.",
        "Fine, perhaps I am not suited to liking anyone either. I give up. This is really terrible."
      ],
      "reasoning": "There are signs of withdrawal, emotional shutdown, repeated nighttime collapse-
recovery cycles, and relational disengagement, suggesting impaired coping and social functioning."
    },
    "dim_7_stressors_comorbidity": {
      "evidence": [
        "If you had not kept trying to enter my heart, this would not have happened.",
        "Every word and action of yours constantly stirs the place where my feelings rest.",
        "If I could choose, I certainly would not choose to fall in love with you.",
        "No matter how I turn, I cannot escape the knot tied to you.",
        "Let my happiness remain only in thought. Let everything arrive with the wind and leave with it. I
never possessed it, yet lost it hundreds of times. This bond ends here."
      ],
      "reasoning": "Interpersonal stress and romantic distress are prominent and appear emotionally
intense, though panic symptoms or other comorbid syndromes are not directly described."
    }
  },
  "overall_assessment": {
    "is_depressed": true,
    "clinical_summary": "The post history supports a depressive-risk signal through sustained low mood
and loss of pleasure in dim 1, pessimistic and self-critical thinking in dim 2, explicit death-related and
self-negating language in dim 4, plus withdrawal and emotional shutdown in dim 6. Interpersonal stress
in dim 7 further reinforces a pattern of persistent inner depletion."
  }
}
```


E.4.2 eRisk

AlignDx complete response.


```
{
  "dimensions_analysis": {
    "dim_1_core_affective": {
      "evidence": [],
      "reasoning": "No posts express sustained sadness, emptiness, despair, or anhedonia. The content is
primarily news-based and factual, with no first-person emotional reporting of low mood or loss of
interest in activities."
    },
    "dim_2_cognitive_pattern": {
      "evidence": [],
      "reasoning": "There are no negative self-views (worthlessness), hostile views of the world, or
hopelessness about the future. While some posts reference trauma (e.g., deaths, abuse), these are
reported as external events, not internal cognitive distortions. No rumination or metacognitive loss of
control is present."
    },
    "dim_3_somatic_executive": {
```

```
    "evidence": [],
    "reasoning": "No posts mention sleep disturbance, severe fatigue, concentration decline, or appetite/
weight changes. All reported symptoms are either external (e.g., 'storm Ashley batters Britain') or related
 to physical incidents, not self-reported somatic or executive dysfunction."
   },
   "dim_4_severe_risk": {
    "evidence": [],
    "reasoning": "There are no explicit references to suicidal ideation, self-harm, death-related
thoughts, or profound worthlessness. While several posts mention deaths (e.g., in Channel crossings, care
 homes, or shootings), these are contextualized as news events, not personal distress or risk of harm."
   },
   "dim_5_linguistic_features": {
    "evidence": [],
    "reasoning": "No posts contain high-frequency absolutist language (e.g., 'always', 'never', '
completely') or elevated first-person singular pronouns indicating excessive self-focus. The language is
neutral, journalistic, and third-person oriented, with no self-referential emotional language."
   },
   "dim_6_behavioral_coping": {
    "evidence": [],
    "reasoning": "No posts indicate destructive compensatory behaviors (e.g., substance misuse, alcohol
 abuse) or social withdrawal (e.g., inability to maintain daily life, staying in bed). All behaviors reported
are either public, institutional, or tied to external events, not personal coping mechanisms."
   },
   "dim_7_stressors_comorbidity": {
    "evidence": [],
    "reasoning": "While several posts describe negative life events (e.g., deaths, police shootings, care
home tragedies), these are presented as news items, not internalized stressors with severe anxious distress
 or panic. No personal stressor is linked to emotional breakdown or psychological distress."
   }
  }
}
```


GPT-5.4-mini complete response.


```
{
  "dimensions_analysis": {
   "dim_1_core_affective": {
    "evidence": [
     "I feel like I have failed everyone' says Nigerian nurse accused of cheating after arriving in UK",
     "'I despair as debts mount': teachers' pension delays put lives on hold",
     "Simon Cowell 'feels empty' after Liam Payne's *****"
    ],
    "reasoning": "The posts contain clear negative affect, including failure, despair, and emptiness.
