SuiChat-CN: Benchmarking Contextual Suicide Risk Assessment in Chinese Group Chats

Abstract

Suicide is a critical global public health challenge, causing approximately 720,000 deaths each year and calling for timely, effective prevention strategies. Existing computational studies primarily focus on post-based social media platforms such as Twitter and Weibo, leaving instant messaging environments such as Telegram underexplored. Yet group chats pose distinct challenges: messages are short, fragmented, multi-party, and often rely on implicit or culturally specific expressions, making isolated post-level analysis insufficient. We introduce **SuiChat-CN**, a Chinese group-chat benchmark for contextual suicide risk assessment. We collect public Telegram group-chat data, construct coherent conversational segments through signal-word extraction and bidirectional context expansion, and annotate user risk levels with an expert-validated, LLM-assisted paradigm. SuiChat-CN contains 13,312 contextual segments from 1,406 users, covering 258,228 raw chat messages. Extensive experiments with PLMs and more than 40 LLMs demonstrate that contextual information is essential for reliable risk assessment, while fine-tuning and partial-context evaluation further reveal the challenges of early detection in multi-party conversations. Due to ethical and sensitivity concerns, the dataset is not publicly released but will be shared with accredited mental health and suicide-prevention research institutions upon reasonable request.

⚠ **Content Warning:** This paper contains examples of harmful language related to suicide and self-harm.

1 Introduction

Mental health disorders are widely recognized as major global public health challenges, with suicide among the most urgent. According to the World Health Organization, approximately 720,000 people die by suicide each year, making it one of the leading causes of death among young people. Early suicide risk detection is therefore essential for enabling timely and targeted intervention.

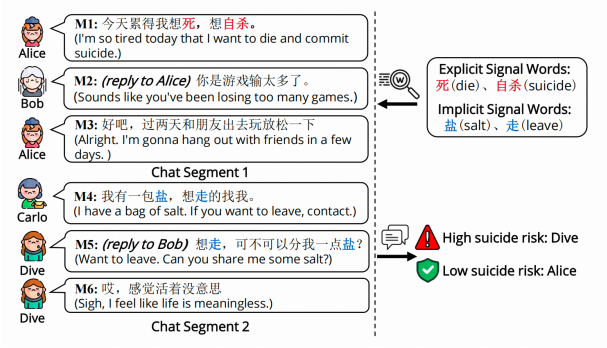


Figure 1: Fragments of two chat segments from a Chinese suicide group with expert-verified risk labels, necessitating contextual information to identify suicidal intent.

The widespread adoption of instant messaging platforms such as Telegram has enabled users to participate in suicide-related discussions within public group chats, including communities centered on suicide support or method sharing. In these spaces, users may disclose suicidal intent, seek validation, or exchange harmful information, making the analysis of group-chat conversations important for identifying warning signs and supporting prevention-oriented research. However, most existing datasets and detection models are built from post-based social media platforms such as Twitter and Weibo. The linguistic and interactional properties of public posts differ substantially from those of group chats, where messages are short, fragmented, multi-party, and highly context-dependent. Accurate risk assessment in this setting therefore requires modeling conversational context rather than isolated messages.

Figure 1 illustrates two chat segments from a public Chinese suicide group, highlighting the indispensable role of context in identifying suicidal intent. In Chat Segment 1, Alice uses explicit and

Lang	Dataset	Source	Type	Classes	Size	Balance	ML	Language
C	BinarygrainedSD(2023)	Weibo	Post	2	1,250	48.12%	✗	Chinese
	FinegrainedSD(2024)	Weibo	Post	10	1,250	48.12%	✓	Chinese
	PsySUICIDE(2024)	Weibo	Post	11	15,010	20.68%	✓	Chinese
NC	Mishra et al.(2019)	Twitter	Post	2	34,306	11.60%	✗	English
	SDCNL(2021)	Reddit	Post	2	1,895	41.53%	✗	English
	DepressEmo(2024)	Reddit	Post	8	6,037	6.80%	✓	English
	Pantaporn et al.(2023)	Twitter	Post	6	5,134	24.46%	✓	Thai
	MentalRiskES(2024)	Telegram	Chat	3	45,000	26.75%	✓	Spanish
C	*SuiChat-CN*(2026)	Telegram	Chat	6	258,228 (13,312)	47.16%	✓	Chinese

Table 1: Comparative overview of typical suicide risk assessment datasets. Abbreviations: **Lang** (Language Type), C (Chinese), NC (Non-Chinese), ML (Multi-Label Classification). The **Balance** column shows the proportion of suicidal ideation, behavior and attempt categories among all categories.

alarming terms such as "die" and "suicide" in M1. In Chat Segment 2, by contrast, Dive uses seemingly benign but contextually risky expressions such as "salt" (referring to drugs) and "leave" (implying death) in M5. Interpreting these messages in isolation would be misleading; a correct assessment requires the surrounding conversational history. In the first case, psychological experts judged Alice as low risk because the utterance reflects transient frustration over a game loss, as clarified by her friend in M2. In the second case, diving is considered high risk because the surrounding context reveals a sustained expression of meaninglessness in life (M6).

This contrast illustrates two core challenges of group-chat suicide risk assessment: risk signals may be explicit or implicit, and their interpretation often depends on multi-turn context. These characteristics weaken the assumptions underlying post-based detection methods. There is therefore a pressing need for context-rich datasets tailored to group-chat environments, as the scarcity of such resources limits progress on realistic suicide risk detection in instant messaging scenarios.

To bridge this gap, we construct **SuiChat-CN**, a Chinese group-chat dataset for contextual suicide risk assessment. We first collect raw chat logs from public Chinese suicide-related Telegram groups. We then use large language models (LLMs) to extract explicit and implicit suicide-related signal words, which are validated by human experts and used to identify suspicious messages and their corresponding at-risk users. For each suspicious message, we perform bidirectional context expansion to form a coherent chat segment. These segments are then annotated through a rigorous two-stage process consisting of pilot annotation and main annotation, ensuring high-quality and consistent labels. The final dataset comprises 13,312 annotated chat segments covering 258,228 raw messages.

Our main contributions are as follows:

- We formulate suicide risk assessment in Chinese group chats as a contextual, multi-turn classification problem and design a construction pipeline based on signal-word discovery, expert validation, and bidirectional context expansion.
- We introduce **SuiChat-CN**, a Chinese group-chat benchmark containing 13,312 contextual segments from 1,406 users, annotated with a six-class suicide risk taxonomy.
- We conduct extensive experiments with PLMs and more than 40 LLMs under prompting, ablation, partial-context, and fine-tuning settings, demonstrating the importance of conversational context for reliable risk assessment.

2 Related Work

2.1 Suicide Risk Taxonomy

Existing studies often adopt binary taxonomies that distinguish low-risk from high-risk content (Mishra et al., 2019; Qi et al., 2025; Sawhney et al., 2018). While simple and easy to annotate, such schemes fail to capture the varying urgency and behavioral specificity of suicidal expressions. Qiu et al. (Qiu et al., 2024) introduced an 11-level taxonomy based on the Columbia-Suicide Severity Rating Scale (C-SSRS) (Posner et al., 2011), but annotation consistency challenges at this granularity led to final evaluation under a binary setting. In contrast, we design a six-class taxonomy that balances clinical

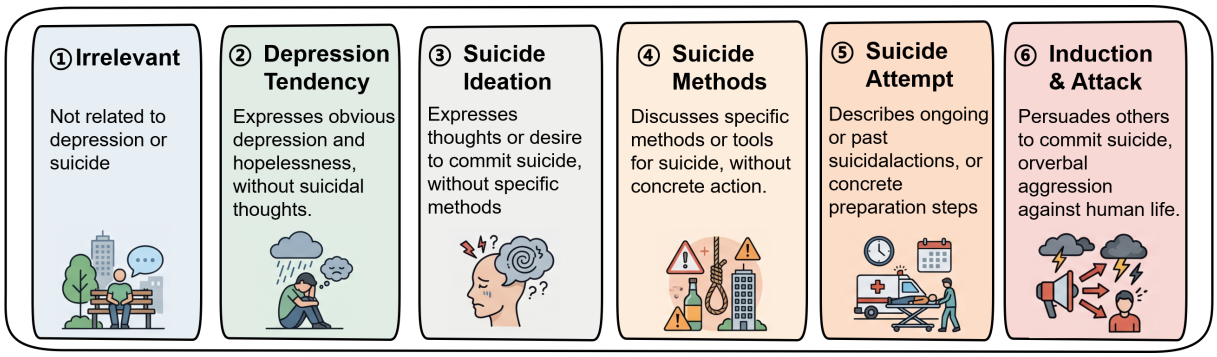


Figure 2: Suicide risk classification taxonomy of the **SuiChat-CN** dataset. Adapted from the C-SSRS and tailored to the linguistic and contextual characteristics of Chinese Telegram group chats, this taxonomy covers a complete risk spectrum from irrelevant content to induction of suicide or verbal aggression. Each level is defined with clear boundaries and accompanied by an illustrative icon.

interpretability, annotation reliability, and practical utility for group-chat risk assessment.

2.2 Suicide Risk Assessment Datasets

Benchmark datasets are critical for developing and evaluating automated suicide risk assessment models. As summarized in Table 1, existing corpora are dominated by isolated posts from public social networks and cover languages such as English, Chinese, and Thai (Benjachairat et al., 2024; Haque et al., 2021; Mishra et al., 2019; Rahman et al., 2024). For Chinese, representative resources include BinarygrainedSD (Qi et al., 2025), its multi-class successor FinegrainedSD (Qi et al., 2024), and PsySUICIDE (Qiu et al., 2024), all derived from platforms such as Weibo or Zhihu. These datasets, however, do not capture the fragmented and multi-party dynamics of group chats. MentalRiskES (Mármol Romero et al., 2024) provides a Spanish Telegram benchmark, but linguistic and cultural differences limit direct transfer to Chinese scenarios. SuiChat-CN addresses this gap by focusing on Chinese group-chat suicide risk assessment.

2.3 Suicide Detection Methods

Suicide risk detection has evolved from classical machine learning with hand-crafted features (Burnap et al., 2015; O'Dea et al., 2015) to deep neural models (Devika et al., 2023; Rahman et al., 2024; Sawhney et al., 2020). More recently, LLMs have shown strong potential due to their general language understanding and instruction-following abilities (Li et al., 2025). Ghanadian et al. (Ghanadian et al., 2023) evaluated ChatGPT under zero-shot and few-shot settings (Shing et al., 2018; Zirikly et al., 2019), Qiu et al. (Qiu et al., 2024) benchmarked LLMs on PsySUICIDE, and other works fine-tuned LLMs on specialized mental health corpora (Menon, 2024; Srivastava, 2025). Despite these advances, suicide risk detection in multi-party group chats remains underexplored, largely because suitable contextual datasets are scarce.

Overall Statistics			
#Samples	13,312		
Avg. Tokens	331.17	Avg. Turns	15.01
Avg. Keyword Count	7.10	Avg. Characters	451.46
Class Distribution (%)			
Irrelevant	35.26	Methods	26.08
Depression	6.44	Attempt	5.85
Ideation	15.25	Induction	11.12
Per-Class Average (Turns/Tokens/Characters)			
Class 0	14.29/270.52/339.57	Class 3	15.08/352.33/520.55
Class 1	15.36/313.79/402.66	Class 4	17.46/407.33/586.37
Class 2	15.72/340.93/461.71	Class 5	14.69/430.50/587.48

Table 2: Statistics of **SuiChat-CN** Dataset.

3 SuiChat-CN Dataset

3.1 An Overview of SuiChat-CN

We introduce **SuiChat-CN**, a Chinese group-chat dataset for contextual suicide risk assessment. The dataset is collected from 20 public Chinese suicide-related Telegram groups, covers chat logs from January 2020 to June 2025, and contains 258,228 raw messages from 1,406 anonymized users.

Using signal-word extraction and bidirectional context expansion, we construct 13,312 annotated contextual segments. Each segment contains, on average, 19.4 messages, 535 Chinese characters, and 4.6 participants. Building on the Columbia-Suicide Severity Rating Scale (C-SSRS) and adapting it to Chinese group-chat discourse, we define

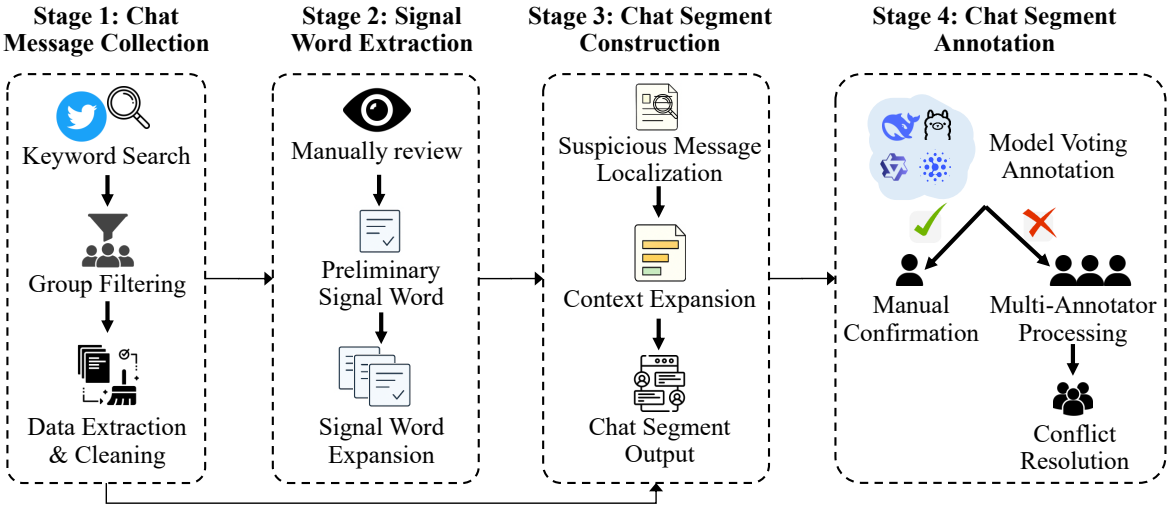


Figure 3: Overview of the *SuiChat-CN* dataset construction pipeline.

six fine-grained risk categories: no risk, depressive emotion, suicidal ideation, suicidal behavior, suicide attempt, and suicide induction or verbal aggression. High-risk samples (ideation, behavior, and attempt) account for 47.16% of the dataset. Detailed statistics are shown in Table 2.

