Disagreement-Driven Joint Refinement of Retrieval and Decision Rules for Imbalanced Counseling Risk Classification

Zhihao Shao[1], **Ryo Sekizaki**[2], **Shengzhou Yi**[1], **Toshihiko Yamasaki**[1]
[1]The University of Tokyo
[2]Welcome to Talk Co., Ltd.
{shao,yishengzhou,yamasaki}@cvm.t.u-tokyo.ac.jp
sekizaki-ryo@welcometotalk.co.jp

Abstract

With the rapid growth of online counseling services, timely and reliable risk classification of counseling records is essential for supporting early screening and prioritizing limited intervention resources. High-risk samples refer to high-acuity suicide risk and require expedited human review. However, this task is challenging due to severe class imbalance (93% low-risk and 7% high-risk samples) and complex decision boundaries. Large language models (LLMs) exhibit unstable predictions and systematic errors in such imbalanced clinical-text settings. To address this issue, we propose Disagreement-Driven Joint Refinement (DDJR), an iterative, parameter-free refinement framework. It uses prediction disagreement between two inference settings, zero-shot and retrieval-augmented in-context learning, as the primary signal for identifying high-value instances. These disagreement-identified instances are transformed into adaptive refinement signals and used to jointly update both the exemplar pool and an executable rule set, thereby sharpening decision boundaries and improving prediction stability. Experiments on 6,481 real-world counseling records demonstrate that the proposed DDJR outperforms existing methods, achieving an accuracy of 0.915 and a Matthews Correlation Coefficient (MCC) of 0.583. These results demonstrate that DDJR achieves more stable and reliable predictions for high-stakes counseling risk classification in real-world settings.

Keywords: Mental Health Intervention, Large Language Models, Self-correction, Exemplar Retrieval

1. Introduction

The World Health Organization (WHO) estimates that approximately 970 million people worldwide are affected by mental disorders. Among these, depression and anxiety are the most prevalent, and suicide remains one of the leading causes of death among individuals aged 15 to 29 (World Health Organization, 2023; Shen et al., 2024). Text-based online counseling has improved access to mental health support while offering greater privacy. However, the absence of nonverbal cues in text-only interactions means that emotional assessment must rely primarily on linguistic signals (Park et al., 2019; Zhang et al., 2022). Therefore, reliable analysis of counseling dialogues is critical for triage and early intervention, as high-risk samples may require expedited human review. In this work, counseling risk is defined in terms of suicide-risk acuity: high-risk denotes high acuity, whereas low-risk denotes low acuity.

As illustrated in Figure 1, automatic counseling risk classification of dialogues can support professionals by identifying high-risk samples and prioritizing them for intervention. Early rule-driven and statistical methods, followed by neural models such as recurrent neural networks (Elman, 1990), improved linguistic modeling but remained limited in capturing long-range context and complex affective dependencies (Kermani et al., 2025). More recently, transformer-based models and large language models (LLMs) have strengthened contextual modeling, motivating increased interest in applying LLMs to counseling risk classification in mental health settings (Matero et al., 2019; Li et al., 2024). In this domain, interpretability and evidence grounding are critical. Nevertheless, counseling dialogue classification remains challenging because distress signals are often indirect and distributed across multiple turns, while LLMs may overemphasize surface cues or generate weakly supported rationales. This challenge is further exacerbated by severe class imbalance, as high-risk samples are rare but clinically critical.

To address these challenges, we propose Disagreement-Driven Joint Refinement (DDJR), an iterative inference-time optimization framework for imbalanced counseling risk classification. Our main contributions are as follows:

1. Disagreement-driven contrastive typing supervision is proposed to assign each instance to one of three types based on prediction disagreement, which provides signals for refinement without parameter updates.
2. Joint refinement of the exemplar pool and the rule set is proposed to enable closed-loop refinement without parameter updates.
3. Type-weighted exemplar retrieval is proposed to prioritize discrepancy and failure-oriented evidence from samples with different learning values, improving calibration for imbalanced counseling risk classification.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 281–289
16 May 2026.

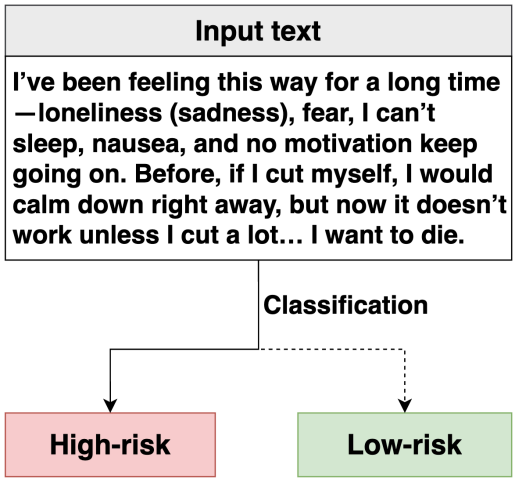

Figure 1: Suicide-risk acuity classification for counseling dialogues.