While some items are news headlines rather than personal status updates, the language still reflects
emotionally heavy content and sadness-related themes."
   },
   "dim_2_cognitive_pattern": {
    "evidence": [
     "'I feel like I have failed everyone' says Nigerian nurse accused of cheating after arriving in UK",
     "'I despair as debts mount': teachers' pension delays put lives on hold"
    ],
    "reasoning": "There are signs of self-blame and hopeless framing. The text suggests negative self-
evaluation and pessimistic appraisal of circumstances, though the evidence is limited and often
embedded in quoted news material."
   },
   "dim_3_somatic_executive": {
    "evidence": [],
    "reasoning": "No reliable evidence of sleep disturbance, fatigue, appetite change, or concentration
and decision-making decline appears in the posts."
   },
   "dim_4_severe_risk": {
    "evidence": [],
```

```
    "reasoning": "There is no direct evidence of suicidal ideation, self-harm, death wishes, or other
severe risk statements in the text."
   },
   "dim_5_linguistic_features": {
    "evidence": [
     "I feel like I have failed everyone' says Nigerian nurse accused of cheating after arriving in UK",
     "'I despair as debts mount': teachers' pension delays put lives on hold"
    ],
    "reasoning": "The corpus includes emotionally loaded first-person language and some absolutist
framing such as 'everyone' and 'despair', but overall the posts are predominantly news headlines rather
than a consistent self-focused personal stream."
   },
   "dim_6_behavioral_coping": {
    "evidence": [],
    "reasoning": "No evidence of substance misuse, marked withdrawal, staying in bed, or inability to
maintain daily life is present."
   },
   "dim_7_stressors_comorbidity": {
    "evidence": [
     "'I feel like I have failed everyone' says Nigerian nurse accused of cheating after arriving in UK",
     "'I despair as debts mount': teachers' pension delays put lives on hold"
    ],
    "reasoning": "The posts reference concrete stressors such as accusation-related distress and debt
pressure, which can contribute to emotional strain. However, anxious distress or panic is not clearly
evidenced."
   }
  }
}
```


E.4.3 WU3D

AlignDx complete response.


```
{
 "dimensions_analysis": {
  "dim_1_core_affective": {
   "evidence": [],
   "reasoning": "No verifiable statements of sustained sadness, emptiness, despair, or anhedonia are
present in the posts. While themes of meaninglessness and existential reflection appear, they are framed
as artistic or philosophical expressions, not as core affective states of low mood or emotional depletion."
  },
  "dim_2_cognitive_pattern": {
   "evidence": [
    "Young people easily fall into anxiety about emptiness. Adulthood brings autonomy, yet standing at
the edge of youth also brings confusion. Would everything look different from the endpoint, and does time
 even have an endpoint? At twenty-two, everything has only just begun."
   ],
   "reasoning": "The post expresses existential uncertainty and a sense of being on the edge of life,
with reflections on meaninglessness and the passage of time. While this reflects philosophical
contemplation, it is not equivalent to negative self-view, hostile worldview, or hopelessness as required
for clinical cognitive distortions. The tone is reflective, not self-critical or despairing."
  },
  "dim_3_somatic_executive": {
   "evidence": [],
   "reasoning": "There are no mentions of sleep disturbance, fatigue, concentration decline, or appetite
/weight changes. The user's posts do not describe any physical or cognitive symptoms that impair
executive function or daily functioning."
  },
  "dim_4_severe_risk": {
   "evidence": [],
   "reasoning": "There is no mention of suicidal ideation, self-harm, death-related thoughts, profound
 worthlessness, or specific plans. The phrase 'meaningless' or 'numb' appears in a poetic or artistic
context, not as a sign of severe psychological distress."
```

```
    },
    "dim_5_linguistic_features": {
      "evidence": [
        "Meaningless Reallife Numb Booooom Ots Sins72 Oasis",
        "MEANINGLESS in everyday life 24089102",
        "No content.",
        "That is really impressive.",
        "Really good.",
        "I learned from this.",
        "Sounds good."
