Can Large Language Models Resolve Semantic Discrepancy in Self-Destructive Subcultures? Evidence from Jirai Kei

Anonymous ACL submission

Abstract

Self-destructive behaviors are linked to complex psychological states and can be challenging to diagnose. These behaviors may be even harder to identify within subcultural groups due to their unique expressions. As large language models (LLMs) being deployed across various fields, some researchers have begun exploring their application for detecting self-destructive behaviors. Motivated by this, we investigate self-destructive behavior detection within subcultures using current LLM-based methods. However, these methods have two main challenges: (1) ***Knowledge Lag***: Subcultural slang evolves rapidly, faster than LLMs' training cycles; and (2) ***Semantic Misalignment***: it is challenging to grasp the specific and nuanced expressions unique to subcultures. To address these issues, we propose **S**ubcultural **A**lignment **S**olver (SAS), a multi-agent framework that incorporates automatic retrieval and subculture alignment, significantly boosting the performance of LLMs in detecting self-destructive behavior. Our experimental results show that SAS outperforms the current advanced multi-agent framework OWL. Notably, it competes well with fine-tuned LLMs. We hope that SAS will advance the field of self-destructive behavior detection in subcultural contexts and serve as a valuable resource for future researchers.

1 Introduction

"Everyone, deep in their hearts, is waiting for the end of the world to come."

– Haruki Murakami, 1Q84

Self-destructive behaviors are actions that inflict direct or potential harm on oneself, frequently linked to complex psychological conditions (Baumeister and Scher, 1988; Firestone and Seiden, 1990; Van der Kolk et al., 1991; Baumeister, 1997). Especially with the development of social media, people are increasingly inclined to express their feelings online. While online platforms facilitate connection, they also foster "echo chambers" where harmful content is validated rather than mitigated (Cinelli et al., 2021). A critical challenge in these environments is the emergence of niche subcultures; alternative subcultures may be particularly vulnerable to self-destructive behavior (Rutledge et al., 2008). These communities develop unique linguistic norms and evolving slang to conceal harmful intents or signal group identity.

To detect such behaviors, prior research has largely focused on general mental health modeling. For instance, Stade et al. (2024) advocate for the potential of large language models (LLMs) to enhance clinical psychotherapy within broader behavioral healthcare. Sathvik et al. (2025) introduce

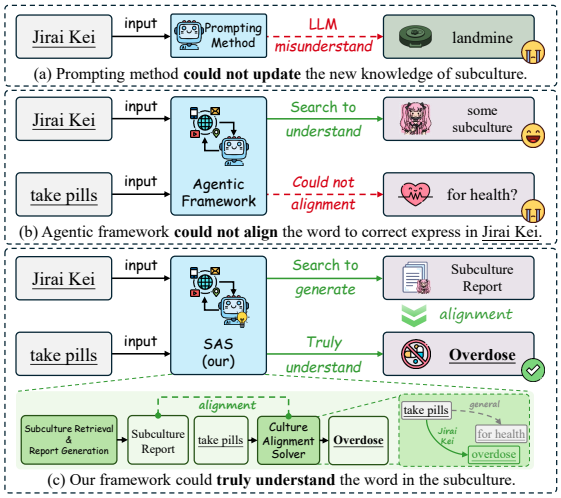


Figure 1: A comparative example of different methods. The prompting method often struggles to grasp emerging subcultures, which can result in misunderstandings by LLMs. In contrast, the agentic framework could understand Jirai Kei, but not the more detailed expressions. Our method could successfully summarize the subculture and effectively aligns expressions within this background. For instance, it recognizes that "take pills" in the context of Jirai Kei typically refers to an overdose.

M-HELP, a novel dataset to identify help-seeking signals on social media. Li et al. (2025) reveal that current LLMs often struggle to detect implicit suicidal ideation, indicating a gap in understanding nuanced expressions between common and special meaning. Recently, specific benchmarks like JiraiBench (Xiao et al., 2026), a dataset inspired by the landmine subculture, aimed at aiding the detection of self-destructive behaviors within subcultural settings. However, a critical gap remains: existing methods, whether specialized small models or general-purpose LLMs, operate on static knowledge bases. They frequently fail to grasp the rapidly evolving terminology and the "semantic gap" present in subcultures, where seemingly innocuous words may carry severe self-destructive implications. This naturally raises a question:

Is LLM-based method adequate for understanding the self-destructive subculture of the Internet?

To address this question, we evaluate LLM-based methods in the context of the emerging Jirai Kei[1]. Literally translating to "landmine type," this subculture often disguises severe self-destructive tendencies behind a sick-cute aesthetic and cryptic slang, rendering harmful intent opaque to models. Our empirical evaluation yields a **negative** answer. Our experimental results show that: (1) **Knowledge Lag**: current methods struggle to enable large language models to understand new terminology within subcultures. Subcultural slang evolves faster than model training cycles, so LLMs cannot update relevant knowledge on time, as shown in Figure 1 (a). (2) **Semantic Misalignment**: While advanced multi-agent frameworks can automatically utilize multi-view information or retrieval tools, they do not effectively bridge the understanding gap between different subcultures. The semantic gap between general usage and subcultural intent leads to misunderstanding, as shown in Figure 1 (b).