2. Related Work

2.1. Mental Health Intervention Systems

Dialogue-based mental health intervention systems have been explored as scalable support in text-based settings. Early systems such as Woebot (Fitzpatrick et al., 2017) demonstrated the feasibility of providing cognitive behavioral therapy-inspired guidance through scripted dialogue management and lightweight personalization. Subsequent reports on Tess (Rauws et al., 2019) described affect-aware interactions through emotion detection from user messages together with multilingual delivery. Systematic evidence indicated that mental-health chatbots could improve symptoms in some settings, but they also revealed substantial heterogeneity in study quality and limited safety reporting (Abd-Alrazaq et al., 2020). More broadly, conversational agents in healthcare remained an emerging area in which rigorous evaluation protocols and safety assessments have not yet been consistently standardized (Laranjo et al., 2018). Within psychiatry, recent reviews have further summarized the use of chatbot for both screening and intervention, emphasizing gaps in real-world validation and reporting (Vaidyam et al., 2019).

Recent work with large language models explored reasoning-oriented prompting for empathetic response generation, including Chain-of-Empathy (CoE), which performed explicit emotion analysis before response generation (Lee et al., 2023), and approaches that integrated rational cognition with value-oriented dialogue to promote well-being (Mercer, 2022). In contrast, our work targets risk triage for counseling dialogues, where decision stability under severe class imbalance and evidence alignment were central requirements for safe escalation to human review.

2.2. Retrieval-Augmented In-Context Learning for Counseling Dialogues

In high-stakes settings, it was desirable to ground model outputs in inspectable external sources rather than relying solely on parametric memory, thereby motivating retrieval-augmented approaches for more reliable and transparent predictions (Guu et al., 2020). Beyond open-domain retrieval-augmented generation (RAG) (Lewis et al., 2020), retrieval-augmented in-context learning (RA-ICL) retrieved labeled exemplars and uses them as in-context demonstrations.

Retrieval-augmented pipelines have been studied as a grounding strategy in mental health text classification, aiming to improve reliability and produce more explainable outputs (Kermani et al., 2025). However, RA-ICL was highly sensitive to demonstration selection. Dr.ICL, a retrieval-based in-context learning method, showed that retrieval-based demonstrations could outperform random sampling (Luo et al., 2024).

Motivated by this sensitivity, DDJR improves robustness under severe class imbalance by refining the RA-ICL exemplar pool to better capture discrepancy- and failure-oriented evidence, mitigating the under-coverage of hard cases during retrieval.

2.3. Self-Refinement in LLMs

Self-improvement strategies for LLMs can improve behavior either at inference time without parameter updates or through self-training with parameter updates. Reflection-based methods such as Reflexion (Shinn et al., 2023) and SELF-REFINE (Madaan et al., 2023) generated critiques to guide iterative revision without changing model weights, while STaR (Zelikman et al., 2022) improved reasoning by generating rationales and iteratively fine-tuning the model on them.

Recent work has further introduced structured self-verification to improve reliability. ProCo (Wu et al., 2024) used key-condition masking and iterative verification to refine outputs in complex reasoning tasks, and PAG (Jiang et al., 2025) alternated between policy and generative-verifier roles to detect errors and revise outputs across multi-turn interactions. Constraint-guided frameworks such as Constitutional AI externalized decision rules to produce more stable and interpretable outcomes (Bai et al., 2022).

In contrast, DDJR targeted risk classification in counseling dialogues under severe class imbalance. It uses disagreement signals to jointly refine the exemplar pool and the rule set, thereby improving evidence alignment for risk triage.