3.2 Dataset Construction Pipeline

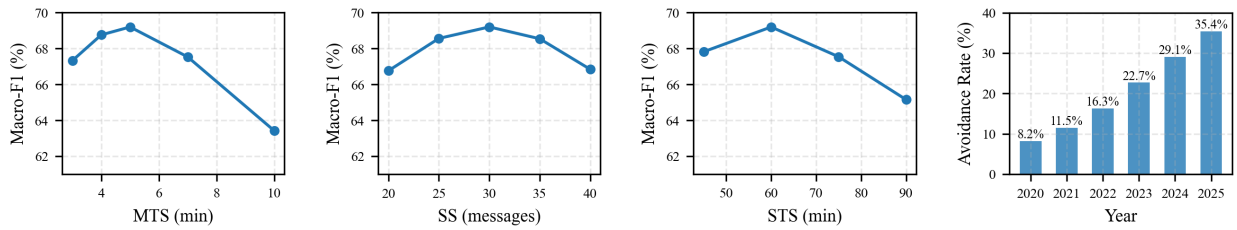
The construction of SuiChat-CN follows the four-stage pipeline illustrated in Figure 3, spanning raw data collection, signal lexicon construction, context expansion, and final annotation.

Compliant Data Collection and Structured Preprocessing. We begin by querying suicide-related Chinese keywords on platforms such as X and Reddit and extracting Telegram group invitation links from public posts and comments. We retain 20 public Chinese suicide-related groups that are active, topically relevant, and accessible without identity verification. Data collection is conducted with the Telethon library in accordance with Telegram's terms of service and data usage policies. To protect user privacy, all user IDs, group IDs, and potentially identifying information are irreversibly anonymized using SHA-256 hashing. The raw data is then converted into a standardized five-tuple format: message_id, user_id, reply_id, timestamp, and content. Preprocessing includes converting Traditional Chinese to Simplified Chinese, removing messages shorter than three characters, and filtering emojis and hyperlinks. This process yields 258,228 clean and structured chat records.

Hierarchical Signal Lexicon Construction with Expert Validation. To address the implicit nature of suicide-related expressions in Chinese group chats, we construct a hierarchical signal lexicon covering both explicit and implicit signals. Explicit signals include direct suicide-related terms, suicide methods and substances, and expressions of nihilism. Implicit signals include Chinese internet abbreviations, homophones, slang, and euphemistic expressions. Because implicit signals are difficult to enumerate manually, we use GLM4-32B-0414 with structured chain-of-thought prompting and few-shot examples to extract candidate signal words from real chat data. All candidate entries are cross-validated by two psychology experts, and false positives are removed. The final lexicon con-

Algorithm 1 Chat Segment Construction

Input: Initial seed m, message list M.
Params: MTS: max gap; SS: max size; STS: max span.
$S \leftarrow \{m\}, i \leftarrow \text{Find}(m, M), L, R \leftarrow i, i$
while $|S| < SS$ **and** $\text{Span}(S) < STS$ **do**
  $changed \leftarrow$ false
  if $L > 0$ **and** $\text{Diff}(M[L-1], M[L]) < MTS$ **then**
    $S \leftarrow S \cup \{M[L-1]\}$, $L \leftarrow L - 1, changed \leftarrow$ true
  end if
  if $R < |M| - 1$ **and** $\text{Diff}(M[R], M[R+1]) < MTS$ **then**
    $S \leftarrow S \cup \{M[R+1]\}$, $R \leftarrow R + 1, changed \leftarrow$ true
  end if
  $reps \leftarrow \{M[msg.RID] \mid msg \in S, msg.RID \neq \text{null}, M[msg.RID] \notin S\}$
  if $reps \neq \emptyset$ **then**
    $S \leftarrow S \cup reps, changed \leftarrow$ true
  end if
  if not $changed$ **then break**
end while
return S



(a) Effect of MTS (b) Effect of SS (c) Effect of STS (d) Effect of XYZ

Figure 4: Parameter sensitivity analysis of our proposed method.

tains 3,508 entries, including 2,142 explicit signals and 1,366 implicit signals.

Suspicious Message Identification and Bidirectional Context Expansion. Using the signal lexicon, we identify suspicious messages containing risk-related signals and mark their senders as potential at-risk users. To overcome the limitations of single-message analysis, we propose a bidirectional context expansion algorithm with temporal constraints. Centered on each suspicious message, the algorithm expands both backward and forward to collect related messages and also retrieves referenced historical messages through reply links to preserve conversational completeness. After merging overlapping segments and filtering extreme-length samples, we obtain 13,312 candidate segments. Each segment is represented as a four-tuple: chat_segment, suspicious_message, at_risk_user, and risk_level. The resulting segments contain an average of 19.4 messages and 4.6 participants.

Two-Stage Annotation and End-to-End Quality Control. We adopt a three-step annotation paradigm consisting of pilot annotation, main annotation, and conflict resolution to ensure label consistency and reliability. In the pilot stage, three trained graduate students in psychology independently annotate 300 samples. Through three rounds of guideline refinement, inter-annotator agreement (Fleiss' κ) improves from 0.57 to 0.73, reaching substantial agreement. In the main annotation stage, the remaining samples are divided into four batches, with 10% of samples randomly selected for periodic review. The κ score remains stable between 0.84 and 0.88, indicating near-perfect agreement. For samples with annotation disagreements, we apply majority voting; ambiguous samples that fail to reach consensus are discarded.

3.3 SuiChat-CN Analysis

This section analyzes SuiChat-CN from three complementary perspectives: context-construction hyperparameters, the signal-word framework, and the agreement structure of multi-model voting annotation. Supplementary analyses are provided in Appendix F and Appendix J.

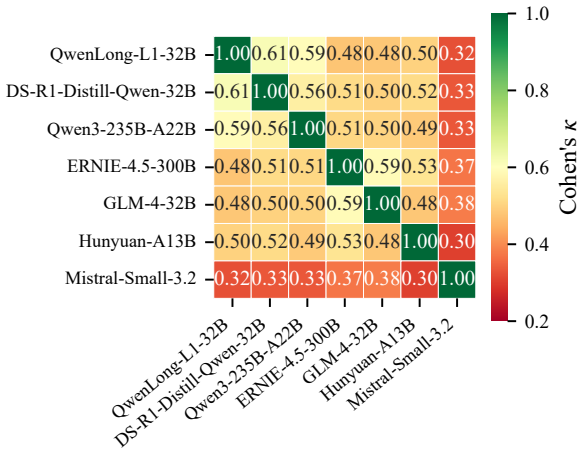


Figure 5: Pairwise inter-model agreement (Cohen's κ) among a representative subset of voter models used in the silver-annotation pipeline.

Hyperparameter selection for Data Construction. We decouple *context construction* from *formal annotation* and adopt a two-stage pipeline for hyperparameter selection and quality control. We first sample 50 instances as a development set, which are independently annotated by three annotators, with disagreements resolved through discussion. Based on this set, we select the hyperparameters for bidirectional context expansion by reconstructing contexts under each candidate configuration and evaluating the zero-shot detection performance of Qwen-32B using Macro-F1. The search space contains 100 configurations ($5\times5\times4$): MTS$\in$ $\{3, 4, 5, 7, 10\}$, SS$\in$ $\{20, 25, 30, 35, 40\}$, and STS$\in$ $\{45, 60, 75, 90\}$. The optimal configuration is MTS=5min, SS=30, and STS=60min, which we fix for full-scale context construction. Figure 4 presents the single-variable sensitivity curves of the key hyperparameters.

Model	Standard			Few Shot (FS)			w/o Keywords			w/o Context		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Proprietary (Closed-Source) Models												
GPT-5	72.7	70.5	70.7	73.1	71.2	72.0	71.8	69.4	70.1	71.3	68.6	69.2
GPT-4o-Mini	67.7	64.3	64.2	68.3	64.8	65.7	67.2	64.1	65.1	62.3	59.8	60.8
Gemini-2.5-Pro	76.7	75.7	75.9	78.4	76.5	76.4	75.8	74.3	74.5	74.6	73.1	73.4
Qwen3.5-Plus	76.9	73.9	74.2	78.9	73.7	73.3	77.3	75.1	75.1	74.9	72.3	72.2
Qwen-Max	74.4	69.2	68.6	77.2	73.9	73.6	76.5	73.6	73.1	74.2	68.6	68.0
Seed-OSS-36B-Instruct	76.8	74.2	74.5	80.2	77.1	77.3	76.2	74.3	74.2	72.7	70.7	71.3
DeepSeek-V3	74.3	70.8	70.6	79.7	75.3	75.1	70.2	68.8	68.9	73.1	71.5	70.7
GLM-5.1	74.6	69.8	69.5	77.5	71.1	71.1	77.3	75.1	75.1	74.0	71.6	72.3
Kimi-K2-Instruct	74.1	80.0	76.3	77.4	72.9	72.8	76.2	74.6	74.7	76.1	71.0	72.4
Hunyuan-A13B-Instruct	55.4	54.2	53.4	69.3	61.9	61.9	65.0	59.9	58.0	54.4	54.6	52.9
MiniMax-M2.7	55.4	51.0	50.1	69.8	63.4	62.8	60.8	57.5	56.8	63.9	59.0	57.9
Open-Weight Models												
Llama-3.3-70B	70.4	73.9	71.6	73.8	70.5	71.2	72.1	69.4	69.8	68.7	71.3	68.9
Llama3-8B	51.7	26.9	23.3	70.0	33.4	32.5	58.7	24.7	18.5	54.2	28.1	24.9
gemma-3-27b-it	51.9	73.9	49.2	60.4	64.2	58.6	56.3	60.7	54.8	52.7	62.3	51.4
Qwen3-32B	69.0	72.0	70.2	69.7	63.8	62.5	67.4	61.2	60.4	67.3	72.6	67.9
Qwen3-8B	65.1	55.8	55.3	70.8	62.6	64.0	69.4	64.5	64.8	71.5	63.0	64.8
DeepSeek-R1	68.9	75.2	71.5	76.3	76.0	74.1	75.4	77.6	76.3	70.4	73.8	71.1
DeepSeek-R1-Distill-Qwen3-8B	68.0	61.4	60.0	74.2	63.5	65.7	69.7	59.3	58.9	71.8	65.1	65.5
Phi4	72.1	56.6	59.0	73.2	59.0	61.6	72.3	53.0	55.5	72.1	57.0	59.1
mistral-small-3.2	56.6	72.7	58.1	64.3	63.5	62.1	60.7	61.4	59.6	55.9	60.4	56.8
GLM4-Z1-9B	65.2	60.8	58.4	68.4	67.1	65.8	70.4	65.1	65.5	68.0	63.6	63.2

Table 3: Performance Comparison of Representative Models Across Different Scenarios.

Signal-word framework analysis. Unlike traditional approaches that rely on explicit keywords, high-risk users in Telegram group chats frequently express suicidal intent through implicit expressions such as metaphors, homophones. For the analysis below, we focus on a curated subset of 442 validated signal words, including 52 explicit and 390 implicit signals spanning six categories.

Statistical analysis shows that implicit signals appear $2.3\times$ more frequently than explicit signals in high-risk samples (Levels 2–4), while 37.6% of high-risk samples contain only implicit signals. Further analysis reveals strong cultural specificity in Chinese group-chat risk expressions, including references to non-controlled drugs and evasive variants based on pinyin or homophones. These findings highlight the importance of modeling implicit, culturally grounded, and evolving risk signals for realistic suicide risk detection.

Agreement structure of multi-model voting annotation. To examine the reliability of the silver-annotation process, we recruit $M = 17$ open-weight LLMs to annotate 11,189 multi-turn topics under the same six-class risk schema. A topic is accepted as a silver-labelled sample only when the weighted consensus ratio reaches $\rho \geq 2/3$, yielding 7,148 accepted samples and a 63.9% acceptance rate. The whole-pool Fleiss' κ is 0.422, indicating moderate inter-model agreement, while the remaining 36.1% no-consensus cases are excluded from training and evaluation. Beyond these aggregate statistics, Figure 5 presents a subset of the pairwise Cohen's κ heatmap to illustrate the structure of model agreement. The heatmap reveals a clear block pattern: models from similar families tend to cluster together, while cross-family agreement remains consistently above chance. This suggests that the final silver labels are not dominated by a single model lineage, but are instead supported by complementary decision boundaries across diverse backbones.

4 Evaluation

4.1 Experiment Setup

We evaluate more than 40 LLMs, including both proprietary API-based systems and open-weight models. These models span multiple families, including GPT, Gemini, GLM, DeepSeek, Kimi, and MiniMax, and cover a broad range of parameter scales. We also evaluate classic PLMs as supervised baselines.

Our experiments cover four scenarios: zero-shot standard evaluation, few-shot standard evaluation with three expert-validated examples per risk level,

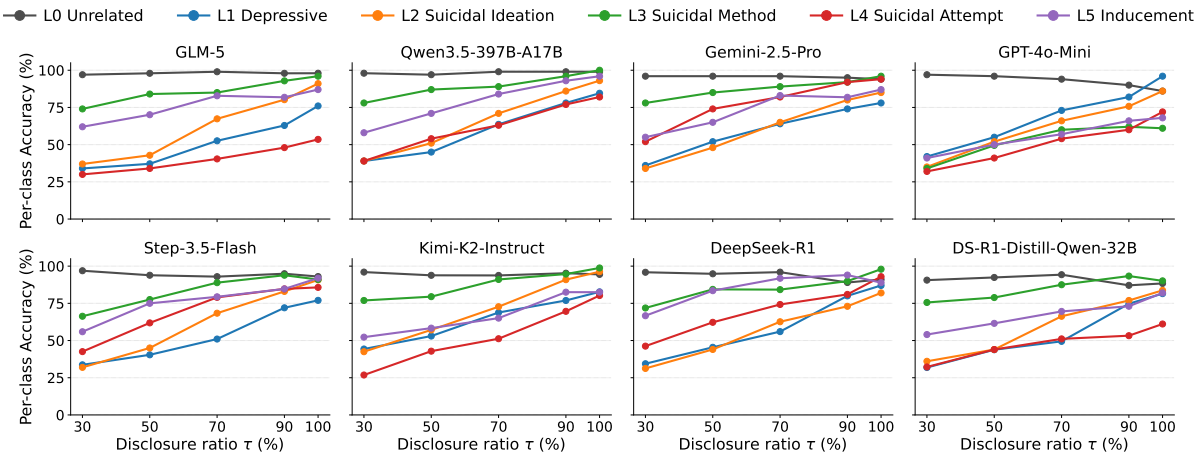


Figure 6: Per-class accuracy trajectories under partial context disclosure for eight representative models.

few-shot evaluation with all suicide-related signal words masked, and few-shot evaluation with only the target user's suspicious message retained. We use macro-averaged precision, recall, and F1-score as primary metrics.

4.2 Main Results

The main results are reported in Table 3. For clarity, we present a subset of representative models.

Closed-source proprietary models consistently outperform open-weight models, with leading models demonstrating superior risk detection capabilities across scenarios. As shown in Table 3, proprietary models generally achieve F1-scores above 70% across settings, outperforming most open-weight counterparts. Top-performing systems such as Seed-OSS-36B-Instruct, Gemini-2.5-Pro, and Kimi-K2-Instruct rank near the top in different scenarios. Among open-weight models, only DeepSeek-R1 and Llama-3.3-70B approach the performance of mid-tier proprietary systems, while smaller open models exhibit substantial gaps.