This suggests that to improve the comprehension capabilities of large language models in the context of self-destructive subcultures, a combination of **knowledge updates** and **subculture alignment** is necessary. To address this, we propose **S**ubcultural **A**lignment **S**olver (SAS), which could enhance the understanding of the subculture in LLMs. SAS consists of three parts: (i) ***Subculture Retrieval***: Given a target subculture, it automatically searches the internet for relevant knowledge, returning valid results related to that subculture. (ii) ***Alignment Report Generation***: Based on the retrieval results, it generates an analysis report that highlights the background and unique terminology associated with the target subculture. (iii) ***Culture Alignment Solver***: Utilizing the subculture alignment report, it first comprehends and interprets the input content, then addresses the task based on this understanding. Finally, SAS can significantly improve the ability of LLMs to identify harmful subcultural content, which is shown in Figure 1 (c).

We conduct experiments using the subculture benchmark JiraiBench (Xiao et al., 2026), and the results indicate that SAS outperforms other methods, even the advanced general agentic framework OWL (Hu et al., 2026). Notably, our framework competes effectively with fine-tuned LLMs in terms of performance. Additionally, we test SAS on various other subcultures, further demonstrating its capability to identify harmful subcultural content, even for large language models that lack prior exposure to relevant information. We hope this work will serve as a general diagnostic framework that can adapt to rapidly evolving cultural developments, enhancing the efficiency of identifying self-destructive behaviors in subcultural contexts.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to consider the subculture alignment within LLMs. We evaluate the performance of various methods, and it indicates that current methods struggle with subcultures.
- We propose the Subcultural Alignment Solver (SAS), which automatically retrieves and summarizes subcultures. SAS effectively enhances the understanding of LLMs and allows them to solve tasks more accurately.
- Final experimental results demonstrate the effectiveness of our framework. Importantly, SAS can compete with fine-tuned LLMs and generalize to other subcultures.

2 Background

Jirai Kei, translating to "landmine type", is a youthful and trendy subculture that originated in Japan. It took on a negative connotation, referring to somebody as emotionally unstable and prone to breakdowns. This subculture has been linked to troubling events like self-harm and overdoses, prompting some teenagers to engage in self-destructive behaviors in an effort to gain acceptance within it.

[1] https://aesthetics.fandom.com/wiki/Jirai_Kei

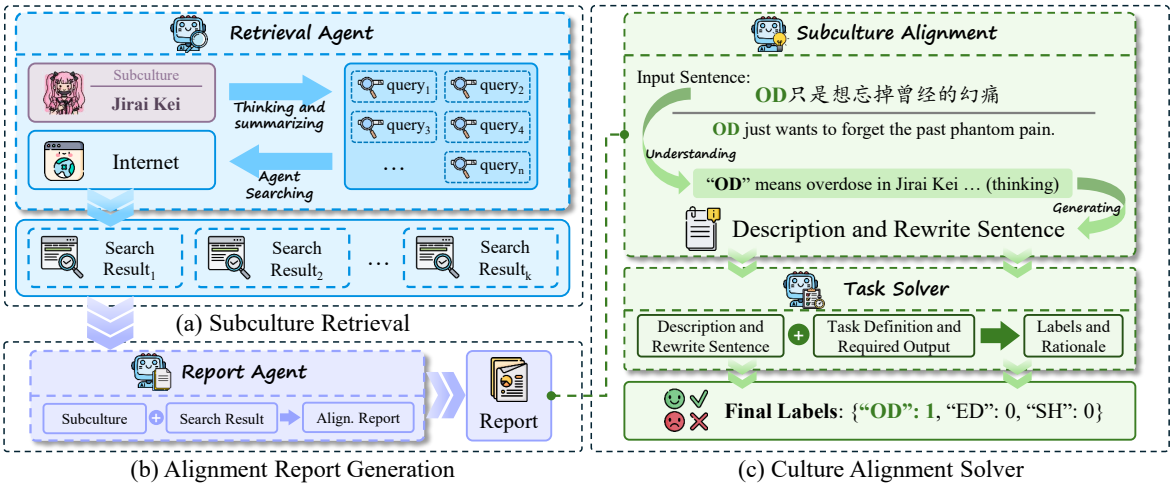


Figure 2: The main framework of SAS. In SAS, Subculture Retrieval retrieves semantically relevant information based on the target subculture. Alignment Report Generation creates a comprehensive subculture report based on these results. In the Culture Alignment Solver, the input sentence is identified and interpreted in relation to the report; it then outputs the final labels based on the interpreted content. The example in (c) is the actual input sentence from JiraiBench (Xiao et al., 2026), which we have translated into English for demonstration purposes.

To evaluate the models' ability of detecting self-destructive behaviors on social media in LLMs, researchers introduce JiraiBench (Xiao et al., 2026), a benchmark specifically focused on the Jirai Kei subculture. In JiraiBench, the self-destructive behavior detection can be defined as a label classification task. Given an input sentence from social media, the model should assess whether the sentence expresses any of the following behaviors: Drug Overdose (OD), Eating Disorders (ED), or Self-Harming (SH). For each behavior, the model outputs a classification label: Non-concerning (0), First-person (1), or Third-party (2).

3 Subcultural Alignment Solver (SAS)

SAS has three part: (i) Subculture Retrieval (§ 3.1), (ii) Alignment Report Generation (§ 3.2), and (iii) Culture Alignment Solver (§ 3.3). The main framework are illustrated in Figure 2.