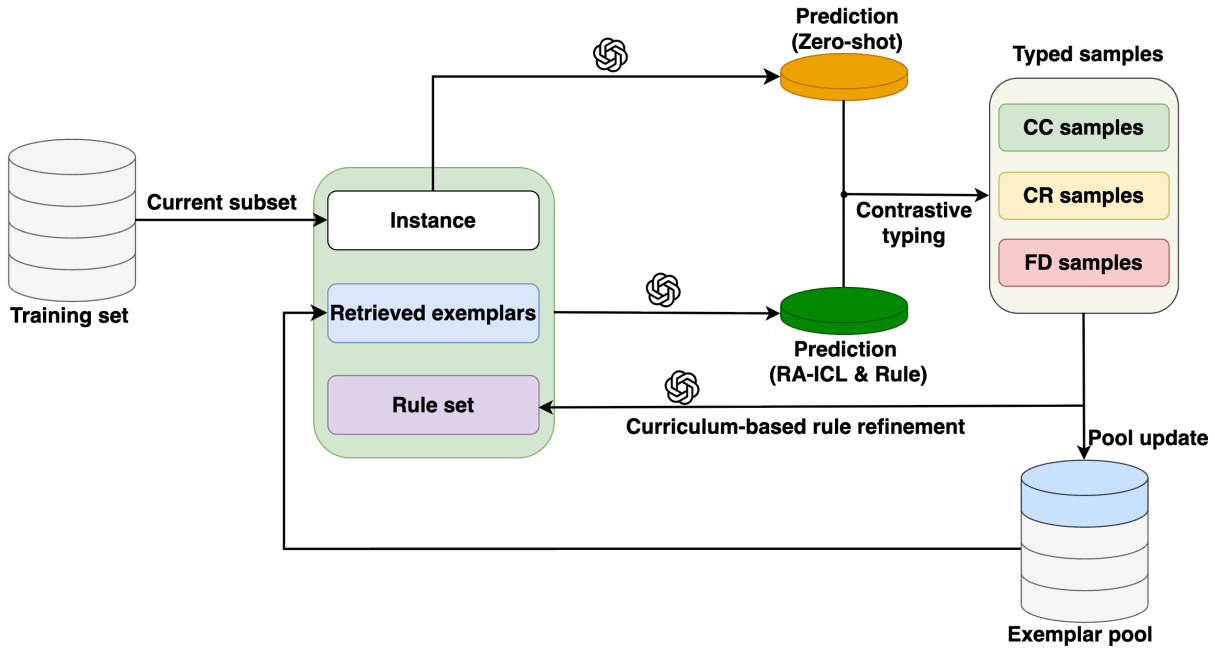

Figure 2: The overview of the Disagreement-Driven Joint Refinement (DDJR) framework.

3. Methodology

3.1. Overview

As shown in Figure 2, the proposed framework is a parameter-free iterative refinement loop that operates on the training set without model weight updates. It consists of three steps: typing instances, updating an exemplar pool, and refining a rule set. The exemplar pool and the rule set are updated jointly across refinement rounds.

3.2. Contrastive Typing Supervision

We obtain typed supervision signals by running dual inference for each labeled instance and using prediction disagreement in correctness outcomes as the primary refinement signal.

Let the training dataset be denoted as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{M} \tag{1}$$

where x_i represents a counseling dialogue and $y_i \in \{0, 1\}$ denotes its risk label, with 0 indicating low-risk and 1 indicating high-risk.

For each input text x_i, we run dual inference: (i) zero-shot inference, yielding prediction $\hat{y}_i^{(0)}$; and (ii) RA-ICL inference, yielding prediction $\hat{y}_i^{(1)}$. The goal is to derive structured supervision signals for exemplar pool and rule refinement, rather than performing prediction ensembling.

To characterize the outcomes of dual inference, we define a correctness indicator for each configuration:

$$c_i^{(k)} = \mathbf{1}\big(\hat{y}_i^{(k)} = y_i\big), \quad k \in \{0, 1\}, \tag{2}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. We then construct a correctness vector

$$c_i = \big(c_i^{(0)}, c_i^{(1)}\big) \in \{0, 1\}^2 \tag{3}$$

By construction, each instance must fall into exactly one of the following mutually exclusive categories:

$$c_i \in \{(1, 1)\} \cup \{(1, 0), (0, 1)\} \cup \{(0, 0)\} \tag{4}$$

Note that this contrastive typing requires gold labels and is performed offline on labeled data during the refinement stage.

Based on this correctness vector, we define three types of supervision signals:

(i) Consistently-Correct samples. When $c_i = (1, 1)$, both configurations predict correctly. We directly construct a standard Consistently-Correct (CC) sample (x_i, y_i). Such instances reflect consistent behavior across prompting conditions and typically correspond to relatively clear decision boundaries.

(ii) Contrastive Rationalization samples. When $c_i \in \{(1, 0), (0, 1)\}$, exactly one configuration predicts correctly. We construct a Contrastive Rationalization (CR) sample consisting of the input text, the ground-truth label, and two rationales: one generated by the correct configuration and the other by the incorrect configuration. Because the input is identical while correctness differs, these samples

highlight discrepancies in evidence selection and contextual interpretation, providing supervision on reasoning patterns to encourage or discourage.

(iii) Failure-Diagnostic samples. When $c_i = (0, 0)$, both configurations predict incorrectly. We treat the instance as a hard case and construct a Failure-Diagnostic (FD) sample. In this setting, we provide the gold label together with the incorrect rationales, enabling the model to summarize the likely missing cues, misleading cues, or boundary confusions shared by both failed predictions. Each FD instance therefore comprises the input text, the ground-truth label, and the generated error diagnosis.