      ],
      "reasoning": "The posts contain frequent use of absolutist or emotionally charged terms like ' meaningless', 'numb', and 'no', but these are used in a stylistic, artistic, or ironic manner. The use of 'no content' (no) and 'really' (really) is common in social media and does not indicate excessive self-focus or linguistic patterns of depression. The linguistic features are not elevated or self-referential in a clinically significant way."
    },
    "dim_6_behavioral_coping": {
      "evidence": [],
      "reasoning": "There is no evidence of destructive coping behaviors (e.g., substance misuse) or social withdrawal that impairs daily functioning. The user engages in sharing creative work, participating in events, and interacting with others through likes, forwards, and comments, indicating active social participation."
    },
    "dim_7_stressors_comorbidity": {
      "evidence": [],
      "reasoning": "No concrete negative life events (e.g., bereavement, breakup, job loss) are mentioned. The posts do not describe panic, anxiety, or distress in response to external stressors. The content is primarily artistic, celebratory, or informational."
    }
  }
}
```


GPT-5.4-mini complete response.


```
{
  "dimensions_analysis": {
    "dim_1_core_affective": {
      "evidence": [
        "Meaningless Reallife Numb Booooom Ots Sins72 Oasis",
        "When we are young, we easily fall into anxiety about emptiness.",
        "I hope those tired of meaning itself can set aside its pressure and indulge themselves while experiencing this work."
      ],
      "reasoning": "The posts show repeated themes of emptiness, numbness, existential fatigue, and tension around meaning. This supports a subclinical negative affective pattern rather than a one-off complaint."
    },
    "dim_2_cognitive_pattern": {
      "evidence": [
        "Meaningless Reallife Numb Booooom Ots Sins72 Oasis",
        "When we are young, we easily fall into anxiety about emptiness.",
        "I hope those tired of meaning itself can set aside its pressure and indulge themselves while experiencing this work."
      ],
      "reasoning": "There is clear cognitive coloration toward meaninglessness and existential depletion, with a worldview framed around 'meaningless' and 'numb'. The text suggests rumination on meaning and pressure, but not overt self-worth devaluation or hopeless future planning."
    },
    "dim_3_somatic_executive": {
      "evidence": [],
      "reasoning": "No clear evidence of sleep disturbance, appetite change, fatigue, or concentration and decision-making decline is present in the posts."
```

```
    },
    "dim_4_severe_risk": {
      "evidence": [],
      "reasoning": "No explicit self-harm, suicidal ideation, death wishes, or profound worthlessness
statements are present."
    },
    "dim_5_linguistic_features": {
      "evidence": [
        "Meaningless Reallife Numb Booooom Ots Sins72 Oasis",
        "MEANINGLESS in everyday life 24089102"
      ],
      "reasoning": "The language includes repeated absolutist-style, globally negative terms such as '
Meaningless' and 'MEANINGLESS', indicating elevated negative framing. First-person singular
pronoun use is limited and not prominent, so this dimension is only partially supported."
    },
    "dim_6_behavioral_coping": {
      "evidence": [],
      "reasoning": "There is no evidence of substance misuse, marked withdrawal, inability to function, or
 staying in bed all day. Several posts instead reflect active posting and creative production."
    },
    "dim_7_stressors_comorbidity": {
      "evidence": [
        "The animated music video I spent two years drawing is finally complete. Doing it alone was
difficult, but I am also very happy.",
        "When we are young, we easily fall into anxiety about emptiness."
      ],
      "reasoning": "The text contains a concrete long-duration work stressor and emotional strain
around creative effort, alongside existential anxiety. This is a mild stress-and-distress signal, but not a
severe comorbidity picture."
    }
  }
}
```


E.5 Conclusion

The three cases exhibit the behaviors targeted by AlignDx's token-weighted reward and step-wise fusion. In MDPD, AlignDx integrates repeated, low-intensity descriptions of fatigue and inner depletion as Somatic–Executive evidence even without a standardized symptom statement, whereas GPT-5.4-mini leaves the dimension unsupported. In eRisk, first-person clauses in third-party headlines are not treated as user disclosures. In WU3D, negative artistic vocabulary is interpreted alongside the surrounding artwork description and the later report of completion and happiness. These examples show how AlignDx captures distributed implicit signals, separates evidence by clinical role, and incorporates attribution and cross-post context during generation.