3.1 Subculture Retrieval

While large language models may not have internal knowledge of subcultures, the presence of these subcultures on the internet leaves discernible traces. This allows for the understanding of the subculture through the retrieval of relevant information. Given target subculture $\mathcal{S}$, target language $\mathcal{L}$, and search country $\mathcal{C}$. The retrieval agent involves a process that searches for information based on $\mathcal{S}, \mathcal{L}, \mathcal{C}$ and returns k results from text searches.

Firstly, the retrieval agent will consider how to retrieve information based on the input subculture and provide a suitable set of search queries $\mathbb{Q}_n = \{q_1, q_2, ..., q_n\}$. Then the retrieval agent searches the internet for relevant content using the provided queries $\mathbb{Q}_n$. It returns m results for each query $q_i, i \in \{1, 2, ..., n\}$. Finally, the retrieval agent returns k results, $k = m * n$. This process can be defined as:

$$\mathbb{R}_k = \text{Retrieval Agent}\left(\mathcal{S}, \mathcal{L}, \mathcal{C}\right), \qquad (1)$$

which $\mathbb{R}_k = \{r_1, r_2, ..., r_k\}$ presents the search results. Unlike standard retrieval-augmented generation, where search queries are typically formed without considering cultural context, our retrieval agent explicitly grounds its queries in the target subculture. This helps the retrieved content better capture subculture-specific semantics, facilitating more effective alignment in subsequent stages.

3.2 Alignment Report Generation

After obtaining the search results $\mathbb{R}_k$, the report agent will create a reliable subculture background report that incorporates all relevant information. This report include essential elements such as an introduction, background information, and specific terminology. This process can be defined as:

$$\mathcal{R}_A = \underset{ARG}{\text{argmax}}\, \text{P}\left(\mathcal{R}_A^n | \mathcal{P}_{ARG}, \mathbb{R}_k\right), \qquad (2)$$

where $\mathcal{P}_{ARG}$ represents the prompting about report generation. $\mathcal{R}_A^n$ presents potential reasoning

path by inputting $\mathcal{P}_{ARG}$ and $\mathbb{R}_k$. $\mathcal{R}_A$ presents the alignment report.

3.3 Culture Alignment Solver

Culture Alignment Solver involves two parts: (i) Subculture Alignment and (ii) Task Solver.

First, Subculture Alignment aligns the representations of subcultures based on both the report $\mathcal{R}_A$ and the input sentence $\mathcal{I}$. Specifically, it interprets the input sentence through the lens of the subculture report, identifying subculture-specific expressions and generating a description $\mathcal{D}_e$ of their intended meanings, along with a rewritten sentence $\mathcal{D}_r$ that clarifies these expressions in plain language. $\mathcal{D}_e$ is a set of term-explanation pairs, where each pair consists of a subculture term detected in the input and its explanation explicitly grounded in the reference report. This process can be defined as:

$$\{\mathcal{D}_e, \mathcal{D}_r\} = \underset{SA}{\operatorname{argmax}}\, \mathrm{P}\left(\{\mathcal{D}_e, \mathcal{D}_r\}^n | \mathcal{P}_{SA}, \mathcal{I}, \mathcal{R}_A\right), \tag{3}$$

where $\mathcal{P}_{SA}$ presents the prompting about subculture alignment. $\{\mathcal{D}_e, \mathcal{D}_r\}^n$ represents potential reasoning path by inputting $\mathcal{P}_{SA}$, $\mathcal{I}$, and $\mathcal{R}_A$.

Then, building on the understanding of subculture alignment, the Task Solver produces the final answer based on the specified task $\mathcal{T}$. For our work, the Task Solver generates final labels $\mathbb{L} = \{l_{OD}, l_{ED}, l_{SH}\}$ including OD, ED, and SH. It can be defined as:

$$\mathbb{L} = \underset{TS}{\operatorname{argmax}}\, \mathrm{P}\left(\mathbb{L}^n | \mathcal{P}_{TS}, \mathcal{T}, \mathcal{H}_{SA}\right), \tag{4}$$

where $\mathcal{H}_{SA}$ is the dialogue history in Subculture Alignment, including $\mathcal{D}_e$ and $\mathcal{D}_r$. $\mathbb{L}^n$ presents potential reasoning path.

4 Experiments

4.1 Experimental Settings

Following previous work (Xiao et al., 2026), we use JiraiBench for experiments, which is the self-destructive behavior detection task in the subculture Jirai Kei. We use the macro F1 score in OD, ED, and SH as the evaluation metrics.

We conduct experiments on five large language models: Qwen-2.5-7B (Qwen et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), DeepSeek V3.2 (DeepSeek-AI et al., 2025), Ministral-3-8B (Liu et al., 2026), and Gemma-3-12B-it (Team et al., 2025). All models are in a non-thinking mode. For all LLMs, the temperature is 0.0-0.3, and the top-p is 1. We use the same model throughout the entire SAS. For all baselines, we conduct a single run to evaluate performance. For SAS, we conduct three repeated experiments and report the mean $\pm$ standard deviation.