Through this correctness-vector induced mechanism, CC samples capture stable patterns, CR samples provide contrastive reasoning supervision, and FD samples summarize shared failure modes. Together, they form a structured supervision pool that supports both rule refinement and exemplar pool updates.

3.3. Exemplar Pool Refinement

The few-shot exemplar pool is initialized with the full training set and its size is kept fixed across iterations. In each iteration, every newly added CR or FD sample replaces one CC sample, so the pool gradually shifts toward more informative hard cases without changing its total size.

Each CC, CR, or FD sample is stored in a unified exemplar format. Each sample is stored as a retrievable exemplar containing the original counseling dialogue x, the gold label y, and type-specific auxiliary fields when needed. For CC samples, we store only the text and label (x, y); for CR samples, we store both the correct and incorrect rationales to provide contrastive guidance; for FD samples, we store an error diagnosis to surface likely failure causes and boundary-confusing patterns. This unified sample representation enables the retriever to return heterogeneous yet complementary exemplars.

At test time for a new query x_q, we retrieve the top-k exemplars from the exemplar pool and include them as in-context demonstrations. In our experiments, we fix $k = 64$. We further apply type-weighted retrieval to prioritize informative CR and FD samples during inference.

3.4. Curriculum-based Rule Construction and Refinement

We refine an executable rule set in a label-aware manner to mitigate rule bias under severe class imbalance. Besides the learning-value types derived from contrastive typing, we additionally condition rule induction on the gold risk label. This design is motivated by the asymmetric utility of errors in risk triage: under low high-risk prevalence, unconstrained aggregation across labels tends to overrepresent majority (low-risk) regularities, yielding conservative constraints that can suppress minority detection. By separating high-risk (y=1) and low-risk (y=0) evidence during refinement, we encourage the rule refiner to extract label-specific decision cues and to preserve high-risk sensitivity while retaining low-risk exclusion criteria.

Formally, we split the constructed samples by gold label, yielding six partitions:

$$\mathcal{S} = \{\text{CC-0},\ \text{CC-1},\ \text{CR-0},\ \text{CR-1},\ \text{FD-0},\ \text{FD-1}\}.$$

Within each label-specific partition, we further organize instances by a correctness axis that indicates whether the corresponding inference configuration predicts the gold label correctly. To avoid ambiguity with class polarity, we denote this axis as correct and incorrect. Correct instances provide higher-reliability evidence for initial constraint formulation, whereas Incorrect instances expose boundary confusions and missing conditions that are informative for subsequent rule revision.

We adopt a two-level curriculum schedule that is both label-aware and difficulty-aware, and proceeds in an explicit easy-to-hard, as shown in Figure 3. Here, difficulty is determined by the correctness outcomes of the two inference configurations. Samples for which both configurations are correct are treated as easy (CC), those for which only one configuration is correct are treated as intermediate (CR), and those for which both configurations are incorrect are treated as hard (FD). Within each gold-label group, evidence is further organized by correctness, and later stages incorporate less separable cases to revise and qualify previously induced constraints.

For each label group (y=0 and y=1), refinement starts from more discriminative and reliable evidence and progressively incorporates less separable evidence to amend previously induced constraints. Specifically, we introduce CC subsets first, then CR subsets, and finally FD subsets. CC cases are easy because both configurations are correct, CR cases are intermediate because only one is correct, and FD cases are hard because both are incorrect. This progression stabilizes early rule formation, while later stages emphasize Incorrect cases to introduce disambiguation constraints and exception constraints that improve robustness on borderline dialogues.

To control partition imbalance, we sample up to a predefined upper bound from each partition and process instances in mini-batches of 25.

Let Rule_t denote the rule set at step t. At each curriculum step, the rule-update prompt takes as input (i) the current rule set and (ii) a mini-batch of typed samples from the current subset. The LLM