The few-shot setting notably boosts performance for most models, particularly in improving precision. Comparing zero-shot and few-shot results, nearly all models show a clear precision increase after labeled exemplars are introduced. For example, Seed-OSS-36B-Instruct improves from 76.8% to 80.2% precision, leading to a corresponding F1 gain. Although a few models show minor recall fluctuations, few-shot prompting remains an effective strategy for reducing false positives and improving overall detection performance.

Dialogue context and suicide-related keywords are critical for risk detection, and their absence universally degrades model performance. Ablation studies show that removing dialogue context or masking suicide-related keywords leads to widespread performance degradation. Context removal has the stronger effect: most models exhibit notable F1 declines because isolated messages fail to capture implicit and multi-turn risk signals. Only a few models, such as DeepSeek-R1, remain robust under the keyword-masked setting, whereas many models suffer recall drops, indicating reliance on explicit surface cues.

4.3 Per-Class Accuracy Analysis

Table 5 shows that class-wise performance varies substantially even among models with similar overall few-shot accuracy, revealing clear specialization patterns. GLM-5.1 and QWEN3.5-PLUS are especially strong on structurally easier classes such as L0 and L3, reaching 92.0% and 94.0%, respectively, whereas KIMI-K2 performs best on L2 (86.0%), suggesting an advantage in identifying suicidal ideation before explicit action cues appear. In contrast, DEEPSEEK-R1 achieves the best accuracy on the most severe classes, L4 (81.8%) and L5 (84.2%), indicating stronger robustness on suicidal attempt and inducement cases. These results show that models with comparable micro-averaged accuracy can behave quite differently across risk levels: some are more effective at filtering low-risk or method-related content, while others are more reliable on the highest-risk categories that matter most for intervention.

4.4 Context Sensitivity under Partial Disclosure

To understand how different risk categories depend on conversational context, we further evaluate mod-

Model	Standard	Few Shot (FS)	FS w/o Keywords	FS w/o Context
GLM4-Z1-9B	58.4 → 86.8 (+28.4)	65.8 → 86.9 (+21.1)	65.5 → 82.0 (+16.5)	63.2 → 80.6 (+17.4)
Gemma-2-9B	48.2 → 88.2 (+40.0)	56.3 → 83.2 (+26.9)	33.2 → 85.1 (+51.9)	43.3 → 74.4 (+31.1)
Qwen3-32B	70.2 → 86.5 (+16.3)	62.5 → 84.2 (+21.7)	60.4 → 81.6 (+21.2)	67.9 → 74.9 (+7.0)
Llama3-8B	23.3 → 82.7 (+59.4)	32.5 → 79.2 (+46.7)	18.5 → 80.1 (+61.6)	24.9 → 73.5 (+48.6)
DeepSeek-R1-Distill-Qwen3-8B	60.0 → 79.5 (+19.5)	65.7 → 80.5 (+14.8)	58.9 → 73.9 (+15.0)	65.5 → 73.4 (+7.9)
Mistral-Small-3.2	58.1 → 62.6 (+4.5)	62.1 → 64.1 (+2.0)	59.6 → 64.5 (+4.9)	56.8 → 61.2 (+4.4)
Phi-4	59.0 → 80.8 (+21.8)	61.6 → 81.7 (+20.1)	55.5 → 79.1 (+23.6)	59.1 → 78.9 (+19.8)

Table 4: F1 comparison before and after supervised fine-tuning. Each cell reports prompt-only F1 → fine-tuned F1, with the absolute change in parentheses.

Model	L0	L1	L2	L3	L4	L5	Acc.
DeepSeek-V3	72.7	**79.9**	84.8	92.1	44.8	77.4	**75.5**
DeepSeek-R1	75.0	50.0	76.9	88.2	**81.8**	**84.2**	74.8
Qwen-Max	77.8	71.1	84.3	88.2	45.5	75.0	73.6
Qwen3.5-Plus	89.9	61.7	76.9	**94.0**	47.7	71.8	73.6
Kimi-K2	77.2	65.5	**86.0**	91.7	58.2	58.9	72.8
GLM-5.1	**92.0**	42.2	67.3	89.9	59.1	76.4	71.4

Table 5: Per-class few-shot accuracy (%) of six representative models.

els under partial disclosure, where only the first $\tau \in \{30, 50, 70, 90, 100\}\%$ of each group-chat dialogue is visible. Figure 6 visualizes the per-class accuracy trajectories of eight representative models as the available context gradually increases. The figure shows that context sensitivity is highly class-dependent. L0, corresponding to non-risk or irrelevant cases, remains consistently easy to recognize even at τ=30%, indicating that benign keyword matches can often be filtered with limited context. By contrast, L1 and L2 exhibit much steeper upward trends, suggesting that depressive expressions and suicidal ideation require evidence accumulated across multiple turns. The most important pattern appears for L4: despite being the highest-risk class, suicidal-attempt cases are poorly recognized under early disclosure and become substantially more detectable only after most of the conversation is observed. This finding challenges the assumption that urgent suicide-risk cues always appear at the beginning of a dialogue. Instead, in group-chat settings, high-risk evidence is often delayed, implicit, or distributed across interactions, making long-context modeling and continuous monitoring essential for reliable early-warning systems.

4.5 Fine-Tuning Performance

To assess whether task-specific supervision can further improve models on SuiChat-CN, we select a subset of representative model, and fine-tune them under the same evaluation scenarios.

Table 4 compares prompt-only and fine-tuned F1 scores using aligned backbone names. Supervised fine-tuning brings substantial gains for most open-weight models, especially smaller or weaker prompt-only backbones. For example, Llama3-8B improves by +59.4 F1 points in the Standard setting and by +61.6 points when keywords are masked, while Gemma-2-9B gains +40.0 and +51.9 points under the same two settings. The DeepSeek row corresponds to DeepSeek-R1-Distill-Qwen3-8B and also benefits consistently from fine-tuning, although its gains are smaller than those of Llama3-8B and Gemma-2-9B. Overall, fine-tuning greatly narrows the gap between open-weight models and stronger proprietary systems, but the remaining drop under the context-removed setting shows that task-specific training cannot replace multi-turn conversational evidence.

5 Conclusion

We introduce **SuiChat-CN**, a Chinese group-chat benchmark for contextual suicide risk assessment. Unlike prior datasets built from isolated social media posts, SuiChat-CN targets multi-party instant messaging scenarios where risk cues are implicit, fragmented, and distributed across turns. Constructed via signal-word extraction, bidirectional context expansion, and expert-validated annotation, the dataset contains 13,312 contextual segments from 1,406 users and 258,228 messages. Experiments with PLMs and over 40 LLMs demonstrate that contextual modeling is critical: removing dialogue context degrades performance, and high-risk cues emerge only after multiple turns. Fine-tuning improves open-weight models, especially prompt-only backbones, but does not eliminate the need for evidence. SuiChat-CN aims to support work on realistic, context-aware, and responsible suicide risk assessment in group-chat.

Limitations

This work has several limitations. First, SuiChat-CN is built from public Chinese Telegram groups; therefore, its linguistic patterns, platform norms, and community dynamics may not fully generalize to private conversations, other languages, or other social platforms. Second, although our context-construction strategy captures multi-turn interactions, it does not explicitly model real-time message latency or long-term user history. Third, due to the sensitivity of suicide-related data, the dataset cannot be openly released, which may limit immediate reproducibility. We will instead provide controlled access to qualified research institutions under strict data-use agreements.

Ethics Statement

This study involves sensitive content related to suicide and self-harm. We adopt several safeguards to reduce potential harm and protect user privacy. All data are collected only from publicly accessible Telegram groups and anonymized before analysis by removing or replacing user identifiers, group identifiers, and other potentially identifying information. Annotators are informed of the sensitive nature of the task, and safeguards are used to reduce psychological burden during annotation. The dataset will not be publicly released; access will be restricted to accredited mental-health and suicide-prevention research institutions that agree to responsible use, non-disclosure, and ethical data-handling requirements.

References

Pantaporn Benjachairat, Twittie Senivongse, Nattasuda Taephant, Jiratchaya Puvapaisankit, Chonlakorn Maturosjamnan, and Thanakorn Kultananawat. 2024. Classification of suicidal ideation severity from twitter messages using machine learning. *International Journal of Information Management Data Insights*, 4(2):100280.

Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM Conference on Hypertext and Social Media*, pages 75–84. ACM.

SP Devika, MR Pooja, MS Arpitha, and Ravi Vinayakumar. 2023. Bert-based approach for suicide and depression identification. In *Proceedings of 3rd International Conference on Advances in Computer Engineering and Communication Systems (ICACECS)*, pages 435–444. Springer.

Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. 2023. Chatgpt for suicide risk assessment on social media: Quantitative evaluation of model performance, potentials and limitations. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 172–182.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. 2021. Deep learning for suicide and depression identification with unsupervised label correction. Berlin, Heidelberg. Springer-Verlag.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Tong Li, Shu Yang, Junchao Wu, Jiyao Wei, Lijie Hu, Mengdi Li, Derek F Wong, Joshua R Oltmanns, and Di Wang. 2025. Can large language models identify implicit suicidal ideation? an empirical evaluation. arXiv preprint arXiv:2502.17899.

GR Menon. 2024. Mental health mistral 7b instructv0.2 finetuned v2. https://huggingface.co/GRMenon/mental-health-mistral-7b-instructv0.2finetuned-V2.

Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. 2019. SNAP-BATNET: Cascading author profiling and social network graphs for suicide ideation detection on social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 147–156, Minneapolis, Minnesota. Association for Computational Linguistics.

Alba María Mármol Romero, Adrián Moreno-Muñoz, Flor Miriam Plaza-Del-Arco, M Dolores Molina-González, and Arturo Montejo-Ráez. 2024. Mentalriskes: A new corpus for early detection of mental disorders in spanish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 11204–11214.

Bridianne O'Dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

Kelly Posner, Gregory Brown, Barbara Stanley, Kseniya Yershova, Maria Oquendo, Glenn Currier, Glenn Melvin, Laurence Greenhill, Sa Shen, and J. Mann. 2011. The columbia-suicide severity rating scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *The American journal of psychiatry*, 168:1266–77.

Hongzhi Qi, Guanghui Fu, Jianqiang Li, Changwei Song, Wei Zhai, Dan Luo, Shuo Liu, Yijing Yu, Bingxiang Yang, and Qing Zhao. 2025. Supervised learning and large language model benchmarks on

mental health datasets: Cognitive distortions and suicidal risks in chinese social media. *Bioengineering*, 12(8).

Hongzhi Qi, Hanfei Liu, Jianqiang Li, Qing Zhao, Wei Zhai, Dan Luo, Tian He, Shuo Liu, Bing Yang, and Guanghui Fu. 2024. Sos-1k: A fine-grained suicide risk classification dataset for chinese social media analysis. pages 3781–3786.

Huachuan Qiu, Lizhi Ma, and Zhenzhong Lan. 2024. Psyguard: An automated system for suicide detection and risk assessment in psychological counseling.

Abu Bakar Siddiqur Rahman, Hoang-Thang Ta, Lotfollah Najjar, Azad Azadmanesh, and Ali Saffet Gönul. 2024. Depressionemo: A novel dataset for multilabel classification of depression emotions. *Journal of Affective Disorders*, 366:445–458.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7685–7697.

Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal. 2018. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 91–98, Melbourne, Australia. Association for Computational Linguistics.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36.

Saurav Kumar Srivastava. 2025. Gemma-3 270m mental health fine-tuned model. `https://huggingface.co/Skshackster/gemma3-270m-mental-health-finetuned-gguf`.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33.

A Prompt for Signal Word Extraction

Figure 7 outlines the templates for prompting methods that utilize chain-of-thought and few-shot learning with Large Language Models in the signal word extraction task. The chain-of-thought approach enables the model to perform step-by-step reasoning, while the few-shot learning component provides annotated examples to guide the model effectively

B Prompt for LLM Suicide Risk Annotation

Figure 8 illustrates the Chain-of-Thought prompt designed for suicide risk annotation using Large Language Models. This prompt operates through a structured three-step process. First, the model reads the chat context to comprehend the conversation and identifies the specific message containing the signal word. Second, based on our risk taxonomy, the model analyzes the target user's statements as well as the overall chat context to categorize their suicide risk level. However, in a group chat conversation, multiple risk categories may be present. Therefore, we apply priority rules so that if multiple risk categories exist, the model assigns the highest risk category as the annotation.

C LLM Details

Table 8 provides a comprehensive specification of the LLMs employed in our evaluation. To ensure robust representativeness across different architectures and parameter scales, our selection integrates both closed-source commercial systems and open-weight alternatives. We classify the evaluated models into two distinct categories based on their accessibility.

- **API Models**: This category encompasses state-of-the-art closed-source models, including offerings from OpenAI, Anthropic, Google, Grok, and so on. These models are accessed via cloud-based inference endpoints provided by their respective developers.
- **Open-Weight Models**: This category features high-performance models with publicly available weights, such as Qwen and DeepSeek. These models allow for local deployment and independent verification, serving as baselines for open-source capability.

D Annotation Platform and Consensus Review System

We developed an online annotation platform and a consensus-based review system to support the construction of high-quality evaluation data. As shown in Figure 9, the platform provides a standardized workflow for task assignment, independent annotation, disagreement detection, expert review, and final label confirmation. This design ensures that

Table 6: Complete Performance Comparison of All Models Across Different Scenarios. Metrics are Precision (%), Recall (%), and F1 (%).