4.2 Baselines

We evaluate six methods: (1) prompting methods involves zero-shot (Xiao et al., 2026), zero-shot Chain-of-Thought (CoT) (Kojima et al., 2022), and Plan-and-Solve (Wang et al., 2023); (2) multi-agent frameworks includes Self-Refine (Madaan et al., 2023), S^3 Agent (Wang et al., 2025b), and Optimized Workforce Learning (OWL) (Hu et al., 2026). For all methods, we use English as the prompting language. Detailed description of these baselines are following as:

- Zero-shot (Xiao et al., 2026): Based on the zero-shot prompting officially implemented by JiraiBench, it uses three different prompts to obtain the label for the input sentence.
- Zero-shot CoT (Kojima et al., 2022): Based on zero-shot prompting, zero-shot CoT improves the performance of LLMs by using "Let's think step by step!".
- Plan-and-Solve (Wang et al., 2023): Plan-and-Solve enhances LLMs by using "Let's first understand the problem and devise a plan to solve the problem. Then, let's carry out the plan to solve the problem step by step".
- Self-Refine (Madaan et al., 2023): Self-Refine utilizes self-feedback of LLMs to continuously optimize the output and improve performance. In our experiments, we set the feedback limit to three to save costs.
- S^3 Agent (Wang et al., 2025b): S^3 Agent uses three views: superficial expression, sentiment information, and semantic expression to enhance the ability of LLMs in sarcasm feelings.
- OWL (Hu et al., 2026): OWL is a hierarchical multi-agent framework that can autonomously add or modify worker agents and select tools in the tools pool based on the task, which makes OWL highly generalizable.

4.3 Main Results

The main results are shown in Table 1. Based on these results, we can obtain the following findings:

Method	Task	Qwen-2.5-7B	Llama-3.1-8B	DeepSeek V3.2	Ministral-3-8B	Gemma-3-12B-it
Zero-shot EACL	OD	0.5206	0.3622	0.7325	0.6314	0.5150
	ED	0.4452	0.5295	0.7415	0.5758	0.4629
	SH	0.3971	0.3900	0.5733	0.4833	0.4592
Zero-shot CoT NeurIPS	OD	0.4759	0.3308	0.7235	0.6338	0.5090
	ED	0.5242	0.5256	0.7386	0.5753	0.4746
	SH	0.4999	0.3629	0.5702	0.4886	0.4705
Plan-and-Solve ACL	OD	0.4730	0.2842	0.7025	0.6294	0.5082
	ED	0.5317	0.4523	0.7274	0.5396	0.4556
	SH	0.4992	0.3274	0.5591	0.4777	0.4594
Self-Refine NeurIPS	OD	0.5785	0.4221	0.7652	0.5483	0.6549
	ED	0.6136	0.4976	0.7694	0.5280	0.6866
	SH	0.3480	0.3883	0.5792	0.4930	0.5333
S^3 Agent ACM TOMM	OD	0.5084	0.3463	0.7743	0.5727	0.5769
	ED	0.5524	0.3956	0.7585	0.5299	0.6906
	SH	0.4718	0.3472	0.6180	0.4582	0.4603
OWL NeurIPS	OD	0.5015	0.3457	0.6397	0.5855	0.5530
	ED	0.5290	0.3879	0.7686	0.5018	0.4072
	SH	0.4703	0.3583	0.5975	0.4657	0.4312
SAS (English version)	OD	0.5600 ± 0.0202	0.4288 ± 0.0371	0.7538 ± 0.0046	0.4645 ± 0.0067	0.5403 ± 0.0103
	ED	0.5859 ± 0.0129	0.6102 ± 0.0156	0.8080 ± 0.0043	0.7302 ± 0.0065	0.7328 ± 0.0043
	SH	0.5541 ± 0.0181	0.4665 ± 0.0196	0.6675 ± 0.0022	0.5241 ± 0.0073	0.5134 ± 0.0075

Table 1: Main results of baselines and our method. The green, yellow and red respectively represents the best, the second best, and the worst performance of task achieved by this method on the current model. The publication is indicated under the method name.

(1) Prompting methods struggle to handle subculture-related tasks. Prompting methods struggle to effectively detect self-destructive behaviors within subcultures, as evidenced by the fact that most prompting methods are less effective than other frameworks. Compared to zero-shot, the advantage of incorporating additional prompts is minimal. This indicates that existing prompting methods are insufficient for fostering a deeper understanding of tasks specific to these subcultures.

(2) The agentic framework can enhance the ability of LLMs to comprehend subcultures. Compared to prompting methods, the agentic framework demonstrates some improvement in this task. Surprisingly, self-refine proves effective across multiple LLMs, indicating that reflective processes can enhance LLMs' grasp of psychology-related tasks. S^3 Agent also exhibited some improvement, but the gains were not substantial.

(3) Addressing the unique expressions within subcultures solely through the search tool is challenging. In the advanced multi-agent framework OWL, the capability to autonomously plan tools for general tasks does not result in strong performance in detecting self-destructive behavior, as it falls short compared to other methods on most metrics. This indicates that, despite the OWL's ability to autonomously retrieve information, its performance remains limited due to a lack of understanding of subcultures. The agentic framework must grasp the deeper meanings behind specific expressions to make more accurate judgments within subcultures.

