# Adaptive Multi-Agent LLM Pipelines for Reliable Self-Harm Risk Screening

Meghana Karnam
Johns Hopkins University
Baltimore, USA

Ananya Joshi
Johns Hopkins University
Baltimore, USA

## Abstract

Hospitals and crisis organizations are deploying multi-agent LLM pipelines for self-harm risk screening, but existing approaches rely on heuristic majority voting with no principled basis for determining when a decision is reliable. Incorrectly flagging safe content increases clinician burden and risks alert fatigue; missing self-harm risk can delay crisis response. We present a statistical framework for multi-agent LLM pipelines structured as directed acyclic graphs (DAGs) that addresses both failure modes. We model each agent as a stochastic categorical decision and introduce (1) tighter agent-level confidence bounds using the DKW inequality, (2) a bandit-based adaptive sampling strategy that allocates LLM calls based on input difficulty, and (3) a novel identify-or-escalate termination condition that routes genuinely uncertain cases to human clinician review rather than forcing a potentially wrong label. On the AEGIS 2.0 behavioral health benchmark (N=161), our approach reduces the false positive rate from 0.159 to 0.095 relative to single-agent models, a 40% reduction in unnecessary escalation of safe content, with no change in false negative rates. Results hold on a held-out SWMH Reddit sample (N=250), suggesting the finding generalizes across source domains. These results indicate that principled adaptive sampling offers a meaningful path toward reliable clinical decision support in self-harm screening.

## CCS Concepts

• **Computing methodologies** → **Machine learning**.

## Keywords

multi-agent LLM, adaptive sampling, self-harm risk screening, bandit algorithms, clinical decision support, behavioral health AI

**ACM Reference Format:**
Meghana Karnam and Ananya Joshi. 2026. Adaptive Multi-Agent LLM Pipelines for Reliable Self-Harm Risk Screening. In *Proceedings of AI4Mental Workshop on AI for Cognitive and Mental Health Support (KDD'26).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## Introduction

Hospitals and behavioral health organizations are increasingly exploring AI systems to support psychiatric intake, triage, and crisis response [14, 16, 17]. In self-harm screening, these systems must follow strict, institution-specific standard operating procedures (SOPs) that define how risk is assessed, when escalation is required, and what constitutes a safe response. The consequences of failure are asymmetric and severe: missing self-harm risk in a patient interaction can delay crisis response with potentially life-threatening consequences, while excessive escalation of safe content creates alert fatigue and strains already limited clinical resources [1].

Addressing this gap requires systems that are not only accurate but scalable, capable of maintaining reliability as patient interaction volume grows without proportional increases in clinician burden. Recent work proposes operationalizing these SOPs using multi-agent LLM systems, where each specialized agent handles a distinct clinical responsibility and uncertain cases are passed along structured workflows [12]. Modular agent designs mirror real-world care coordination, a frontline counselor screens incoming content, uncertain cases proceed to a clinical supervisor, and edge cases reach a compliance officer before any content reaches a patient. These designs improve robustness over single models and are especially relevant to behavioral health, where decision-making is inherently staged and safety-critical.

The core problem is that current multi-agent systems are largely heuristic. They rely on simple majority voting over a fixed number of stochastic LLM calls, with no principled way to know when a decision is reliable, how many samples are sufficient, or how uncertainty propagates through the pipeline. Without formal guarantees, practitioners must rely on ad hoc choices that may produce either missed detections or excessive escalation. Existing empirical evaluations offer only point estimates of performance—insufficient for clinical deployment, auditing, or regulatory compliance.

This motivates a statistical framework for reliable clinical decision support that addresses:
(A) *When should a behavioral health AI system act on its own assessment of patient risk, and when should it defer to a human clinician for review?*
(B) *How can a system reduce missed detections of self-harm risk without increasing unnecessary escalation burden on clinicians?*
(C) *How do system-level errors accumulate over time as patient interaction volume grows under real-world deployment?*

*Contributions.*

- **Confidence-guided clinical decision thresholds:** We introduce a method that quantifies system confidence at each routing decision, distinguishing cases that can be safely resolved from those requiring clinician review. This provides explicit reliability guarantees at each stage of care rather than relying on fixed heuristics.