Model	Standard			Few Shot (FS)			w/o Keywords			w/o Context		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Proprietary (Closed-Source) Models												
Gemini-2.5-Pro	76.7	75.7	75.9	78.4	76.5	76.4	75.8	74.3	74.5	74.6	73.1	73.4
Seed-OSS-36B-Instruct	76.8	74.2	74.5	80.2	77.1	77.3	76.2	74.3	74.2	72.7	70.7	71.3
Qwen3.5-Plus	76.9	73.9	74.2	78.9	73.7	73.3	77.3	75.1	75.1	74.9	72.3	72.2
Qwen-Max	74.4	69.2	68.6	77.2	73.9	73.6	76.5	73.6	73.1	74.2	68.6	68.0
Qwen3.5-397B-A17B	72.4	69.7	69.5	76.2	72.9	72.4	75.0	73.3	72.9	71.8	68.9	68.6
GPT-5	72.7	70.5	70.7	73.1	71.2	72.0	71.8	69.4	70.1	71.3	68.6	69.2
Kimi-K2-Instruct	74.1	80.0	76.3	77.4	72.9	72.8	76.2	74.6	74.7	76.1	71.0	72.4
DeepSeek-V3	74.3	70.8	70.6	79.7	75.3	75.1	70.2	68.8	68.9	73.1	71.5	70.7
GLM-5.1	74.6	69.8	69.5	77.5	71.1	71.1	77.3	75.1	75.1	74.0	71.6	72.3
DeepSeek-V4-Pro	65.6	64.9	65.0	71.3	68.4	68.2	69.7	69.1	68.7	66.3	67.5	66.7
Doubao-Seed-2.0-pro	71.6	59.9	59.5	77.3	70.6	70.6	73.7	64.9	64.7	70.2	60.2	64.5
GLM-5	70.8	65.6	65.5	74.2	67.7	67.2	73.0	70.1	69.9	72.8	67.0	66.6
GLM-4.7	70.4	63.3	63.4	71.9	66.9	66.4	70.4	65.0	65.1	72.7	65.2	67.3
Kimi-K2.5	75.8	70.9	71.1	75.3	73.6	73.4	72.2	69.4	69.3	76.4	73.0	73.0
Kimi-K2.6	72.7	69.6	69.4	75.9	72.1	71.9	72.4	68.9	68.6	73.5	76.2	73.3
GPT-4o-Mini	67.7	64.3	64.2	68.3	64.8	65.7	67.2	64.1	65.1	62.3	59.8	60.8
DeepSeek-V4-Flash	64.6	61.3	61.3	68.7	64.3	64.1	63.4	61.8	62.1	62.8	59.8	59.8
DeepSeek-V3.2	63.9	60.6	60.4	74.7	70.1	69.9	70.8	67.7	67.6	64.7	61.8	61.5
MiMo-V2-Flash	70.8	61.5	59.5	71.1	61.0	58.2	67.6	59.8	57.9	70.9	62.1	59.5
MiniMax-M2.7	55.4	51.0	50.1	69.8	63.4	62.8	60.8	57.5	56.8	63.9	59.0	57.9
MiniMax-M2.5	62.3	53.4	52.6	68.0	60.0	59.8	61.7	54.2	53.4	69.2	48.7	51.1
Hunyuan-A13B-Instruct	55.4	54.2	53.4	69.3	61.9	61.9	65.0	59.9	58.0	54.4	54.6	52.9
Qwen3-235B-A22B	73.7	70.6	70.5	72.7	70.9	70.8	71.7	70.6	70.5	71.2	69.3	69.2
GLM-4-32B-0414	73.1	77.2	74.7	75.0	68.3	67.6	71.9	71.8	71.6	71.4	69.1	70.8
QwenLong-L1-32B	72.8	78.6	75.2	75.6	73.4	73.1	73.9	71.8	71.4	70.5	73.2	70.7
GLM-4.6	71.9	73.9	69.1	74.3	70.2	69.3	71.8	68.9	69.0	70.4	68.1	69.1
Ring-flash-2.0	72.1	69.4	68.7	76.7	76.9	76.0	70.3	67.0	66.4	58.5	68.4	57.8
Open-Weight Models												
DeepSeek-R1	68.9	75.2	71.5	76.3	76.0	74.1	75.4	77.6	76.3	70.4	73.8	71.1
Llama-3.3-70B	70.4	73.9	71.6	73.8	70.5	71.2	72.1	69.4	69.8	68.7	71.3	68.9
Qwen3-32B	69.0	72.0	70.2	69.7	63.8	62.5	67.4	61.2	60.4	67.3	72.6	67.9
Phi4	72.1	56.6	59.0	73.2	59.0	61.6	72.3	53.0	55.5	72.1	57.0	59.1
Qwen3-8B	65.1	55.8	55.3	70.8	62.6	64.0	69.4	64.5	64.8	71.5	63.0	64.8
GLM4-Z1-9B	65.2	60.8	58.4	68.4	67.1	65.8	70.4	65.1	65.5	68.0	63.6	63.2
DeepSeek-R1-Distill-Qwen3-8B	68.0	61.4	60.0	74.2	63.5	65.7	69.7	59.3	58.9	71.8	65.1	65.5
DeepSeek-R1-Distill-Llama-8B	52.6	46.4	43.7	54.8	49.4	49.3	50.4	45.9	43.9	52.6	44.2	42.1
DeepSeek-R1-Distill-Qwen-32B	65.7	73.1	68.9	72.1	69.9	69.5	63.5	60.8	60.4	67.7	65.5	65.3
mistral-small-3.2	56.6	72.7	58.1	64.3	63.5	62.1	60.7	61.4	59.6	55.9	60.4	56.8
gemma-3-27b-it	51.9	73.9	49.2	60.4	64.2	58.6	56.3	60.7	54.8	52.7	62.3	51.4
gemma-2-9b-it	45.2	68.4	47.1	55.7	60.3	54.5	51.4	57.8	50.9	48.2	60.5	49.8
Qwen3-30B-A3B-Instruct-2507	69.4	63.5	63.8	73.2	66.7	67.0	70.6	67.3	66.5	58.7	66.4	56.8
Ling-flash-2.0	61.3	55.7	54.9	64.7	58.5	57.9	60.4	56.0	55.1	58.2	55.9	52.9
Gemma2	71.5	46.9	48.2	69.4	54.7	56.3	70.3	33.2	33.2	69.8	42.0	43.3
Llama3-8B	51.7	26.9	23.3	70.0	33.4	32.5	58.7	24.7	18.5	54.2	28.1	24.9
Mistral-7B-Instruct	56.8	27.7	24.2	64.4	36.8	35.0	60.3	38.0	37.2	57.8	27.9	24.5

Chain-of-Thought and Few-Shot Learning Prompt for Signal Word Extraction

你是一个中文分析专家，任务是从给定聊天中识别与自杀或抑郁情绪相关的显式或隐喻信号词，提取真实意图、原文自杀抑郁信号词和表达动因。表达动因需尽量使用简洁明了的方式表达合理意义。按以下步骤推理，以 JSON 格式输出三元组，形式为 < 真实意图, 原文自杀抑郁信号词, 表达动因 >:

(You are a Chinese language analysis expert. Your task is to identify explicit or implicit signal words related to suicide or depressive emotions from a given conversation. Extract the true intention, the original suicide or depression-related signal words from the text, and the underlying motive for the expression. The motive should be expressed clearly and concisely, using reasonable metaphors where applicable. Follow the reasoning steps below and output a triplet result in JSON format in the form of <True Intention, Original Suicide/Depression Signal Words, Underlying Motive>.)

推理步骤：

1. 分析聊天语义：判断是否表达自杀或抑郁情绪，关注显式和隐喻表达。
2. 识别信号词：提取与自杀或抑郁相关的显式或隐喻信号词，分析语义。
3. 确定真实意图：识别显式或隐喻表达指代的真实对象或概念。
4. 推导表达动因：描述真实意图与原文自杀抑郁信号词之间的语义联系，使用简洁明了的方式表达动因。
5. 输出：JSON 格式三元组。

(Reasoning Steps:)

1. Analyze the conversation semantics: Determine whether it expresses suicidal or depressive emotions, focusing on both explicit and implicit expressions.
2. Identify signal words: Extract explicit or implicit signal words related to suicide or depression and analyze their meanings.
3. Determine the true intention: Identify the actual object or concept referenced by the explicit or implicit expressions.
4. Infer the underlying motive: Describe the semantic connection between the true intention and the original suicide or depression-related signal words using a clear and concise explanation.
5. Output: Return a single triplet in JSON format.

少样本示例：

- 输入：谁有很强的氰化钾卖，输出：< 自杀药物，氰化钾，自杀方式 >
- 输入：跳进岩浆什么感觉，输出：< 自杀方式，跳进岩浆，自杀意图 >
- 输入：我还想有人打死我勒，输出：< 自杀方式，打死我，自杀意图 >
- 输入：十楼以上几乎百分百必 si，输出：< 自杀结果，si，汉语拼音隐晦表达自杀 >
- 输入：一点也不像紫砂的人，输出：< 自杀行为，紫砂，汉语谐音隐晦表达自杀 >

(Few-shot Examples:)

- Input: Who has potent potassium cyanide for sale, Output: <Suicide method, cyanide, Suicide means>
- Input: What does jump into magma feel like, Output: <Suicide method, jump into magma, Suicide intention>
- Input: I still wish someone would kill me, Output: <Suicide method, kill me, Suicide intention>
- Input: Almost 100% sure to die on floors above the tenth, Output: <Suicide outcome, die, Pinyin for "死" (death) as a euphemism for suicide>
- Input: Not at all like someone who loves Zisha, Output: <Suicide behavior, Zisha, Homophonic euphemism for suicide in Chinese>

Figure 7: Chain-of-thought and few-shot learning prompt for signal word extraction.

each sample is annotated in a traceable and verifiable manner.

- **Online Annotation Platform**: This component provides a unified web-based interface for annotators. Each task contains the source text, annotation guidelines, predefined label options, and optional evidence or rationale fields. The platform records annotation results, annotator information, timestamps, task status, and revision history, allowing the annotation process to be systematically tracked and audited.

- **Consensus Review System**: This component is used to resolve annotation disagreements and improve label reliability. As shown in Figure 10, multiple annotators first complete the same task independently. The system then compares their labels and automatically separates agreement cases from disagreement cases. Samples with inconsistent annotations are submitted to a review stage, where senior annotators or domain experts determine the final label based on the original text, candidate labels, and annotator rationales.

The consensus process follows a structured procedure. First, each annotator completes the assigned task independently without access to other

Table 7: Overall per-class few-shot accuracy (%). Each cell is class-conditional accuracy on successfully parsed predictions; the rightmost column is micro-averaged accuracy. Best per column in **bold**.

Model	L0	L1	L2	L3	L4	L5	Acc.
STEP-3.5-FLASH	80.9	64.7	76.1	89.0	73.2	71.4	**75.9**
DEEPSEEK-V3	72.7	**79.9**	84.8	92.1	44.8	77.4	75.5
DEEPSEEK-R1	75.0	50.0	76.9	88.2	**81.8**	**84.2**	74.8
QWEN-MAX	77.8	71.1	84.3	88.2	45.5	75.0	73.6
QWEN3.5-PLUS	89.9	61.7	76.9	**94.0**	47.7	71.8	73.6
QWEN3.5-397B	89.8	62.3	74.4	91.0	48.6	71.5	72.8
KIMI-K2	77.2	65.5	**86.0**	91.7	58.2	58.9	72.8
KIMI-K2.6	88.7	50.0	72.1	87.1	71.3	63.4	72.2
GLM-5.1	**92.0**	42.2	67.3	89.9	59.1	76.4	71.4
DOUBAO-SEED-2.0	90.8	43.9	76.2	85.0	70.9	56.8	70.5
DEEPSEEK-V3.2-EXP	87.2	62.2	77.9	86.0	45.6	61.5	70.4
DS-R1-DISTILL-QWEN-32B	79.0	56.7	77.6	89.0	51.4	65.9	69.9
GLM-4.6	80.6	54.5	82.7	91.0	43.2	69.0	69.8
DEEPSEEK-V4-PRO	85.9	51.6	64.5	80.1	55.3	73.1	68.2
GLM-5	87.5	53.6	73.8	91.5	43.1	56.7	67.7
GLM-4.7	77.0	55.0	74.6	89.9	41.6	63.0	66.7
DEEPSEEK-V4-FLASH	74.6	62.1	72.7	77.9	40.0	58.2	64.2
QWEN3-32B	78.0	77.5	83.3	66.0	27.5	50.3	63.4
MINIMAX-M2.7	91.7	54.8	67.9	74.4	50.3	41.3	62.9
MIMO-V2-FLASH	86.6	69.0	75.6	74.8	12.1	48.0	60.8
MINIMAX-M2.5	**92.6**	38.6	56.1	75.9	59.5	37.6	59.9
DS-R1-DISTILL-LLAMA-8B	33.8	53.7	75.5	56.6	40.6	36.0	49.2

annotators' decisions. Second, the system aggregates annotations and computes the agreement status at the sample level. Third, disagreement cases are flagged for manual review. Finally, after discussion or expert adjudication, the final label is confirmed and locked in the system. This process ensures that the final dataset is not solely determined by a single annotator's judgment, but is instead based on independent annotation, disagreement detection, and consensus-based verification.

In addition, the platform maintains an audit trail for each sample, including the original annotations, reviewer decisions, comments, and final labels. This design improves the transparency and reproducibility of the annotation process. By combining structured online annotation with consensus-based review, our system provides stronger quality assurance for constructing reliable evaluation data.

E Representative Examples for Each Risk Level

To further illustrate our annotation scheme, we provide representative examples for each risk level in the dataset. These examples are selected to demonstrate the distinctions among different categories, ranging from non-risk or contextually benign expressions to explicit self-harm ideation, method-related discussions, reported self-harm behavior, and harmful peer responses. Each table contains three representative chat segments for one risk level.

For each example, we present the original Chinese chat segment, its English translation, and a brief category summary. To improve readability, risk-related keywords and the primary risk-bearing speaker are highlighted when applicable. These examples are intended to clarify the practical annotation boundaries used in our dataset construction and to support a more transparent interpretation of the risk taxonomy.

Table 9 shows examples of Level 0 cases, where potentially sensitive words appear in non-risk or contextually benign settings. Table 10 presents cases involving psychological distress, loneliness, anxiety, or depression-related expressions without explicit suicidal intent. Table 11 includes examples with direct or impulsive expressions of suicidal ideation. Table 12 contains method-related or planning-oriented discussions. Table 13 reports cases involving prior or recent self-harm behavior. Finally, Table 14 presents harmful peer responses, including encouragement, abuse, or coercive language directed toward self-harm.

F Multi-Model Consensus Annotation

To scale risk annotation beyond what manual labelling could cover, we recruit a pool of $M = 17$ open-weight large language models (LLMs) as silver annotators on the raw dialogue corpus ($N = 11{,}189$ multi-turn topics). Each model is prompted with the same six-class risk schema used by human annotators (Appendix D; classes 0–5 as detailed in Appendix E) and emits a single label $v_m \in \{0, 1, 2, 3, 4, 5\} \cup \{-1\}$, where -1 denotes a parse failure or an explicit refusal. Let w_m be a per-model weight (we set w_m=1 in this work). The *weighted consensus ratio* for the winning class is

$$\rho = \max_{c \in \{0,\ldots,5\}} \frac{\sum_{m:\, v_m = c} w_m}{\sum_{m:\, v_m \neq -1} w_m}.$$

A topic is accepted as a silver-labelled sample iff $\rho \geq 2/3$ (≈ 11 of the 17 raters concur). Topics that fail this gate are stored as a *No-consensus* pool which is excluded from training and evaluation but retained for future hard-case analysis. Out of the 11,189 topics, 7,148 (63.9%) reached consensus and form the silver-labelled set used in our downstream experiments.

F.1 Voter Pool

Table 15 lists every voter model with its parameter scale, label-coverage (the fraction of items

Table 8: Detailed specifications of the LLMs included in the evaluation. 'Unknown' indicates undisclosed parameter sizes. For MoE models, both total and active parameters are reported when available.