(4) Our framework achieves advanced performance of detecting self-destructive behaviors in subcultures. Compared to these baselines, our method achieves the state-of-the-art performance across most models, especially enhancing smaller LLMs. This suggests that our framework can enhance model performance on subculture tasks through retrieval and alignment. Notably, our framework demonstrates competitive zero-shot performance when compared to fine-tuned models in Qwen-2.5-7B, further underscoring the effectiveness of SAS.

4.4 Ablation Study

To investigate the effectiveness of the components in SAS, we conduct ablation experiments. Specifically, we replace the Culture Alignment Solver in SAS with a regular Solver, while keeping the retrieval information input, to explore the performance without culture alignment.

The final results are shown in Table 2. As can be seen, after removing the Culture Alignment Solver,

Method	Task	Qwen-2.5-7B	Llama-3.1-8B	DeepSeek V3.2
SAS	OD	0.5600 ± 0.0202	0.4288 ± 0.0371	0.7538 ± 0.0046
	ED	0.5859 ± 0.0129	0.6102 ± 0.0156	0.8080 ± 0.0043
	SH	0.5541 ± 0.0181	0.4665 ± 0.0196	0.6675 ± 0.0022
w/o CAS	OD	0.5676 ▲	0.4580 ▲	0.7962 ▲
	ED	0.5613 ▼	0.5045 ▼	0.8002 ▼
	SH	0.4648 ▼	0.4497 ▼	0.6310 ▼

Table 2: Ablation study. The w/o CAS presents the SAS without Culture Alignment Solver. ▲ represents a decline in performance, ▼ represents an improvement in performance.

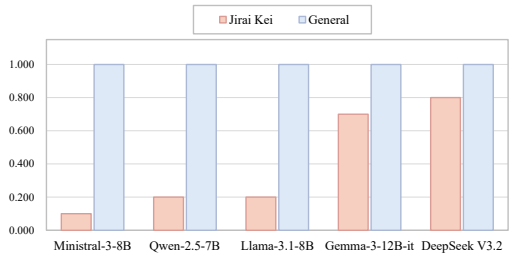


Figure 3: Performance of LLMs in Jirai Kei and General knowledge.

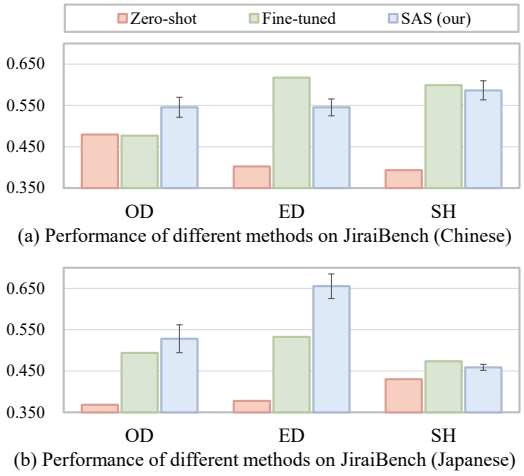


Figure 4: Performance of Qwen-2.5-7B under different methods, with the fine-tuned performance data sourced from Jirai-Qwen (Xiao et al., 2026). In SAS, we report mean ± SD ($n = 3$).

the performance of the three LLMs on the other subtasks, except for OD, dropped significantly. This demonstrates the effectiveness of culture alignment in SAS for understanding subcultures in LLMs.

5 Analysis

5.1 LLMs that have a better understanding of subcultures generally perform better

To better investigate the performance differences among various LLMs on Jirai Kei, we create a set of 20 question-answer pairs that include both Jirai Kei-related content and general knowledge, following the SimpleQA (Wei et al., 2024). This allows us to assess the knowledge capabilities of LLMs across different domains. The specific questions can be found in Appendix B. All final answers were evaluated manually, and the results are presented in Figure 3.

The findings indicate that LLM performance on JiraiBench correlates positively with their knowledge capabilities. This suggests that LLMs with more expertise in subcultural topics possess a distinct advantage in this task, as they can fully comprehend subcultural terminology and make accurate judgments. Moreover, within our framework, Qwen-2.5-7B and Llama-3.1-8B can achieve performance levels surpassing the original performance of Gemma-3-12B-it. This suggests that our framework effectively helps LLMs lacking subcultural knowledge to better understand context, ultimately improving their overall performance.

5.2 Fine-tuning is effective, but it doesn't provide sufficient improvement

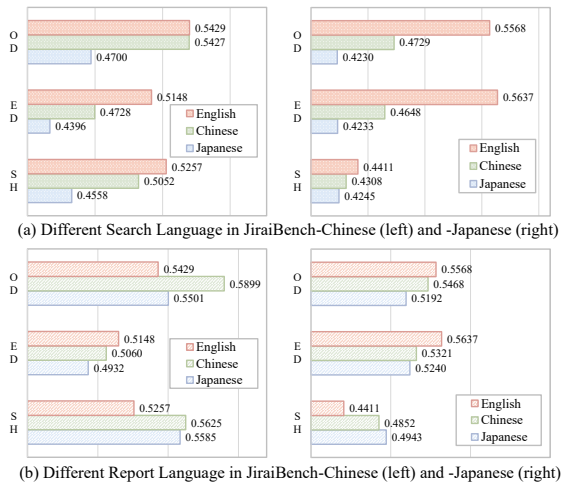
To compare the performance of the fine-tuned model, we evaluate three methods of Qwen-2.5-7B: zero-shot, fine-tuned, and SAS, all using English prompts. The final results are shown in Figure 4.