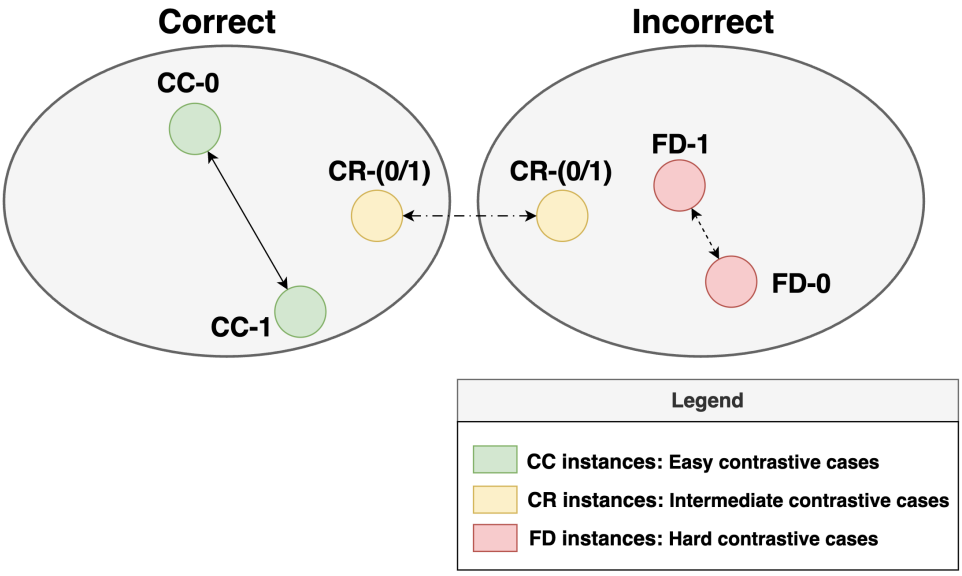

Figure 3: Label-aware easy-to-hard curriculum for rule refinement.

then regenerates the complete rule set in natural language and organizes it into four parts: high-risk indicators, low-risk indicators, false-positive prevention, and false-negative prevention. Given newly proposed constraints $\mathrm{New\_Rule}_t$ summarized from the current batch, we update:

$$\mathrm{Rule}_t = \mathrm{Refine}\big(\mathrm{Rule}_{t-1} \cup \mathrm{New\_Rule}_t\big). \quad (5)$$

We implement rule updates by prompting the LLM with the previous rule set and the newly introduced typed samples, then asking it to regenerate a revised full rule set. The updated rule set removes redundant constraints, merges overlapping rules, and relaxes overly specific formulations to improve generalizability.

3.5. Type-Weighted Exemplar Retrieval and Rule Injection

At test time, DDJR performs RA-ICL inference with type-weighted exemplar retrieval and full injection of the refined rule set. Given a query text x_q, each case s in the exemplar pool has type $\tau(s) \in \{\mathrm{CC}, \mathrm{CR}, \mathrm{FD}\}$ and similarity score $\mathrm{sim}(x_q, s)$. We apply a type-dependent weight, $w_{\tau(s)}$, to the similarity score:

$$\mathrm{score}(x_q, s) = \mathrm{sim}(x_q, s) \cdot w_{\tau(s)} \quad (6)$$

We select the top-k samples according to $\mathrm{score}(x_q, s)$ as in-context demonstrations, where higher weights emphasize CR and FD to surface discrepancy and failure cues for borderline inputs.

We assign larger type weights to CR and FD cases so that discrepancy and failure-oriented exemplars are prioritized during retrieval and injects the full refined rule set Rule_t into the prompt as explicit decision guidance to encourage consistent and evidence-aligned predictions.

4. Experiments

4.1. Implementation

Experiments were conducted on an annotated corpus of online student counseling records, consisting of 6,481 records labeled by trained annotators. All records were anonymized prior to use, with personally identifiable information removed. We adopted the binary risk setting of low-risk versus high-risk, where high-risk corresponded to high-acuity suicide risk and low-risk corresponded to low-acuity suicide risk.

The dataset was divided into training, validation, and test sets with the ratio of $8:1:1$, stratified to preserve label proportions across splits. The resulting labels were imbalanced, with high-risk samples accounting for approximately 7% of the corpus.

We used GPT-4.1 (Achiam et al., 2023) as the backbone model for all LLM-based methods and compared several prompting-based inference settings under a shared retrieval setup. Unless otherwise specified, RA-ICL retrieved $k = 64$ semantically similar labeled exemplars based on vector similarity for in-context prompting. In all experiments, the type weights were set to $w_{CC} : w_{CR} : w_{FD} = 1 : 1.5 : 1.5$. All methods were assessed with the same backbone model and retrieval setting, and we reported accuracy (ACC), high-risk recall, balanced accuracy (BACC), and Matthews Correlation Coefficient (MCC).

Method	ACC	MCC	Recall	BACC
Zero-shot learning	0.643	0.316	**1.000**	0.808
Few-shot learning (Random)	0.764	0.402	0.978	0.863
ExRAG (Kermani et al., 2025)	0.843	0.453	0.889	0.864
CoE (Lee et al., 2023)	0.737	0.376	0.978	0.848
Fine-tuning (Achiam et al., 2023)	**0.944**	0.561	0.578	0.775
DDJR (Ours)	0.915	**0.583**	0.867	**0.893**

Table 1: Comparison with different methods under the same dataset split, backbone model, and evaluation metrics.