Model	Size	Access	Provider	Model	Size	Access	Provider
GPT-5	Unknown	API	OpenAI	DeepSeek-V4-Flash	284B/13B act.	API	DeepSeek
GPT-4o mini	Unknown	API	OpenAI	DeepSeek-V3.2	Unknown	API	DeepSeek
Gemini-2.5-Pro	Unknown	API	Google	DeepSeek-R1	671B/37B act.	API	DeepSeek
Qwen3.5-Plus	Unknown	API	Alibaba	Doubao-Seed-2.0-Pro	Unknown	API	ByteDance
Qwen-Max	Unknown	API	Alibaba	Seed-OSS-36B-Instruct	36B	Open	ByteDance
Qwen3.5-397B-A17B	397B/17B act.	Open	Alibaba	Step-3.5-Flash	Unknown	API	ByteDance
Qwen3-235B-A22B	235B/22B act.	Open	Alibaba	MiniMax-M2.5	Unknown	API	MiniMax
QwenLong-L1-32B	32B	Open	Alibaba	MiniMax-M2.7	Unknown	API	MiniMax
Kimi-K2-Instruct	1T/32B act.	Open	Moonshot	Hunyuan-A13B-Instruct	80B/13B act.	Open	Tencent
Kimi-K2.5	Unknown	Open	Moonshot	MiMo-V2-Flash	309B/15B act.	Open	Xiaomi
Kimi-K2.6	Unknown	API	Moonshot	Ring-flash-2.0	100B/6.1B act.	Open	InclusionAI
GLM-5.1	Unknown	API	Zhipu	Llama-3.3-70B	70B	Open	Meta
GLM-5	Unknown	API	Zhipu	Llama-3-8B	8B	Open	Meta
GLM-4.7	Unknown	API	Zhipu	Qwen3-32B	32B	Open	Alibaba
GLM-4.6	Unknown	API	Zhipu	Qwen3-8B	8B	Open	Alibaba
GLM-4-32B-0414	32B	API	Zhipu	Qwen3-30B-A3B-Instruct-2507	30B/3B act.	Open	Alibaba
DeepSeek-V3	671B/37B act.	API	DeepSeek	Phi-4	14B	Open	Microsoft
DeepSeek-V4-Pro	1.6T/49B act.	API	DeepSeek	GLM-Z1-9B	9B	Open	Zhipu
DeepSeek-R1-Distill-Qwen3-8B	8B	Open	DeepSeek	Mistral-Small-3.2-24B-Instruct-2506	24B	Open	Mistral AI
DeepSeek-R1-Distill-Llama-8B	8B	Open	DeepSeek	Mistral-7B-Instruct	7B	Open	Mistral AI
DeepSeek-R1-Distill-Qwen-32B	32B	Open	DeepSeek	Gemma-3-27B-IT	27B	Open	Google
Gemma-2-9B-IT	9B	Open	Google	Gemma-2-9B	9B	Open	Google
Ling-flash-2.0	100B/6.1B act.	Open	InclusionAI				

for which it returned a parseable label), abstention rate, and agreement with the final silver label. The pool deliberately spans (i) reasoning-tuned models in the QWENLONG/DEEPSEEK-R1 family, (ii) instruction-tuned dense models from QWEN, GEMMA, GLM-4 and MISTRAL, and (iii) Mixture-of-Experts (MoE) systems such as DEEPSEEK-V3, ERNIE-4.5, MINIMAX-M1 and QWEN3-235B-A22B. The whole-pool Fleiss' $\kappa = 0.422$ (computed by treating -1 as missing data) places inter-model agreement in the *moderate* range, broadly comparable to crowd-sourced human annotation on similar risk-labelling tasks (Landis and Koch, 1977). We observe a clear trade-off between *coverage* and *agreement*: the strongest reasoning models (DEEPSEEK-R1, QWQ-32B) abstain on $\sim 65\%$ of items but agree with the consensus $\geq 84\%$ of the time when they do answer, whereas smaller dense models (GEMMA-2-9B, GEMMA-3-12B) answer every item but agree with the consensus on only $\sim$57–61% of them. This complementarity motivates the use of a heterogeneous pool: pooling high-coverage and high-precision voters together raises the effective *joint* coverage of consensus-grade labels to 63.9%, which would not be reachable with any single annotator.

F.2 Distribution of Silver Labels

Figure 11 summarises the consensus outcome over the full candidate pool. Three trends are visible. First, the dataset is dominated by the *Irrelevant* class (23.8%): keyword search recovers many topics whose surface form mentions risk-related vocabulary but whose discursive context is benign (idioms, song lyrics, gaming slang). Second, ideation-related risk levels are well represented (*Suicidal ideation* 8.4%, *Suicidal method* 15.9%, *Inducement / attack* 9.3%), while the highest-severity *Suicide attempt* class is rare (3.7%; only 410 items), confirming the long-tailed prevalence reported for psychiatric-risk corpora. Third, 36.1% of candidates land in the *No-consensus* pool—these constitute the most ambiguous boundary cases (e.g. borderline ideation vs. method, or sarcastic ver-

类别总结	**Chat Segment (Source Text, ZH)**	**Chat Segment (Translation, EN)**
法律/新闻语境中的敏感词	user_24: 尊嘟假嘟这是能出的吗 **user_379**: 假的不然算贩毒罪处理 **user_632**: 看新闻吗有个人家里人过世出生前用的抗癌药算贩毒判个月 user_143: 其实我可以买下来然后当着你的面扔掉	User 24: For real, can this even be released? **User 379:** No way—otherwise it would count as **drug** trafficking. **User 632:** Did you see the news? Someone's family member passed away; the anticancer drugs used before birth were ruled as **drug** trafficking—a few months' sentence. User 143: Actually I could buy it and throw it away right in front of you.
日常字面义	**user_433**: 最后两块钱 **user_433**: 买了杯奶茶 **user_433**: 好喝 user_533: 哈哈哈 user_533: 到了	**User 433:** My **last** two yuan. **User 433:** Bought a milk tea. **User 433:** Tastes good. User 533: Hahaha. User 533: It arrived.
器物/技术问题	**user_38**: 建议以后买手机时选择手机我这被迫购买了荣耀的手机几个月后手机后背无缘无故发出一阵阵沙铃声 **user_38**: 也就是无缘无故的发响 user_73: 噗 **user_38**: 是的才使用了不到半年就这样了 **user_38**: 并且有了很多次了	**User 38:** Suggest choosing carefully when buying a phone—I was forced to buy a Honor phone; months later the back started making bursts of **sand**-like ringing for no reason. **User 38:** Just ringing out of nowhere. User 73: Pfft. **User 38:** Yeah, less than half a year of use and it's already like this. **User 38:** And it has happened many times.

Table 9: Representative examples for Level 0 suicide risk annotation.

bal attacks vs. genuine inducement) and motivate the human-in-the-loop adjudication described in Appendix D.

F.3 Per-Label Consensus Quality

Table 16 drills down into the strength of agreement *within* each accepted label. For all six risk classes the mean consensus ratio $\bar{\rho}$ exceeds 0.76, and at least 62% of items per class are labelled with $\rho \geq 0.7$, i.e. supported by ≥ 12 of the 17 voters. The *Irrelevant* class is the most strongly endorsed ($\bar{\rho} = 0.830$; 35.4% of items reach $\rho \geq 0.9$), reflecting the fact that most off-topic keyword matches are easy to filter. By contrast, the safety-critical *Suicide attempt* and *Suicidal method* classes show somewhat lower headroom ($\bar{\rho} \approx 0.77$–0.79; only 15.9–19.3% reach $\rho \geq 0.9$), consistent with the cognitive difficulty of distinguishing concrete attempts from descriptions of methods. The *No-consensus* pool sits clearly below the acceptance threshold ($\bar{\rho} = 0.479$), validating the $2/3$ cut-off as a conservative quality gate.

F.4 Per-Voter Behaviour

Figure 12 decomposes every voter into three behavioural buckets relative to the silver label: *agrees* with the consensus (green), *disagrees* (red), or *abstains* (grey). Two qualitative regimes emerge. **High-coverage answerers**—QWENLONG-L1-32B, DS-R1-DISTILL-QWEN-32B, QWEN3-32B, GLM-4-32B and ERNIE-4.5—answer essentially every item with 84–90% agreement, behaving as reliable workhorses of the pool. **Selective reasoners**—DEEPSEEK-R1 and QWQ-32B—abstain on roughly two thirds of the items but exhibit very high precision (84.7% and 87.1% respectively) whenever they do commit, suggesting that they internally apply a confidence-based refusal policy. A third group of weaker dense models (GEMMA-2-9B, GEMMA-3-12B, MISTRAL-SMALL-3.2) achieves universal coverage but considerably lower agreement (57–64%), and would not be safe to use as the sole annotator. Crucially, removing any single voter from the pool changes the final silver label on at most 0.6% of items, confirming the *redundancy* that motivates the multi-model design.

F.5 Inter-Model Agreement Structure

Figure 5 reports pairwise Cohen's κ for every pair of voter models, excluding items on which either model abstained. The matrix exposes three coherent clusters. (i) A tight

类别总结	Chat Segment (Source Text, ZH)	Chat Segment (Translation, EN)
明确抑郁/焦虑表述	**user_1104**: 焦虑症 **user_1182**: 我很痛苦 **user_1320**: 抑郁吗 **user_1320**: 我都不配因为钱的事抑郁烂透的命运	**User 1104: Anxiety disorder**. **User 1182:** I feel so **painful**. **User 1320: Depressed**? **User 1320:** I don't even deserve to be **depressed** over money—my fate is rotten.
生活迷茫与无助	user_106: 既有钱赚要是不顺心还能偷顾客外卖吃不是 **user_659**: 可是一个人生活感觉很心慌 user_529: 如果好不容易吃顿好的被偷了也太崩溃了 **user_659**: 我感觉很痛苦 **user_659**: 我现在觉得我好像一点方向也没有我好无助	User 106: You earn money, and if things don't go **your way** you can even steal customers' food deliveries, right? **User 659:** But living alone makes me feel really **anxious**. User 529: If a hard-won good meal gets stolen, that's too **crushing**. **User 659:** I feel very **painful**. **User 659:** Now I feel I have no direction at all—I'm so **helpless**.
孤独与共情式痛苦	**user_632**: 不都是这样吗 user_488: 也许 **user_632**: 能来到这里或多或少都有痛苦 **user_632**: 我已经孤独好久了 user_488: 我也是	**User 632:** Isn't everyone like this? User 488: Maybe. **User 632:** Anyone who can end up here has some **pain** more or less. **User 632:** I've been **lonely** for a long time. User 488: Me too.

Table 10: Representative examples for Level 1 suicide risk annotation.

Qwen-family cluster at the top (QWENLONG-L1-32B, DS-R1-DISTILL-QWEN-32B, QWEN3-235B-A22B, QWEN3-32B, QWEN2.5-72B) with $\kappa \in [0.52, 0.61]$, indicating substantial agreement consistent with shared pre-training data and instruction tuning. (ii) A mid-tier of large-scale MoE and Chinese-language-tuned models (ERNIE-4.5-300B, GLM-4-32B, MINIMAX-M1, HUNYUAN-A13B, DEEPSEEK-R1) hovering around $\kappa \approx 0.42$–0.55. (iii) A weakly-agreeing *Western-pretrained* cluster at the bottom (GEMMA-2-9B, GEMMA-3-12B, MISTRAL-SMALL-3.2) with $\kappa \approx 0.25$–0.40, which is also where most cross-cluster pairs land. The block-diagonal structure shows that our consensus is not dominated by any single model family: every cluster contributes votes that survive the $2/3$ threshold, and the cross-cluster pairs sit above pure chance (all $\kappa > 0.2$). This makes the multi-voter scheme robust to systematic biases that might be shared within a single LLM lineage.

F.6 Summary

The 17-voter consensus pipeline yields 7,148 silver-labelled dialogues at moderate inter-rater agreement (Fleiss' $\kappa = 0.422$), with most accepted labels supported by ≥ 12 of the 17 models. The complementarity between high-coverage workhorses and high-precision selective reasoners drives the effective 63.9% acceptance rate, while the family-level clustering in Figure 5 shows that the consensus is distributed across model lineages and is therefore unlikely to inherit any single family's blind spots. We use this silver-labelled corpus as the training source for the experiments reported in the main paper, and reserve the $4{,}041$ no-consensus items for human re-adjudication via the platform described in Appendix D.

G Streaming Risk Classification under Partial Disclosure

G.1 Motivation

In a real deployment, a suicide-risk classifier never sees a complete topic block at once: messages arrive in real time, and any helpful intervention must be triggered *before* the conversation has finished. The benchmark in the main paper gives every model the full topic context, which conflates two abilities that are conceptually distinct: (i) *interpreting* a fully revealed dialogue and (ii) *committing* to a risk label after seeing only a partial prefix. The latter ability is operationally far more important, because it determines when a downstream system can escalate to a human. We therefore re-evaluate eight representative open-access and commercial API models in a controlled streaming setting that varies only the prefix length of the input, while keeping everything else identical to the *Few-Shot* configuration of the main benchmark.

类别总结	**Chat Segment (Source Text, ZH)**	**Chat Segment (Translation, EN)**
直接求死意愿	user_489: 老鼠为什么要把死老鼠拖走 **user_470**: 不想活了 **user_470**: 我不活了 user_125: 别死 **user_470**: 我有点死了	User 489: Why would a **rat** drag away a **dead rat**? **User 470:** I **don't want to live anymore**. **User 470:** I'm not **living** anymore. User 125: Don't **die**. **User 470:** I'm kind of **dead** already.
冲突情境冲动表态	user_65: 吵架的可多了 user_65: 干嘛只我 **user_654**: 好吧 **user_654**: 只有我要死 user_65: 别死	User 65: So many people fight. User 65: Why only me? **User 654:** Fine. **User 654:** Only I want to **die**. User 65: Don't **die**.
愤怒宣泄式意愿	**user_1810**: 死爹死妈的狗东西我不想活了 **user_1810**: 死爹死妈的狗东西我不想活了 **user_1810**: 死爹死妈的狗东西我不想活了 user_433: 最后一句 user_507: 不允许你侮辱我你这个精神病	**User 1810:** You **dead**-dad **dead**-mom piece of trash—I **don't want to live anymore**. **User 1810:** [repeats the same line] **User 1810:** [repeats the same line] User 433: **Last** sentence. User 507: You are not allowed to insult me, you mentally **ill** person.

Table 11: Representative examples for Level 2 suicide risk annotation.

G.2 Subset and Protocol

We construct a streaming validation set by re-using the 1,200 *Few-Shot* items already evaluated by 24 baseline models, then ranking each item by the cross-model *correct ratio* (fraction of the 24 baselines that produced the gold label). The top-100 items per class with at least five utterances form a N=600, class-balanced *quality pool*. Average correct ratio on this pool is 86%, so any drop we observe under partial disclosure is attributable to information loss rather than item difficulty.