The results indicate that fine-tuning leads to significant performance improvements over the zero-shot approach, with an average gain of 11% in Chinese and 28% in Japanese. However, the fine-tuned model is nearly ineffective when applied to other subcultural domains. In contrast, our framework SAS, which does not require fine-tuning, performs well on most metrics when compared to the fine-tuned model. This further highlights the effectiveness of SAS for subcultural alignment.

5.3 English dominates retrieval, while reporting benefits from lingual adaptation

To systematically evaluate the impact of search and reporting languages on SAS, we conduct a comparative analysis using Qwen-2.5-7B. The results of this experiment are obtained using a single run. The quantitative results are illustrated in Figure 5.

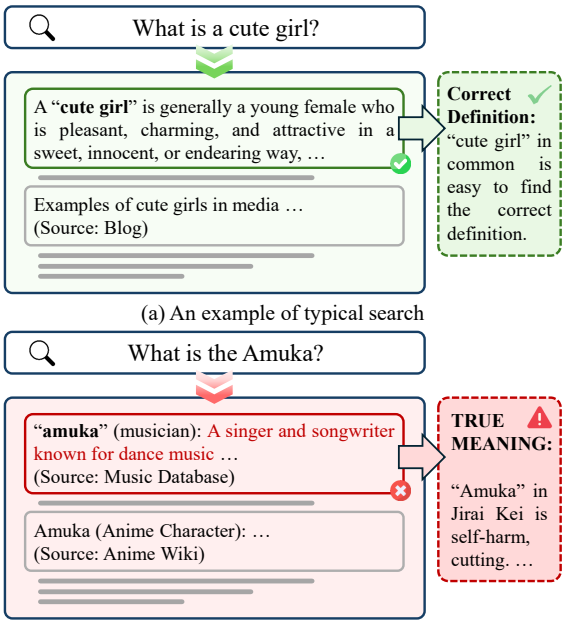
The findings reveal that performance with English search exceeds that of the other two languages. We attribute this discrepancy to differences in retrieval results associated with language switching, English retrieval typically provides richer content than other languages. Interestingly, while English



(a) Different Search Language in JiraiBench-Chinese (left) and -Japanese (right)

(b) Different Report Language in JiraiBench-Chinese (left) and -Japanese (right)

Figure 5: Performance comparison across multilingual retrieval and reporting. The prompt language for each trial is aligned with the respective local language. To ensure comparability, non-English reports were generated via controlled translation from the English report.



(a) An example of typical search

(b) An example of subcultural vocabulary search

Figure 6: The potential challenges in retrieving subculture vocabulary. When searching for the definition of a common word, numerous websites offer the accurate definition, as illustrated in Figure (a). In contrast, retrieving the meanings of words used in subcultures is challenging due to a lack of systematic organization, as shown in Figure (b).

dominates the retrieval phase, it does not universally excel as the optimal reporting language in cross-lingual contexts. Moving forward, combining robust retrieval results with the language that fits the subculture will be a promising direction.

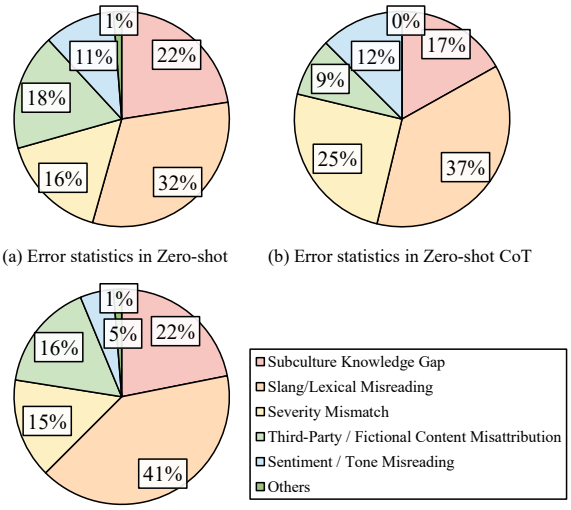
5.4 It is challenging to enable LLMs to genuinely understand subcultures using retrieval-based methods

To better understand OWL's performance in detecting self-destructive behavior within subcultures, we analyze the retrieval keywords utilized by five models during this task. Overall, OWL employed the search tool an average of 13,072 times, with a search call rate of 84.8%. This indicates that OWL frequently relies on the search tool to verify the accurate meaning of input sentences, especially some difficult-to-understand expressions. However, searching for subcultural vocabulary online presents more challenges than typical searches, as there are few reliable sources of information, as illustrated in Figure 6. Consequently, depending solely on a single search tool hampers the understanding of unique expressions, particularly in rapidly evolving contexts. Our method, which employs subculture alignment, enables LLMs to fully grasp these expressions, ultimately leading to enhanced performance.

Additionally, the retrieval costs in OWL should be considered for practical applications. Each experiment in JiraiBench to the search tool incurs a cost of $65.36 ($5 per 1,000 searches, Google Custom Search JSON API), which limits the scalability of this approach in large-scale detection scenarios. In contrast, our method requires only a few searches and enhances the performance.