Sample usage	ACC	MCC	Recall	BACC
CC (Kermani et al., 2025)	0.843	0.453	**0.889**	0.864
CC+FD	**0.923**	0.415	0.467	0.712
CC+CR	0.889	0.523	0.867	**0.879**
CC+FD+CR	0.905	**0.546**	0.844	0.877

Table 2: Impact of FD and CR on sample effectiveness.

All prediction prompts used a shared task-definition template and required the model to output a structured JSON object containing a binary label and a free-text rationale. The zero-shot setting used only the task instruction and output schema, whereas the RA-ICL setting prepended retrieved demonstrations in the same output format. At test time, DDJR made the final prediction with RA-ICL using the refined exemplar pool as in-context demonstrations and the refined rule set as explicit decision guidance.

4.2. Comparison with Baseline Methods

We adopt a hierarchy of increasingly competitive baselines to assess gains from generic prompting, contextual alignment, and iterative self-improvement. We also include a fine-tuned GPT-4o (Achiam et al., 2023) model for reference; without cost-sensitive tuning, fine-tuning may optimize aggregate accuracy while under-detecting the minority high-risk class, leading to lower Recall and BACC despite high ACC. The zero-shot setting relies only on task instructions and the target record, and is therefore susceptible to confusing transient emotional intensity with sustained psychological risk. Few-shot prompting improves label discrimination but can remain unstable due to sensitivity to the selected in-context examples. Building on RA-ICL, DDJR further refines the exemplar pool and rule set before test-time inference.

4.3. Structured Sample Usage

We treat ExRAG (Kermani et al., 2025) as the CC-only variant, where CC refers to standard consistently-correct exemplars. Its exemplar pool contains only text-label exemplars, without the CR (contrastive rationalization) or FD (failure-diagnostic) auxiliary fields used in DDJR.

Method	ACC	MCC	Recall	BACC
CC w/ CoE (Lee et al., 2023)	0.838	0.447	0.889	0.862
CC Non-curriculum strategy	**0.918**	**0.501**	0.667	0.802
CC CBRR	0.827	0.453	**0.933**	**0.876**

Table 3: Comparison of curriculum-based rule refinement (CBRR) starting from the same CC-only instance set.

We examine how different combinations of disagreement-derived sample types affect performance. In this comparison, only the included sample types are varied, while the backbone model, retrieval setting, and downstream inference procedure are kept unchanged, and no additional rule-update rounds are performed.

We compare four settings: (i) CC-only; (ii) CC+FD; (iii) CC+CR; and (iv) CC+FD+CR. This comparison examines whether different disagreement-derived sample types provide complementary benefits. Table 2 indicates that the contribution of typed samples is type-dependent.

Adding FD without CR reduces high-risk recall substantially, from 0.889 to 0.467, and also lowers BACC from 0.864 to 0.712, although ACC increases from 0.843 to 0.923. This suggests that many FD cases primarily emphasize correcting over-sensitive false-positive decisions; in the absence of contrastive guidance, such signals may inadvertently suppress high-risk predictions.

Incorporating CR improves the overall trade-off in MCC and BACC while largely preserving high-risk recall relative to CC-only. This supports the role of CR in providing contrastive evidence where the same input yields different outcomes across prompting configurations thereby encouraging alignment with more appropriate evidence usage.

Combining CR and FD yields the most balanced overall performance, suggesting that CC, CR, and FD provide complementary supervision under class imbalance: CC supplies stable but conservative signals, CR anchors decision boundaries via contrastive instances, and FD complements these signals by summarizing persistent failure patterns that remain even with retrieval augmentation.

4.4. Curriculum-Based Rule Refinement

We evaluate whether curriculum-based rule refinement (CBRR) is necessary for robust risk detection under severe class imbalance. For clarity, both "w/o CBRR" and "w/ CBRR" start from the same CC-only instance set, and differ only in whether the curriculum schedule is applied during rule refinement. We compare two settings: (i) without CBRR, where rule refinement is performed on the

same exemplar pool using a non-curriculum strategy (i.e., removing the easy-to-hard, label-aware progression); and (ii) with CBRR, which follows the proposed curriculum schedule. In both settings, the backbone model, retrieval pipeline, and evaluation protocol are kept identical; the only difference is whether the curriculum schedule is applied during rule refinement.

As shown in Table 3, removing CBRR significantly weakens the model's ability to identify high-risk samples. In contrast, incorporating CBRR better balances detecting high-risk samples against avoiding false alarms on low-risk ones, leading to the highest BACC of 0.876. CBRR improves minority sensitivity, yielding higher Recall and BACC, but also increases false positives, resulting in lower ACC. This trade-off is preferable in practical screening, where missing high-risk samples is much more costly than making errors on low-risk samples.