For every topic block of $|M|$ utterances and every disclosure ratio $\tau \in \{30, 50, 70, 90, 100\}\%$ we reveal the first $\lceil \tau\,|M| \rceil$ utterances (at least one) and ask the model to commit to a label in $\{0, \ldots, 5\}$. The five ratios bracket the regimes that matter operationally: τ=30% is the *early-warning regime* where a deployed system must decide with about a third of the conversation in hand; $\tau \in \{50, 70\}\%$ probe the intermediate trajectory; τ=90% measures whether the final 10% of the conversation carries decisive evidence; and τ=100% recovers the non-streaming upper bound used in the main benchmark. The prompt, demonstrations and decoding hyper-parameters are identical to the *Few-Shot* setting (Appendix C).

We evaluate eight models spanning four provider families and two capacity tiers: flagship API models from Zhipu (GLM-5), Alibaba (QWEN3.5-397B-A17B), Google (GEMINI-2.5-PRO), OpenAI (GPT-4O-MINI), StepFun (STEP-3.5-FLASH), and Moonshot (KIMI-K2-INSTRUCT); the open-weight reasoner DEEPSEEK-R1; and its 32-B dense distillate DS-R1-DISTILL-QWEN-32B. The parent–distill pair lets us quantify the cost of distillation under partial-context decoding, which the literature has not yet examined; GPT-4O-MINI provides a compact commercial baseline against which to compare the larger open-weight and frontier API systems.

G.3 Results

All eight models improve monotonically with τ, but the slope is highly model-dependent. Table 17 and Figure 13 report the headline numbers. The absolute Accuracy gain from τ=30% to τ=100% is +31.3pp on average. The largest jump is QWEN3.5-397B-A17B (+33.9pp, 58.5 → 92.4%), which also attains the highest full-context Accuracy in this panel; STEP-3.5-FLASH is a close second on slope (+33.6pp). The smallest gain is GLM-5 (+27.9pp, 55.7 → 83.6%). At τ=30%, QWEN3.5-397B-A17B and GEMINI-2.5-PRO jointly lead (58.5% each), ahead of DEEPSEEK-R1 (58.0%) and the distilled 32-B model (52.8%). GPT-4O-MINI trails the panel at 46.8% in this early-warning regime yet still gains +31.3pp by full context (78.2% at τ=100%), indicating that its errors are driven primarily by missing late-arriving evidence rather than by a flat inability to use partial prefixes. GEMINI-2.5-PRO dominates the mid-trajectory regime (70.0% at τ=50%, 79.8% at τ=70%), while QWEN3.5-397B-A17B pulls

类别总结	Chat Segment (Source Text, ZH)	Chat Segment (Translation, EN)
药物/渠道信息	**user_15**: 以前泰国兽医商店可以买到戊巴比妥钠现在不知道可不可以了 **user_15**: 和平药丸手册这本书里面提到过泰国可以购买戊巴比妥钠 user_632: 啥枪对头打都够了啊 user_39: 这是最快最不痛苦的啦	**User 15:** Veterinary shops in Thailand used to sell **pentobarbital sodium**; not sure if they still do. **User 15:** The Peaceful **Pills Handbook** mentions buying **pentobarbital sodium** in Thailand. User 632: A **gun** to the head is enough, huh? User 39: This is the fastest and least **painful** way.
化学物自制	user_1455: 有人约的吗 user_1460: 哪里买的 **user_447**: 纯度不高的氰化物可以自制 user_1462: 我也想 user_1460: 不会自制的	User 1455: Anyone want to meet up? User 1460: Where do you buy it? **User 447:** Low-purity **cyanide** can be made at home. User 1462: I want to too. User 1460: I don't know how to make it myself.
烧炭方式细节	**user_425**: 烧炭外面的人察觉的到吗 **user_425**: 木炭烟好像少 **user_425**: 我怕看见烟报警了 user_1419: 窗户窗帘都关着怎么看见	**User 425:** If I burn **charcoal**, can people outside notice? **User 425: Charcoal** smoke seems less. **User 425:** I'm afraid someone sees the smoke and calls the police. User 1419: With windows and curtains closed, how would they see?

Table 12: Representative examples for Level 3 suicide risk annotation.

ahead from τ=90% onward (88.1% $\rightarrow$ 92.4%). The six stronger non-distilled systems cluster at $83.6 - 92.4\%$ at τ=100%, with GPT-4O-MINI (78.2%) and the distillate (81.1%) forming a lower tier. The distillation gap versus DEEPSEEK-R1 is -5.2pp at τ=30% and widens to -8.8pp at τ=100%. The Macro F1 curves in Figure 13 (right) and the heatmap in Figure 14 show the same picture.

Per-class behaviour: easy classes saturate immediately, but the highest-risk class (L4) does *not*. Table 18 and Figure 6 drill into the per-class trajectory. The pattern that emerges has a clear practical implication. Label L0 (*Unrelated*) is identified with $\geq$ 87% accuracy at *every* disclosure ratio for every model: a third of the conversation is already enough to rule out a risk situation. By contrast, the two semantically subtle classes L1 (*Depressive without ideation*) and L2 (*Suicidal ideation without method*) show the steepest slopes, gaining +38 to +59pp from τ=30% to τ=100%. These two classes require cumulative evidence of intent that almost never surfaces in the first three utterances.

Crucially, the highest-risk label L4 (*active suicidal attempt*) does *not* saturate early in our corpus. At τ=30%, L4 Accuracy ranges from 26.8% (KIMI-K2-INSTRUCT) to 52.0% (GEMINI-2.5-PRO); none of the eight models recovers L4 cases reliably until τ=90% ($48 - 92\%$), and DS-R1-DISTILL-QWEN-32B never exceeds 61% even at full context. This contradicts the common intuition that crisis disclosures are front-loaded: in our corpus, explicit method execution and emergency-room cues tend to appear *after* an extended period of depressive build-up, so a streaming detector with a fixed early trigger will systematically miss the most urgent cases. Counting raw errors, the two classes L1 and L5 (verbal abuse / inducement of others) jointly account for 40.0% of all misclassifications at τ=30% across the eight models, while L0 contributes only 1.5%.

G.4 Practical Implications

Three concrete take-aways follow from these numbers. First, *a single disclosure-ratio sweep is a cheap, deployment-relevant robustness check.* We recommend that future evaluations of risk-classification LLMs report at minimum the $\tau \in \{30, 70, 100\}$% triple alongside full-context Macro F1; the gap between τ=30% and τ=100% is a more honest upper bound on real-time deployment risk than the headline full-context number alone. Second, *distillation costs more in long-context regimes than in short-prefix regimes:* the gap between DEEPSEEK-R1 and its 32-B distillate

类别总结	Chat Segment (Source Text, ZH)	Chat Segment (Translation, EN)
既往尝试自述	**user_1472**: 我是因为先天早产双眼失明一点都看不见那种然后还有点轻微的脑瘫四肢协调性不好 user_1472: 这么小的孩子不要离开世界呀我觉得你现实里一定很可爱的如果世界少了你也会觉得很可惜的 **user_24**: 昨天凌晨我才发现我一个人喝盐也有点怕怕的 **user_1472**: 我已经一个半月没来这个软件啦尝试两次自杀之后可能我自己想死但是我希望每个人都活着看到你才岁不知道为什么很难过	**User 1472:** I was born prematurely and am completely blind, with mild **cerebral palsy** and poor coordination. User 1472: Such a young child—don't **leave** this world; you must be lovely in real life; the world would miss you. **User 24:** Only at dawn yesterday did I realize I drank **salt** alone and felt a bit scared. **User 1472:** I haven't been on this app for a month and a half; after two **suicide** attempts I may still want to **die**, but I hope everyone stays alive—you're only [age] years old; I don't know why I feel so **sad**.
烧炭等行为已实施	**user_833**: 越聊越想死了 user_834: 那不聊了这样就不想死了 **user_833**: 那就是想死 **user_833**: 而不是越想死 **user_833**: 昨天去烧炭了	**User 833:** The more we chat, the more I want to **die**. User 834: Then let's stop chatting—that way you won't want to **die**. **User 833:** So it's wanting to **die**. **User 833:** Not "the more I chat the more I want to **die**." **User 833:** I went to burn **charcoal** yesterday.
自伤刚完成	**user_547**: 肚子疼 **user_547**: 还是疼 **user_516**: 恶心死了 **user_516**: 死傻逼 **user_516**: 刚割完手打开一看他妈的给我发信息	**User 547:** My stomach **hurts**. **User 547:** Still **hurts**. **User 516:** So gross I could **die**. **User 516: Die**, idiot. **User 516:** I just finished **cutting** my wrist—opened the app and they messaged me.

Table 13: Representative examples for Level 4 suicide risk annotation.

widens from 5.2pp at τ=30% to 8.8pp at τ=100%, suggesting that whatever the parent model learns in pre-training, it is the ability to *integrate* late-arriving context (not the ability to commit early) that distillation loses first. Third, *class-conditional curves are far from parallel*: any deployment that thresholds on a single global confidence score will systematically under-trigger on L1/L2/L4 in early prefixes while already being confident on L0. A deployment-friendly objective would equalise calibration across τ *and* across classes jointly; we leave this for future work.

G.5 Limitations

The streaming validation set of N=600 items, while balanced across the six classes and statistically well-powered for between-model contrasts (±2.0pp Accuracy SE at N=100 per class), is deliberately curated to be *learnable* under full context: items that even the 24 main-paper baselines fail on are excluded so that any drop at short prefixes can be attributed to information loss rather than intrinsic difficulty. Conclusions about absolute deployment accuracy on raw conversation streams should therefore be treated as an upper bound. A second limitation is that we use prefix truncation as a proxy for streaming, which assumes the speaker order in the corpus reflects real-time arrival. The corpus we use preserves arrival order at message-level granularity but does not record inter-message latency; modelling time-aware arrival is left to future work. A third limitation is that we report the eight models with full N=600 coverage at the time of writing; we release the streaming dataset and the evaluation harness so that follow-on work can extend the panel to additional closed-source flagships (e.g. GPT-5) under their respective API rate limits.

H Qualitative Error Examination

Based on the experimental results, we select the Gemma model as our baseline for a detailed error analysis. We evaluate all 469 incorrect predictions from the test set to categorize error types. Figure 15 shows the confusion matrix heatmap, with true categories on the x-axis and predicted categories on the y-axis. It is evident that the Indicator category is most frequently misclassified into other

类别总结	Chat Segment (Source Text, ZH)	Chat Segment (Translation, EN)
教唆自杀	**user_1017**: 臭傻逼，看不懂头像，觉得你不是个东西，过街老鼠 **user_89**: 你踏马快死了在叫 **user_89**: 赶紧笑死吧 **user_89**: 该死 **user_1017**: 操你妈教唆别人用这个自杀，你怎么不去死呀	**User 1017:** Stupid idiot, can't even read avatars—you're nothing, a street **rat**. **User 89:** You're almost **dead** and still yapping. **User 89:** Hurry up and laugh yourself to **death**. **User 89:** You deserve to **die**. **User 1017:** F— you for coaching others to use this method to **kill themselves**—why don't you go **die**?
诅咒与贬损劝死	**user_318**: 原来是死穷逼一个 **user_318**: 笑死 **user_1201**: 什么时候让我去光顾一下 **user_1201**: 哎呦喂都不让我光顾吗 **user_1201**: 你这个乖儿子要是不敢来的话你和你妈就是卖淫到死	**User 318:** Turns out you're just a broke **dead** loser. **User 318:** LOL **dead**. **User 1201:** When can I come **visit**? **User 1201:** Wow, you won't even let me **visit**? **User 1201:** If you don't dare show up, you and your mom can sell yourselves until you **die**.
对骂式人身攻击	**user_318**: 死妈废物 **user_318**: 死妈废物还不发你黑鸡巴出来 **user_1201**: 像你这种废物就喜欢大黑鸡巴塞到你嘴里面 **user_318**: 死妈废物还不叫	**User 318: Dead**-mom **waste**. **User 318: Dead**-mom **waste**, still won't send your dick pic. **User 1201:** Trash like you loves having a big black dick shoved in your mouth. **User 318: Dead**-mom **waste**, still won't bark.

Table 14: Representative examples for Level 5 suicide risk annotation.

categories, with 205 (43.7%) samples incorrectly predicted. The remaining five categories each account for approximately 10% of the error samples.

Further analysis of the 205 Indicator case errors reveals that the majority are misclassified as Depression and Ideation, comprising 58 (28.3%) and 76 (37.1%) samples, respectively. Through manual examination of these erroneous samples, we find that users often include suiciderelated signal words when expressing frustration or making jokes, without exhibiting genuine suicidal intent. These subtle contextual differences lead to the model's misclassification.

Additionally, we observe significant confusion between the Behavior and Attempt categories. This misclassification may occur because users discussing plans to commit suicide often recount others' past suicide attempts, leading the model to erroneously infer that the user themselves have attempted suicide. These findings indicate that the model struggles to distinguish between classes with overlapping semantics or similar contextual cues, particularly when the distinctions are nuanced.

I Ethical Considerations

We acknowledge that research centered on suicide and severe depressive tendencies is highly sensitive. Therefore, throughout the construction and dissemination of our dataset, we adhere to a series of ethical standards, including data privacy protection, annotator mental health safeguards, and stringent data sharing protocols.

Data Source and Compliance. All Telegram groups involved in this study are public, allowing anyone to join and participate without identity verification or specific approval. Consequently, the content is publicly accessible in accordance with Telegram's platform policies. To protect privacy, all messages are anonymized before analysis by removing group chat IDs, user IDs, and personal information, replacing them with irrelevant sequential identifiers.

Mental Health of Annotators. Prior to annotation, all annotators undergo an initial consultation to assess their psychological suitability for the task. Regular conversations are conducted throughout the annotation process, and a final consultation is arranged to ensure that the work has not adversely affected their mental health.

Data Sharing. Given the dataset's sensitive na-

Table 15: Voter pool used for consensus annotation. *Coverage* is the fraction of the $N = 11{,}189$ items for which the model returned a valid label (i.e. did not abstain or fail to parse). *Agree* is the model's agreement rate with the final silver label, computed on the $N_c = 7{,}148$ items that reached consensus and with abstentions excluded from the denominator.