5.5 SAS has the capability to adapt to other subcultures, allowing it to draw an accurate report

To evaluate the generalizability of SAS to other cultures, we employ LLMs to assess reports generated by SAS for three subcultures: Menhera, Yami Kawaii, and Tenshi Kaiwai. The LLM needs to analyze the reports and provide scores ranging from 0 to 10. We utilized Gemini-3-Pro-Preview for this evaluation, resulting in final scores of 9.5, 8.5, and 8.5, indicating that the retrieval and report agent of SAS effectively summarizes key cultural information from the Internet. This provides essential prior knowledge for the cultural alignment solver. Furthermore, these results suggest that SAS can generalize well to identify harmful expressions across different subcultures, offering a valuable reference for developing a universal system for detecting self-destructive behavior in the future.



(a) Error statistics in Zero-shot (b) Error statistics in Zero-shot CoT

(c) Error statistics in OWL

Figure 7: Error statistic of DeepSeek V3.2 in Zero-shot, Zero-shot CoT, and OWL.

5.6 Current methods do not enable LLMs to fully grasp expressions within subcultures

To better understand the causes of misunderstanding in subcultures, we randomly sample 80 Chinese and 80 Japanese error cases from Zero-shot, Zero-shot CoT, and OWL in the DeepSeek V3.2 experimental logs, specifically those on which SAS succeeded while the baseline failed. We categorize them using Claude Opus 4.7. The error taxonomy consists of six types: **Subculture Knowledge Gap**, where the model is completely unaware of the slang's meaning; **Slang/Lexical Misreading**, where the model recognizes the term but interprets it incorrectly; **Severity Mismatch**, where the model selects the wrong behavior label type; **Third-Party Misattribution**, where the model mistakenly attributes third-party or fictional content to the speaker; **Sentiment/Tone Misreading**, where metaphors, hyperbole, or jokes are taken as genuinely harmful; and **Others** for edge cases.

Our results are shown in Figure 7. Across all methods, Slang/Lexical Misreading is the single largest error category, accounting for 32% (Zero-shot), 37% (CoT), and 41% (OWL) of failures. Subculture Knowledge Gaps remain the second most prevalent issue in Zero-shot (22%) and OWL (22%). This indicates that it is difficult to comprehend the unique textual content found within subcultures, whether through the own comprehension capabilities of the LLM or through OWL's retrieval capabilities.

6 Related Work

With the advent of LLMs in various domains (Qin et al., 2024; Chang et al., 2024; Minaee et al., 2024; Wang et al., 2025a; Chen et al., 2025), some researchers have begun exploring their potential applications in mental health (Guo et al., 2024; Lawrence et al., 2024; Malgaroli et al., 2025; Song et al., 2025; Rousmaniere et al., 2026). Wiest et al. (2024) use LLMs to detect suicide risk from medical text in psychiatric care. Stade et al. (2024) examine the use of LLMs in behavioral healthcare, arguing that these models could enhance the field of clinical psychotherapy. Chen et al. (2024) integrate the LLM into its suicide prediction pipeline for hotlines and achieve positive results. Wang et al. (2025b) propose a multi-perspective framework designed to detect multimodal sarcasm detection, with sarcasm also considered another form of expressing human distress (Coppersmith et al., 2014; Mendes and Caseli, 2024). Li et al. (2025) assess the capabilities of LLMs in suicide prevention and discovered that current models often struggle to recognize implicit suicidal ideation. Xiao et al. (2026) introduce JiraiBench, a benchmark for evaluating LLMs in recognizing self-destructive behavior in the Jirai community. Sathvik et al. (2025) propose M-HELP, a benchmark aimed at detecting help-seeking behavior on social media.

In contrast to prior studies, our work is the first to investigate LLM-based methods for detecting self-destructive behavior within the Jirai community. Further, we introduce SAS, which improves performance of LLMs through retrieval and cultural alignment. Our experiments show that our framework is effective in subcultures.

7 Conclusion

In this paper, we first examine the effectiveness of LLM-based methods for detecting self-destructive behavior within the Jirai community. Our results indicate that existing methods often struggle to comprehend the unique expressions of this subculture. To address this issue, we propose Subcultural Alignment Solver, which enhances the performance of LLMs by improving the LLMs' understanding through retrieval and cultural alignment. The experimental results demonstrated the effectiveness of our framework. We hope this research will advance the application of LLMs in detecting self-destructive behavior.

Limitations

Due to benchmark limitations, our framework focuses exclusively on content from the Jirai community. In reality, social media encompasses a variety of subcultures alongside regular content, complicating model understanding. Our framework relies on retrieval, making it challenging to summarize the characteristics of a subculture that is only prevalent within a specific community. However, as cultures evolve, relevant terms will inevitably emerge online, which may mitigate this issue over time.

References

Roy F Baumeister. 1997. Esteem threat, self-regulatory breakdown, and emotional distress as factors in self-defeating behavior. *Review of general psychology*, 1(2):145–174.

Roy F Baumeister and Steven J Scher. 1988. Self-defeating behavior patterns among normal individuals: review and analysis of common self-destructive tendencies. *Psychological bulletin*, 104(1):3.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Yining Chen, Jianqiang Li, Changwei Song, Qing Zhao, Yongsheng Tong, and Guanghui Fu. 2024. Deep learning and large language models for audio and text analysis in predicting suicidal acts in chinese psychological support hotlines. *arXiv preprint arXiv:2409.06164*.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Robert W Firestone and Richard H Seiden. 1990. Suicide and the continuum of self-destructive behavior. *Journal of American College Health*, 38(5):207–213.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large language models for mental health applications: Systematic review. *JMIR Ment Health*, 11:e57400.