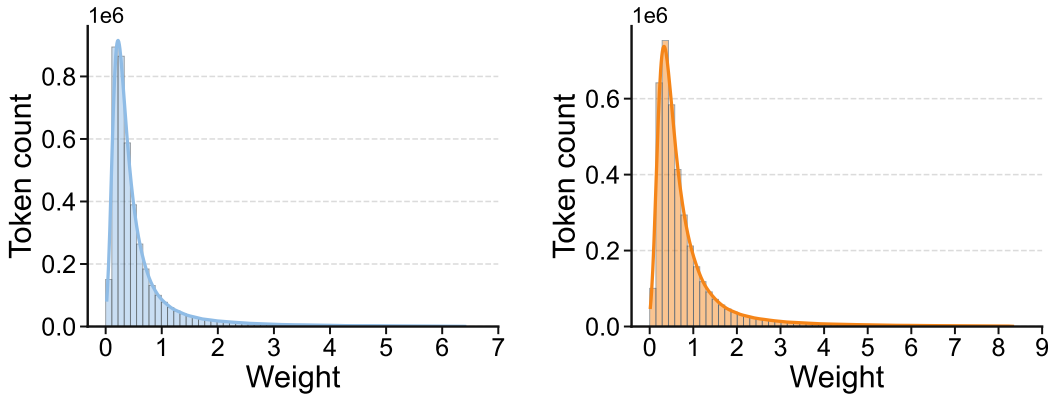
F Analysis of Token-level Importance Weights

To better understand how the proposed gradient-weighted mechanism concentrates the optimization signal, we analyze the learned token-level importance weights ω_t from two complementary perspectives: their overall statistical distribution (F.1) and their concrete allocation over a clinical analysis report (F.2).

F.1 Weight Distribution Analysis

Figure 8 reports the distribution of the learned importance weights ω_t over all tokens in the chosen responses generated by the LLaMA-3.1-8B-Instruct and Qwen3-4B-Instruct-2507 backbones. Two observations are consistent across both models.

First, the distributions are strongly right-skewed and long-tailed. The overwhelming majority of tokens are assigned small weights (concentrated in the low range around and below 1.0), while only a small fraction of tokens receive substantially larger weights, extending out to roughly 7 for LLaMA



(a) LLaMA-3.1-8B-Instruct (b) Qwen3-4B-Instruct-2507

Figure 8: Token-level importance weight distributions of chosen responses generated by Llama and Qwen models.

and 9 for Qwen. Since the weights are normalized to a per-response mean of 1, a weight below 1 indicates a token that contributes less than the average to preference discrimination, whereas the sparse high-weight tail marks the decisive tokens.

Second, the shape of the tail differs across backbones. Qwen exhibits a wider dynamic range and a heavier upper tail than LLaMA, indicating that its gradient-based attribution differentiates more sharply between decisive and non-decisive tokens.

F.2 Qualitative Analysis of Token-level Weight Allocation

To illustrate how importance weights guide the optimization process, Figure 9 presents a representative example of token-level weight allocation over a generated multi-dimensional report. As shown by the highlighting intensity, the mechanism assigns markedly higher weights to tokens that carry clinically meaningful evidence—such as expressions of negative affect, cognitive distortion, somatic symptoms, and suicide-related signals.

This allocation pattern reveals two properties of the proposed mechanism. First, high weights are distributed across multiple clinical dimensions rather than concentrated on a single symptom type, indicating that the gradient attribution captures decisive evidence wherever it appears in the report rather than favoring one particular category. Second, the mechanism attends to the evidence-bearing content of each dimension rather than to surface fluency or lexical form, so that decisive tokens are prioritized even when they are not standard diagnostic keywords.

G Inference Speed Analysis

A naive implementation of reward-guided decoding would run the base branch and the reward branch as two independent serial requests, effectively doubling the per-step forward cost. To further reduce this computational overhead, we maintain two branches of the base model. The primary branch uses the base-model parameters to produce the original next-token distribution π_{base}, whereas the reward branch activates a pretrained reward-model LoRA adapter to produce the preference-guided distribution π_{rm}. Within the SGLang inference engine, this is realized through paired primary–shadow requests. The primary request maintains the final output sequence, while the shadow request produces only the logit signal from the reward branch. The two requests strictly synchronize their token histories and KV-cache updates, minimizing the additional computational overhead of preference guidance.