Model	Family	Scale	Coverage	Abstain	Agree
DeepSeek-R1	DeepSeek	671B MoE (A37B)	35.0%	65.0%	84.7%
DeepSeek-V3	DeepSeek	671B MoE (A37B)	53.8%	46.2%	64.3%
DS-R1-Distill-Qwen-32B	DeepSeek-R1-Distill	32B dense	100.0%	0.0%	88.9%
ERNIE-4.5-300B	ERNIE	300B MoE (A47B)	98.0%	2.0%	85.0%
GLM-4-32B	GLM-4	32B dense	100.0%	0.0%	84.3%
Gemma-2-9B	Gemma	9B dense	100.0%	0.0%	56.9%
Gemma-3-12B	Gemma	12B dense	100.0%	0.0%	61.3%
Gemma-3-27B	Gemma	27B dense	100.0%	0.0%	71.2%
Hunyuan-A13B	Hunyuan	80B MoE (A13B)	100.0%	0.0%	81.8%
MiniMax-M1	MiniMax-M1	456B MoE (A46B)	67.1%	32.9%	82.2%
Mistral-Small-3.2	Mistral	∼24B dense	99.8%	0.2%	64.0%
QwQ-32B	QwQ	32B dense	35.2%	64.8%	87.1%
Qwen2.5-72B-Instruct	Qwen2.5	72B dense	91.8%	8.2%	87.2%
Qwen3-235B-A22B	Qwen3	235B MoE (A22B)	89.4%	10.6%	88.4%
Qwen3-32B	Qwen3	32B dense	100.0%	0.0%	87.9%
QwenLong-L1-32B	QwenLong	32B	100.0%	0.0%	89.7%
QwenLong-CoT	QwenLong-CoT	32B (CoT)	100.0%	0.0%	79.2%
Pool average	—	—	86.5%	13.5%	79.1%
Fleiss' κ (all 17 raters)	$\kappa_{\text{Fleiss}} = 0.422$ (moderate agreement)				

Table 16: Per-label consensus statistics on the $N = 11{,}189$ silver-annotated topics. ρ is the weighted consensus ratio; $\rho \geq 0.7/0.9$ is the share of items with ρ above the corresponding threshold; *Unanimous* counts items on which every non-abstaining model voted for the final label.

ID	Label	# items	Share	Mean ρ	Median ρ	$\rho \geq 0.7$	$\rho \geq 0.9$	Unanimous
0	Irrelevant	2,667	23.8%	0.830	0.840	79.0%	35.4%	18.2%
1	Depression cue	307	2.7%	0.762	0.746	62.2%	17.3%	6.2%
2	Suicidal ideation	937	8.4%	0.802	0.798	73.3%	25.3%	11.1%
3	Suicidal method	1,784	15.9%	0.789	0.788	72.0%	19.3%	5.1%
4	Suicide attempt	410	3.7%	0.768	0.749	65.9%	15.9%	3.7%
5	Inducement / attack	1,043	9.3%	0.785	0.780	68.6%	21.4%	7.7%
–	No-consensus (excluded)	4,041	36.1%	0.479	0.492	—	—	—

ture, access is restricted to accredited research institutions. Applicants must provide valid identification, proof of employment, and research objectives, and sign a data use agreement for non-disclosure. Researchers are expected to use the data responsibly and for socially beneficial purposes.

J Effect of Jargon vs. Non-Jargon Seed Keywords on Classification

J.1 Setup

Our seed lexicon comprises two risk-related subsets used to harvest topic blocks from live chat: **jargon** (indirect or euphemistic cues) and **non-jargon** (literally recognisable cues). For each item in the balanced validation set (4,619 topic blocks), we label the block as **jargon-dominant** or **non-jargon-dominant** according to which subset accounts for more `<keyword>` tags in the dialogue (ties and blocks without either tag type are excluded), yielding 1,576 jargon-dominant and 1,347 non-jargon-dominant topics.

We then measure how well the same $M{=}22$ few-shot models from the main benchmark classify these blocks, using the stored predictions in `evaluation_results_balanced_fewshot_*.json`. Each model$\times$topic pair contributes one accuracy score; we report micro-pooled accuracies over $N_{\text{jargon}}{=}6{,}235$ and $N_{\text{non-jargon}}{=}5{,}762$ predictions. Per-class comparisons use two-proportion z-tests with Bonferroni correction ($\alpha{=}0.05/6$).

Table 17: Streaming risk classification under five disclosure ratios $\tau \in \{30, 50, 70, 90, 100\}\%$ for the four open-access reasoning models, evaluated on the full N=600 streaming validation set. Each cell reports Accuracy / Macro-F1 (%). Δ_{Acc} is the absolute Accuracy gain from τ=30% to τ=100%. Best per column in bold.

Model	τ=30%	τ=50%	τ=70%	τ=90%	τ=100%	Δ_{Acc}
STEP-3.5-FLASH	54.6 / 54.9	65.5 / 65.6	76.6 / 76.8	**85.5 / 85.6**	88.2 / 88.2	+33.6
KIMI-K2-INSTRUCT	55.8 / 55.1	64.2 / 64.3	73.6 / 73.4	85.0 / 85.0	89.1 / 89.1	+33.2
DEEPSEEK-R1	**58.0 / 57.4**	**68.8 / 68.4**	**77.4 / 77.3**	84.5 / 84.4	**89.9 / 89.9**	+31.9
DS-R1-DISTILL-QWEN-32B	52.8 / 52.4	60.8 / 60.4	69.5 / 69.1	76.4 / 76.0	81.1 / 81.0	+28.2

Table 18: Per-class Accuracy (%) at the extreme disclosure ratios τ=30% and τ=100%, and the per-class gain Δ_{cls} (pp). L0–L5 correspond to the six risk levels defined in Appendix G.

Model	τ=30%						τ=100%						Δ_{cls} (pp)					
	L0	L1	L2	L3	L4	L5	L0	L1	L2	L3	L4	L5	L0	L1	L2	L3	L4	L5
STEP-3.5-FLASH	97	34	32	66	43	56	93	77	91	91	86	92	-3.9	+43.3	+58.8	+24.7	+43.2	+36.0
KIMI-K2-INSTRUCT	96	44	42	77	27	52	94	83	96	99	80	82	-1.6	+38.4	+53.8	+21.8	+53.4	+30.2
DEEPSEEK-R1	96	34	31	72	46	67	91	87	82	98	93	89	-4.9	+52.6	+50.7	+26.1	+46.7	+22.2
DS-R1-DISTILL-QWEN-32B	91	32	36	76	32	54	88	82	84	90	61	82	-2.2	+49.7	+47.7	+14.5	+28.9	+28.0

J.2 Results

Overall accuracy. Non-jargon-dominant blocks reach 68.3% pooled accuracy versus 66.9% on jargon-dominant blocks (Δ_{Acc}=+1.35 pp, z=1.58, p=0.12). The corpus-level gap is small and not significant; effects are primarily visible after stratifying by gold risk class.

Class-stratified effects. Figure 16 and Table 19 report the per-class breakdown:

- **L1 (Depressive)** and **L4 (Attempt):** non-jargon-dominant blocks are more accurate by 15.9 pp and 13.1 pp (Bonferroni significant).
- **L0 (Irrelevant)** and **L5 (Inducement):** jargon-dominant blocks are more accurate by 12.7 pp and 14.2 pp (significant).
- **L2 (Ideation)** and **L3 (Method):** no significant split (gaps of 3.5 pp and 0.6 pp); L3 accuracy is $\geq 84\%$ in both groups.

Figure 19 summarises the pooled headline, the cross-model distribution of Δ_{Acc} (non-jargon minus jargon), per-class gaps, and mean absolute label error. Figure 18 gives a compact heatmap view.

Takeaway. Jargon-dominant and non-jargon-dominant blocks are *not* uniformly easier or harder: the direction of the gap reverses across classes. Reporting a single pooled accuracy therefore understates the interaction between seed-word type and risk level; class-conditional analysis is necessary.

Table 19: Per-class pooled few-shot accuracy (%) on jargon-dominant vs. non-jargon-dominant topic blocks. Δ_{Acc} = non-jargon − jargon (pp). * Bonferroni $p < 0.05/6$.

Class	**Non-jargon**	**Jargon**	**Δ**	
L0 Unrelated	68.1	80.9	-12.7^*	
L1 Depressive	70.1	54.2	$+15.9^*$	
L2 Ideation	73.9	77.4	-3.5	
L3 Method	84.4	85.0	-0.6	
L4 Attempt	57.6	44.5	$+13.1^*$	
L5 Inducement	54.2	68.4	-14.2^*	
Pooled	68.3	66.9	$+1.4$	p=0.12

Chain-of-Thought Prompt for LLM Suicide Risk Evaluation

你是一名专业的心理健康风险分类器。你的任务是根据以下一段聊天上下文、用户输入 `<Signal Word>` 以及分析结果内容，对用户关联的自杀风险等级进行判断。请严格遵循分类标准，谨慎评估用户是否存在现实风险，并最终输出以下步骤和最终分类。

(You are a Professional Mental Health Risk Analyst. Your core task is to analyze and assess the suicide risk level of the user based on a specific message containing the `<Signal Word>` tag within a chat context. This work is crucial, and you must perform it with a high degree of diligence and professionalism. Please strictly follow the steps and criteria below for text analysis and risk classification.)

Step 1: Analyze Chat Context Thoroughly and Target Classification

1. **Read the Chat Context:** Carefully and thoroughly read all messages in the chat context provided below to understand the background, emotional changes, and topic development.
2. **Locate the Target User:** In the conversation, find the target message that contains the `<Signal Word>` tag and identify the user who posted this message as the target user.

步骤 1：上下文分析与目标定位

1. **全面阅读上下文：**认真阅读给定聊天中的所有消息，理解对话背景、情绪变化和主题发展。
2. **定位目标用户：**找出含有 `<Signal Word>` 的消息，并将该消息发布者视为目标用户。

Step 2: Risk Category Assessment

Compare the target user's statements with the following risk categories and select the most appropriate risk category based on the defined classification criteria.

- **Category 0: No Evidence:** The target user's statements are completely unrelated to depressive emotions, suicidal thoughts, suicide methods, attempts, inducement, or verbal aggression.
- **Category 1: Depression:** The target user's statements express obvious depressive emotions, persistent low mood, hopelessness, self-negation, or emotional distress, but do not mention suicidal thoughts or behaviors.
- **Category 2: Suicide Ideation:** The target user's statements express thoughts or desires of suicide but do not mention a specific method, plan, or attempt behavior.
- **Category 3: Suicide Behavior:** The target user's statements clearly discuss or mention specific suicide plans but do not mention or imply specific attempts or implementation actions that have already begun.
- **Category 4: Suicide Attempt:** The target user's statements describe specific, ongoing, or already occurred suicide attempt behaviors, or concrete preparatory steps for implementation.
- **Category 5: Suicide Inducement:** The target user's statements do not express their own risk status but include advice, encouragement, or persuasion related to suicide.

步骤 2：风险类别判断

请根据目标用户的表达，在以下风险类别中进行判断：

- **类别 0：无证据：**与抑郁、自杀想法、自杀方法、自杀尝试、自杀诱导或攻击性表达均无关。
- **类别 1：抑郁情绪：**表达明显消极、低落、绝望、自我否定或痛苦情绪，但未出现自杀意图或行为。
- **类别 2：自杀意念：**表达想死、自杀愿望或相关念头，但未出现具体方法、计划或行为。
- **类别 3：自杀计划：**明确提到具体自杀方式或计划，但未体现正在实施或已经实施。
- **类别 4：自杀尝试：**描述正在进行、已经发生或具有明确准备行为的自杀尝试。
- **类别 5：自杀诱导：**并非表达自身风险，而是劝说、诱导、鼓励或美化他人自杀。

Step 3: Apply Priority Rules

If the target user's statements fit multiple categories, classify them according to the following priority order and output only the highest-priority category number:

$$\text{Category 4} > \text{Category 3} > \text{Category 2} > \text{Category 1} > \text{Category 5} > \text{Category 0}$$

中文优先级规则：若同时符合多个类别，请仅输出最高风险类别。

$$\text{类别 4} > \text{类别 3} > \text{类别 2} > \text{类别 1} > \text{类别 5} > \text{类别 0}$$

Analysis Material

The following is a chat context, with each line in the format "UserID: Message Content". Please analyze this conversation to understand the background.

chat_context: `[[input]]`

Analysis Target: Conduct a risk classification for the user associated with the target message containing the `<Signal Word>` tag.

Figure 8: Chain-of-thought prompt for LLM suicide risk evaluation.

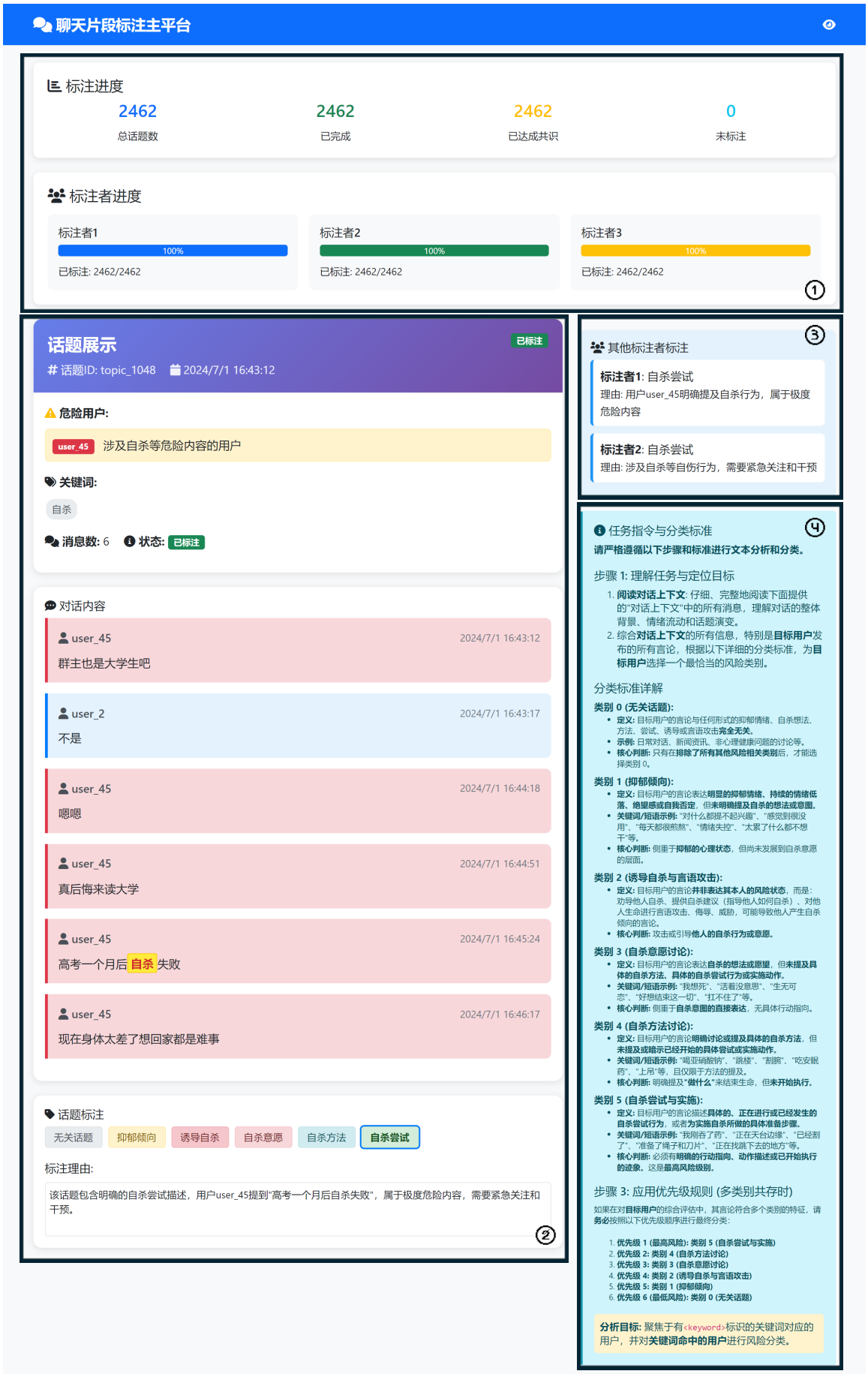


Figure 9: Online Annotation Platform

Figure 10: Consensus Review System

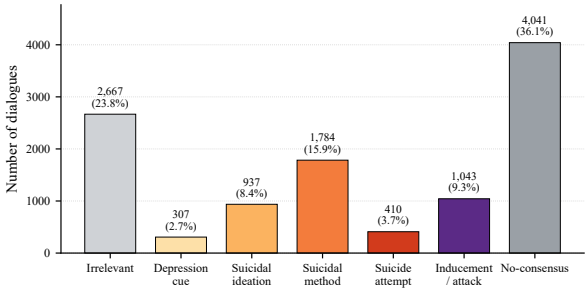


Figure 11: Distribution of final silver labels across the $N = 11{,}189$ candidate topics. *No-consensus* items ($\rho < 2/3$) are excluded from the downstream train/eval splits.