4.5. Ablation Study and Qualitative Analysis

Method	ACC	MCC	Recall	BACC
DDJR (Ours)	**0.915**	**0.583**	0.867	**0.893**
w/o Exemplar pool refinement	0.777	0.404	**0.956**	0.859
w/o Rule refinement	0.905	0.546	0.844	0.877

Table 4: The ablation study of DDJR.

To evaluate the contributions of the two key design choices in DDJR, we conduct an ablation study by removing: (i) exemplar pool refinement, which constructs the types of disagreement-derived instances via dual-inference comparison, and (ii) rule refinement, which refines the rule set following the proposed curriculum schedule. All variants are evaluated under the same LLM, prompt skeleton, and retrieval setting; the only changes are the removal of the corresponding component.

Table 4 reports the results of the ablation study. Removing contrastive typing (i.e. without exemplar pool refinement) leads to a substantial degradation in ACC and MCC, indicating that the construction of the three sample types is critical for producing informative supervision signals and for stabilizing decision boundaries. In contrast, removing rule refinement (without rule refinement) yields a smaller but consistent drop across MCC, Recall, and BACC, suggesting that the curriculum schedule contributes to maintaining balanced performance under class imbalance by refining rules in a structured manner. Collectively, the full DDJR configuration achieves the best performance, demonstrating that both contrastive typing and rule refinement are necessary to obtain robust counseling risk classification. Although DDJR does not achieve the highest high-risk recall, it achieves the best ACC (0.915), MCC (0.583), and BACC (0.893), suggesting that it offers a more balanced trade-off between sensitivity to high-risk cases and overall classification robustness.

Example Analysis

Text:
…But lately, I've started to strongly feel like I want to die. It's not like anything in particular happened—***my relationships and grades at school are decent—but I still feel like I want to die. I can't talk about this with my parents,*** teachers, or friends. Because I'm known as the "cheerful one," if I told people something like this, they'd worry about me. I can't bring myself to say "I want to die," not even as a joke. Why have these suicidal thoughts suddenly come up...

Ground-truth label: Low-risk

Method: Zero-shot learning
Result: High-risk
Reason: There have been ***suicide suggestive remarks from the parent***, such as "Shall we die together?" and "There's no point in living," ***and the person seeking help is under severe stress related to school, future plans, relationships, and the home environment;*** therefore, the suicide risk is assessed as high.

Method: DDJR
Result: Low-risk
Reason: The suicidal ideation ("strongly feel like I want to die") is clearly expressed, but there is no mention of a specific suicide plan, intent to act, self-harm behavior, or an imminent sense of crisis. The main complaint remains confusion about why these feelings arose and the conflict of being unable to seek help. ***Based on the [LowRisk Indicators] [False Positive Fix] rule, the suicide risk is judged to be low at this time.***

Figure 4: Case study.

Figure 4 presents a representative qualitative example illustrating how the proposed components mitigate typical false positives. The record describes a student's distress related to schooling and family pressure, and includes a parent's statement about "dying together". Our method predicts low-risk by attributing the suicidal wording to the parent rather than the student, and by identifying contextual cues indicating agency and forward-looking intent (e.g., a desire for independence and plans to continue living). In contrast, the baseline RA-ICL setting misclassifies the record as high-risk, overemphasizing death-related keywords and failing to account for speaker attribution and broader discourse context. This case suggests that contrastive typing supplies contrastive and failure-oriented exemplars that surface such misinterpretations, while the proposed rule set helps regularize the decision boundary by discouraging keyword-triggered overgeneralization. Together, these com-

ponents promote a more faithful interpretation of emotionally charged language and reduce false positives in risk triage.

5. Conclusion and Limitations

We propose DDJR, an iterative parameter-free refinement framework for imbalanced counseling risk classification that does not update model parameters. The method improves retrieval-augmented inference by refining the exemplar pool and decision rules through disagreement signals derived from zero-shot and RA-ICL predictions. Experimental results showed that DDJR achieved the best overall balance among ACC, MCC, Recall, and BACC, indicating more robust and reliable performance under severe class imbalance.

Several limitations should also be noted. First, the recall of DDJR indicates that some high-risk samples may still be missed. Second, the proposed refinement is conducted offline on labeled training data, while test-time use only relies on the finalized exemplar pool and refined rule set for inference. Third, the study is based on counseling records from a single institution, and the generalizability of the learned retrieval patterns and induced rules to other populations and settings remains unclear. Future work will examine external validation, cross-institutional transfer, and the stability of the refined rule set under distribution shift.

6. Acknowledgements

This work has been approved by the Institutional Review Board of our graduate school: UT-IST-RE-231102. This work was partially financially supported by JST ASPIRE Program, Japan, Grant Number JPMJAP2303 and The Toyota Foundation, Co-Creating New Society with Advanced TechnologiesGrant Number: D24-ST-0006.