We compare the generation throughput of three inference strategies: (1) the base model alone (Base); (2) the base model produces its logits, the LoRA adapter generates preference-guided logits, and the two are subsequently fused in logit space; and (3) Dual-Path Logit Fusion, which computes the Base and Base+LoRA distributions at each decoding step, fuses their logits, and synchronizes the

Table 6: Generation throughput under different inference strategies, measured in tokens/s.

Inference Strategy	LLaMA	Qwen
Base Model	65.47	93.13
Two Single-Path	30.47	41.04
Dual-Path	31.03	41.88

Table 7: BERT classifier results. Qwen and LLaMA denote Qwen3-4B-Instruct-2507 and LLaMA-3.1-8B-Instruct.

Backbone	Dataset	ACC	F1	MCC
Qwen	MDPD	0.9014	0.8404	0.7792
	eRisk	0.8914	0.5966	0.5428
	WU3D	0.9108	0.8628	0.7974
LLaMA	MDPD	0.8384	0.7306	0.6305
	eRisk	0.8753	0.5153	0.4482
	WU3D	0.8914	0.8276	0.7484

Table 8: Qwen3-0.6B classifier results. Qwen and LLaMA denote Qwen3-4B-Instruct-2507 and LLaMA-3.1-8B-Instruct.

Backbone	Dataset	ACC	F1	MCC
Qwen	MDPD	0.9060	0.8553	0.7882
	eRisk	0.8891	0.6080	0.5588
	WU3D	0.9081	0.8656	0.8015
LLaMA	MDPD	0.8593	0.7686	0.6799
	eRisk	0.8562	0.5328	0.4760
	WU3D	0.8880	0.8336	0.7524

sampled token across both paths. The two paths share the base-model weights but maintain separate KV caches.

The results in Table 6 confirm the benefit of this design. Compared with the serial Base+LoRA baseline, Dual-Path Logit Fusion yields a slight improvement in throughput—from 30.47 to 31.03 tokens/s on LLaMA and from 41.04 to 41.88 tokens/s on Qwen.

H Discussion on Classifiers

We evaluate the generated analyses using two downstream classifiers—BERT and Qwen3-0.6B—both applied with a fixed decision threshold of 0.5. Tables 7 and 8 report their performance across the MDPD, eRisk, and WU3D datasets under both the Qwen3-4B-Instruct-2507 and LLaMA-3.1-8B-Instruct generators.

Overall, the two classifiers exhibit highly consistent trends, indicating that AlignDx's gains stem primarily from the quality of the generated multi-dimensional analyses rather than from the choice of a particular downstream classifier. Under both classifiers, the Qwen generator consistently outperforms the LLaMA generator across all three datasets and all metrics, confirming that a stronger, better-aligned analysis backbone yields more discriminative structured reasoning regardless of the classifier used.

Comparing the two classifiers, Qwen3-0.6B slightly outperforms BERT in most settings. For example, with the Qwen generator on MDPD, Qwen3-0.6B improves MCC from 0.7792 to 0.7882 and F1 from 0.8404 to 0.8553; on eRisk it raises MCC from 0.5428 to 0.5588. The advantage is more pronounced under the LLaMA generator: on MDPD, Qwen3-0.6B lifts MCC from 0.6305 to 0.6799 and F1 from 0.7306 to 0.7686. This suggests that a lightweight generative classifier can



Figure 9: Example of token-level importance weight allocation.Tokens associated with clinically relevant evidence receive higher importance weights, demonstrating that the proposed weighting mechanism captures meaningful depression-related signals.

more effectively exploit the structured clinical reasoning, particularly when the underlying analysis is comparatively weaker.

Notably, the performance gap between the two classifiers remains modest relative to the gap between the two generators. This confirms that the decisive factor is the faithfulness and completeness of the generated evidence rather than the classifier architecture, and that AlignDx is robust to the specific choice of downstream classifier.