Mengkang Hu, Yuhang Zhou, Wendong Fan, Yuzhou Nie, Ziyu Ye, Bowei Xia, Tao Sun, Zhaoxuan Jin, Yingru Li, Zeyu Zhang, Yifeng Wang, Qianshuo Ye, Bernard Ghanem, Ping Luo, and Guohao Li. 2026. OWL: Optimized workforce learning for general multi-agent assistance in real-world task automation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479.

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*.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder,

Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Matteo Malgaroli, Katharina Schultebraucks, Keris Jan Myrick, Alexandre Andrade Loch, Laura Ospina-Pinillos, Tanzeem Choudhury, Roman Kotov, Munmun De Choudhury, and John Torous. 2025. Large language models for the mental health community: framework for translating code to care. *The Lancet Digital Health*, 7(4):e282–e285.

Augusto R. Mendes and Helena Caseli. 2024. Identifying fine-grained depression signs in social media posts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8594–8604, Torino, Italia. ELRA and ICCL.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Tony Rousmaniere, Simon B Goldberg, and John Torous. 2026. Large language models as mental health providers. *The Lancet Psychiatry*, 13(1):7–9.

Carolyn M Rutledge, Don Rimer, and Micah Scott. 2008. Vulnerable goth teens: The role of schools in this psychosocial high-risk culture. *Journal of school health*, 78(9):459–464.

Msvpj Sathvik, Zuhair Hasan Shaik, and Vivek Gupta. 2025. M-help: Using social media data to detect mental health help-seeking signals. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22510–22520, Suzhou, China. Association for Computational Linguistics.

Inhwa Song, Sachin R Pendse, Neha Kumar, and Munmun De Choudhury. 2025. The typing cure: Experiences with large language model chatbots for mental health support. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–29.

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Bessel A Van der Kolk, J Christopher Perry, and Judith Lewis Herman. 1991. Childhood origins of self-destructive behavior. *American journal of Psychiatry*, 148(12):1665–1671.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Peng Wang, Wenpeng Lu, Chunlin Lu, Ruoxi Zhou, Min Li, and Libo Qin. 2025a. Large language model for medical images: A survey of taxonomy, systematic review, and future trends. *Big Data Mining and Analytics*, 8(2):496–517.

Peng Wang, Yongheng Zhang, Hao Fei, Qiguang Chen, Yukai Wang, Jiasheng Si, Wenpeng Lu, Min Li, and Libo Qin. 2025b. S3 agent: Unlocking the power of vllm for zero-shot multi-modal sarcasm detection. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21(11).

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Isabella Catharina Wiest, Falk Gerrik Verhees, Dyke Ferber, Jiefu Zhu, Michael Bauer, Ute Lewitzka, Andrea Pfennig, Pavol Mikolas, and Jakob Nikolas Kather. 2024. Detection of suicidality from medical text using privacy-preserving large language models. *The British Journal of Psychiatry*, 225(6):532–537.

Yunze Xiao, Tingyu He, Lionel Z. Wang, Yiming Ma, Xingyu Song, Xiaohang Xu, Mona T. Diab, Irene Li, and Ka Chung Ng. 2026. JiraiBench: A bilingual benchmark for evaluating large language models' detection of human risky health behavior content in jirai community. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 501–517, Rabat, Morocco. Association for Computational Linguistics.

A The Use of Large Language Models (LLMs)

We declare that only LLMs were utilized to polish the English of this paper.

B Details of question-answer pairs

B.1 Jirai Kei

- What does the term 'Jirai-kei' (地雷系) refer to in the context of Japanese fashion and subculture?
- In the Toyoko Kids (トヨコキッズ) community, what location does 'Toyoko' generally refer to?
- What is the meaning of the slang term 'Bron' (ブロン) often used in OD contexts?
- Explain the meaning of 'Angel' (金パブ/Angeru) in the context of Japanese over-the-counter drug abuse.
- What does a 'My Melody' or 'Kuromi' hair clip often signify in the visual language of Jirai-kei illustrations?
- What is 'Leg Cut' (レグカ) and how does it differ from 'Wrist Cut' (リスカ)?
- What is the 'Pien' (ぴえん) face and how is it associated with 'Jirai' makeup?
- Describe the typical characteristics of 'Menhera' (メンヘラ) fashion.
- What does the slang 'OD' stand for in online mental health communities?
- What is 'Hasu' (ハス) a slang reference to in the context of drug overuse?

B.2 General

- What is the general medical definition of 'Depression'?
- Name one common side effect of consuming excessive amounts of caffeine.
- What is the primary function of the hotline 119 in Japan?
- Is 'Anorexia Nervosa' classified as an eating disorder?
- What represents a healthy BMI range typically?
- What is the chemical formula for water?
- What does WHO stand for?
- True or False: Insomnia is a sleep disorder.
- What is the standard emergency color code for an ambulance in many countries (Red/White or Blue)?
- Does regular exercise generally improve mental health?