Alaa A. Abd-Alrazaq, Asma Rababeh, Mohannad Alajlani, Bridget M. Bewick, and Mowafa Househ. 2020. Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 22(7):e16021.

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

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, and J. Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive Science*, pages 179–211.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Mental Health*, 4(2):e19.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *ICML*, pages 3929–3938.

Y. Jiang, Y. Xiong, Y. Yuan, C. Xin, W. Xu, Y. Yue, and L. Yan. 2025. Pag: Multi-turn reinforced llm self-correction with policy as generative verifier. *arXiv preprint arXiv:2506.10406*.

Arshia Kermani, Veronica Perez-Rosas, and Vangelis Metsis. 2025. A systematic evaluation of LLM strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. RAG. pages 172–180.

Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. 2023. Chain of empathy: enhancing empathetic response of large language models based on psychotherapy models. *arXiv preprint arXiv:2311.04915*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*, pages 9459–9474.

Wenjie Li, Tianyu Sun, Kun Qian, and Wenhong Wang. 2024. Optimizing psychological counseling with instruction-tuned large language models. *arXiv preprint arXiv:2406.13617*.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y. Zhao. 2024. Dr.icl: Demonstration-retrieved in-context learning. *Data Intelligence*, pages 909–922.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *NeurIPS*, pages 46534–46594.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *CLPsych*, pages 39–44.

Calvin Mercer. 2022. Mental and spiritual health needs of cognitively enhanced people: a therapeutic and spiritual care model for responding. *Religions*, page 701.

Sungjoon Park, Donghyun Kim, and Alice Oh. 2019. Conversation model fine-tuning for classifying client utterances in counseling dialogues. In *NAACL-HLT*, pages 1448–1459.

Michiel Rauws, John Quick, and Nancy Spangler. 2019. X2 AI Tess: working with AI technology partners. *The Journal of Employee Assistance*, 49(1):22–25.

Hao Shen, Zihan Li, Minqiang Yang, Minghui Ni, Yongfeng Tao, Zhengyang Yu, Weihao Zheng, Chen Xu, and Bin Hu. 2024. Are large language models possible to conduct cognitive behavioral therapy? In *BIBM*, pages 3695–3700.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. *NeurIPS*, pages 8634–8652.

Anand N. Vaidyam, Hannah Wisniewski, John D. Halamka, Matcheri S. Kashavan, and John B. Torous. 2019. Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.

World Health Organization. 2023. Depressive disorder (depression).

Z. Wu, Q. Zeng, Z. Zhang, Z. Tan, C. Shen, and M. Jiang. 2024. Large language models can self-correct with key condition verification. In *EMNLP*, pages 12846–12867.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. STaR: Bootstrapping reasoning with reasoning. In *NeurIPS*, pages 15476–15488.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *npj Digital Medicine*, page 46.

A. Example Rule Set

The rules were originally generated in Japanese and are shown here in English for readability.

High-risk indicators. Specific suicide methods, suicide attempts, overdose (OD), wrist-cutting, use of means such as knives, suicide preparation/planning, a suicide note or will, or a history of emergency transport. If suicidal intent or action is clear, classify the case as high risk.

Low-risk indicators. Expressions such as “I want to die, but I do not want to die,” “Small things make me want to die,” or “I keep going back and forth between wanting to live and wanting to die,” when they reflect only ambivalence or temporary suicidal ideation without any concrete action or plan. In such cases, classify the case as low risk.

False-positive prevention. Even if phrases such as “I want to die” or “I want to disappear” appear, do not classify the case as high risk if there is no specific plan, method, intent, or self-harm behavior, and the text mainly expresses distress or help-seeking.

False-negative prevention. Even if phrases such as “I want to die” or “I want to disappear” are indirect, classify the case as high risk if they are accompanied by suicide preparation, planning, self-harm, suicide attempts, a suicide note, or strong hopelessness and urgency.

B. Cost and Scalability Discussion

DDJR introduces additional overhead compared with zero-shot and standard RA-ICL baselines, primarily because it uses longer prompts and an offline multi-round refinement procedure before final test-time prediction. Since DDJR does not update model parameters, its cost is better understood as online prediction cost plus offline refinement overhead rather than training cost.

Under our setup, all three methods require only one LLM call per test instance, but their prompt sizes differ substantially: approximately 1,000 characters for zero-shot, around 25,000 for standard RA-ICL, and slightly above 50,000 for DDJR. This corresponded to an estimated per-instance input cost of about \$0.003, \$0.04–\$0.05, and \$0.06–\$0.09, respectively. In addition, DDJR incurred a one-time offline refinement cost of approximately \$318 prior to final test-time inference. These observations suggest that the main overhead of DDJR comes from the one-time offline refinement stage, while deployment only uses the finalized exemplar pool and refined rule set for inference.