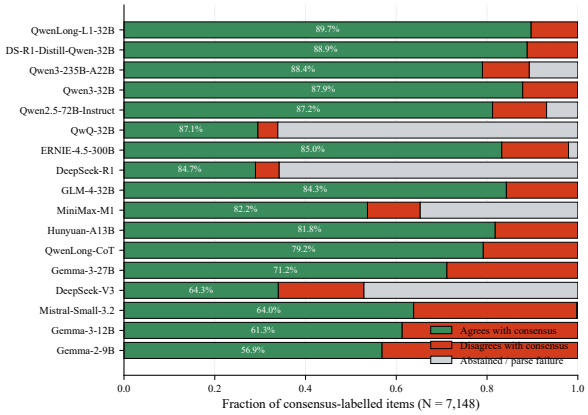


Figure 12: Per-model decomposition of the $N_c = 7{,}148$ consensus-labelled items. In-bar percentages give each model's agreement rate *conditional on a non-abstaining response*; the green portion equals the absolute number of agreed items divided by N_c.

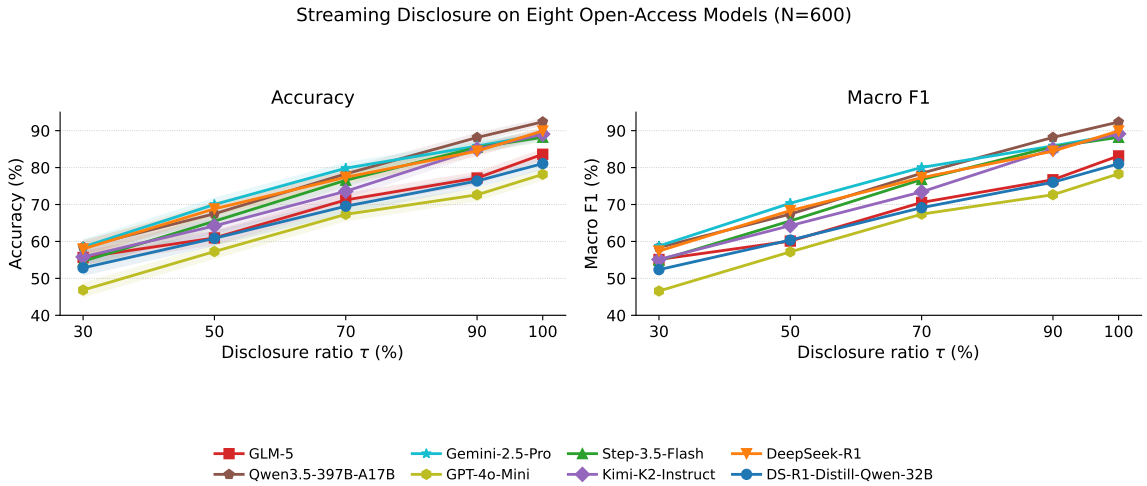


Figure 13: Accuracy (left) and Macro F1 (right) of the eight streaming baselines as a function of the disclosure ratio τ. Shaded bands on the Accuracy panel are ± 1 binomial standard error on the $N{=}600$ streaming validation set.

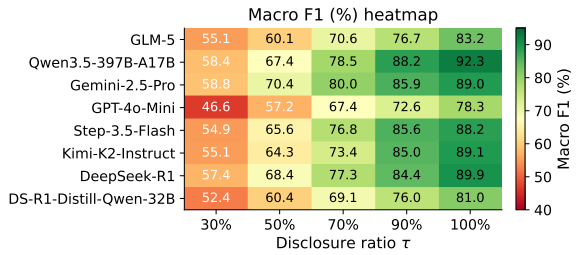


Figure 14: Macro F1 (%) heatmap. Rows are ordered as in Table 17; columns are increasing disclosure τ. The visible vertical gradient confirms a monotone improvement with τ for every model; the lighter bottom two rows visualise the persistent gap of the compact GPT-4O-MINI and distilled 32-B models relative to the six larger systems.

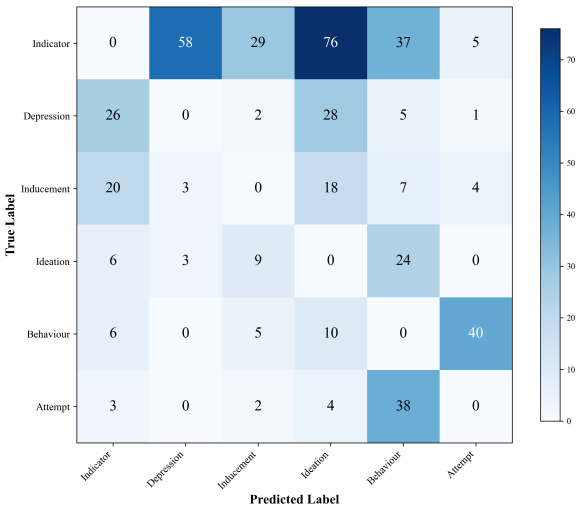


Figure 15: Confusion matrix for suicide risk misclassification

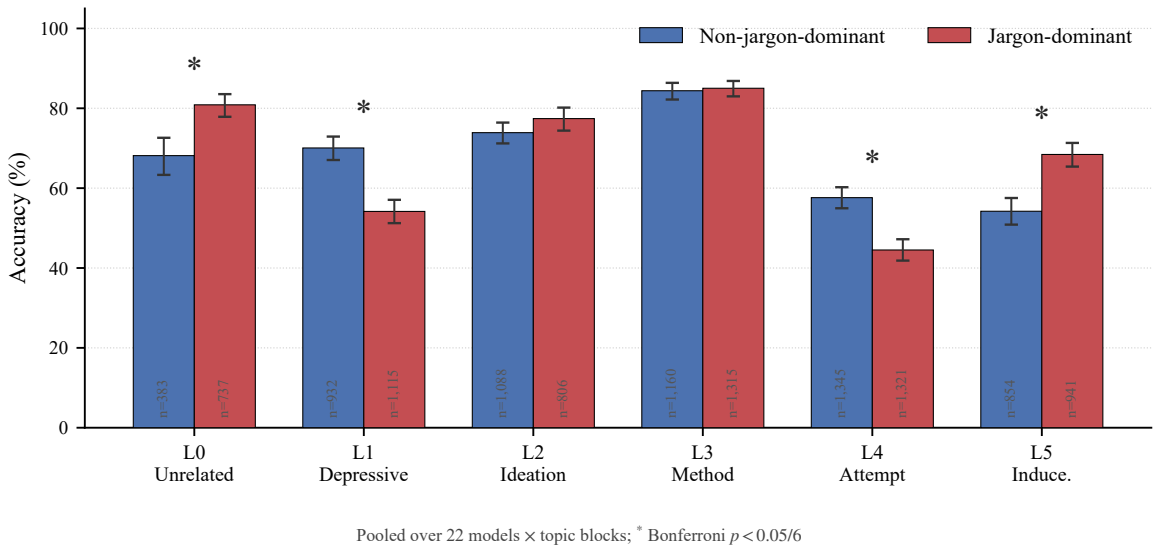


Figure 16: Per-class few-shot accuracy on the balanced validation set, split by jargon-dominant vs. non-jargon-dominant <keyword> tags. Error bars: 95% Wilson intervals; asterisks: Bonferroni-significant class gaps.

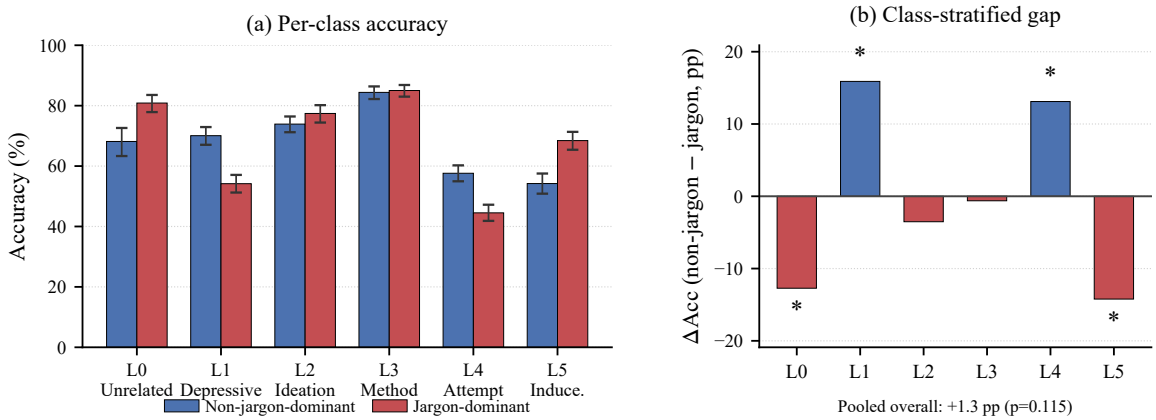


Figure 17: (a) Per-class accuracy by seed-word dominance type; (b) class-stratified Δ_{Acc} (non-jargon − jargon, pp).

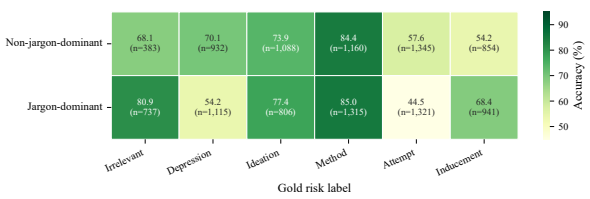


Figure 18: Pooled accuracy (%) by dominance type (rows) and gold label (columns).

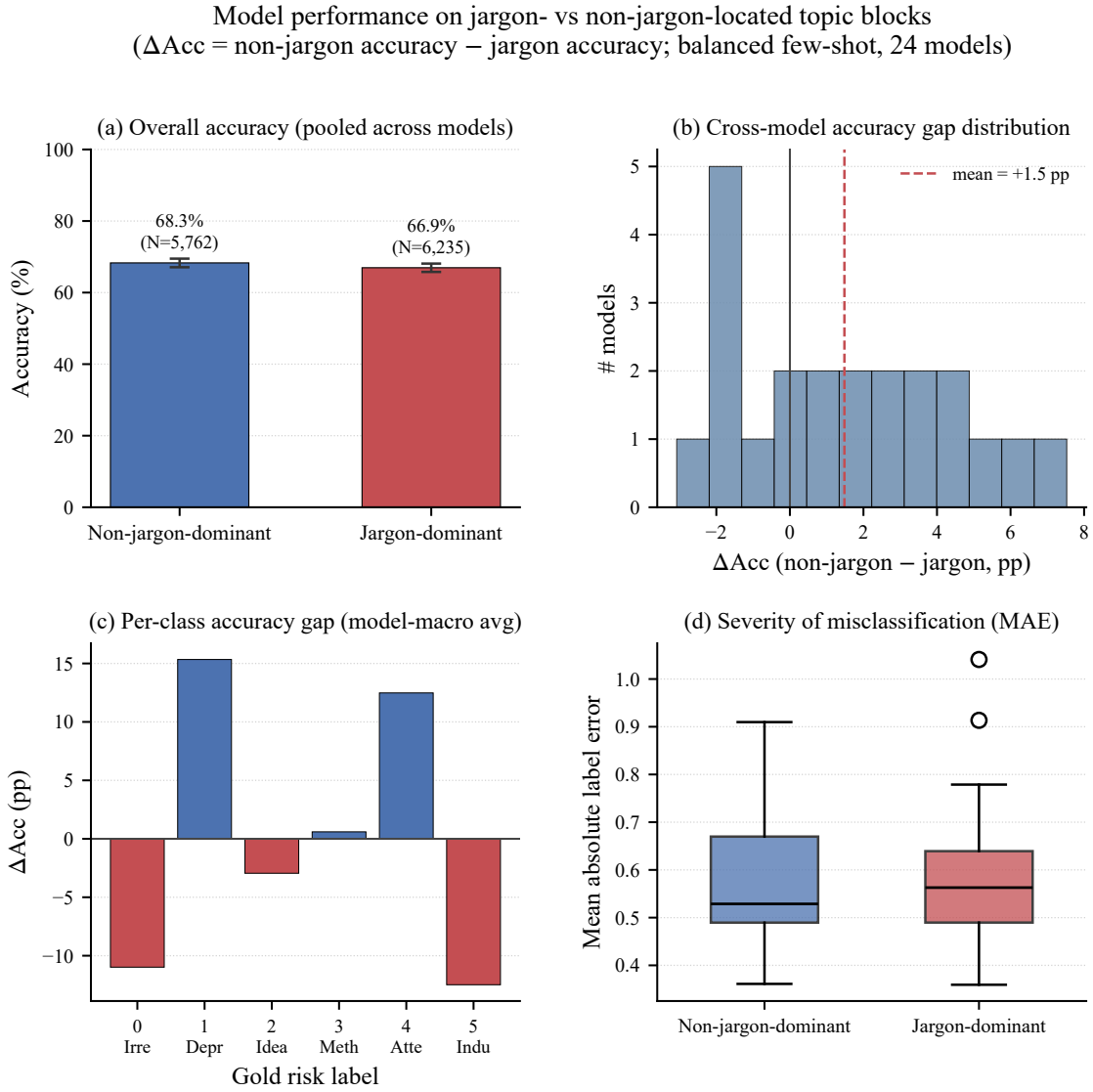


Figure 19: Summary across models: pooled accuracy, per-model Δ_{Acc}, per-class gap, and MAE by dominance type.