*KDD'26, Jeju, Korea*

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- **Adaptive sampling based on case complexity:** We develop a strategy that allocates more LLM calls to ambiguous or high-risk cases while resolving clear ones efficiently, reducing missed high-risk detections without increasing unnecessary clinician burden.
- **Formal guarantees for safe deployment:** We prove that system errors grow logarithmically rather than linearly with patient volume, providing a foundation for auditing and deploying multi-agent AI systems in safety-critical behavioral health settings.
- **Interpretable escalation decisions:** The framework produces explicit confidence estimates at each routing step, distinguishing budget exhaustion (the algorithm is uncertain) from an active `escalate` label (the model is confident the case requires specialist review). This transparency supports clinician trust and auditability in deployment.

## Related Work

Detecting suicide risk and self-harm in clinical and social media text has a long history in NLP. Early approaches applied machine learning to clinical notes [7, 8] and social media platforms [10, 20]. Benchmarks grounded in validated clinical instruments such as the Columbia Suicide Severity Rating Scale (C-SSRS) [5] have established evaluation standards for NLP models in this space. AEGIS 2.0 [6] extends this to LLM safety evaluation, providing human-annotated responses across safety categories including self-harm, and serves as our primary benchmark.

Large language models have introduced new evaluation challenges for safety-sensitive mental health tasks. Most existing approaches rely on single-pass classification (LLM-as-a-judge) or fixed-sample voting (LLM-as-a-jury), with no mechanism for determining how many samples are sufficient or how errors compound when multiple agents are involved. The most closely related system [11] organizes behavioral health evaluation agents into a DAG and demonstrates accuracy gains over single-agent approaches using Monte Carlo sampling at each node. Other multi-agent systems [19] similarly route inputs through structured pipelines. None provide formal guarantees on when a routing decision is trustworthy, how many samples are needed, or how errors accumulate under deployment.

The mathematical tools we apply are well established. The Dvoretzky-Kiefer-Wolfowitz (DKW) inequality [3, 13] bounds estimation error across all categories simultaneously, avoiding the $\ln K$ penalty of Hoeffding-based approaches. Our contribution is instantiating these tools for DAG-structured clinical decision pipelines, introducing an identify-or-escalate termination condition, and proving system-level regret bounds [2, 9] showing logarithmic error growth under deployment.

## Methods

As shown in Figure 1, we organize LLM-based behavioral health evaluation agents into a DAG mirroring the staged human review process used in clinical SOPs. The **Worker** node performs first-pass content screening, analogous to a frontline crisis counselor. Inputs that cannot be confidently resolved escalate to the **Risk** node (secondary clinical review), then to the **Legal** node (institutional compliance review) before any content reaches a human clinician.

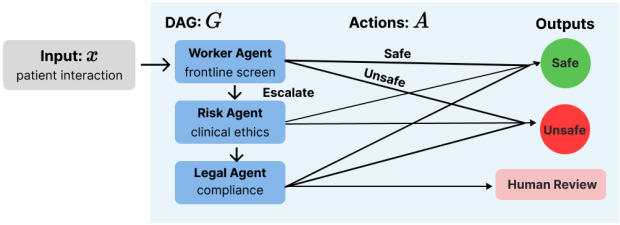


**Figure 1: DAG architecture of the multi-agent evaluation pipeline. Inputs unresolved at the Worker node escalate to Risk, then Legal. If all three nodes return `escalate`, the input is referred to a human clinician; otherwise a label is committed at whichever node first reaches a confident decision.**

| Symbol | Definition |
|---|---|
| $G = (\mathcal{S}, E)$ | DAG of agent nodes and escalation edges |
| $\mathcal{S}, K$ | Agent nodes ($\lvert\mathcal{S}\rvert = 3$); label arms ($K = 3$) |
| $p_s(c\,;x)$ | True probability node $s$ returns label $c$ on input $x$ |
| $c^*(s,x)$ | Most probable label at node $s$ for input $x$ |
| $\Delta_s(x)$ | Gap: $p_s(c^*) - \max_{c\neq c^*} p_s(c)$ |
| $\Delta_{\min}$ | Minimum gap over all nodes and inputs |
| $T, T_c$ | Total deployment episodes; per-arm pulls within one episode |
| $B, \delta$ | Sampling budget per node; confidence failure probability |
| $n$, $w_c$ | Fixed samples (majority vote); confidence width for arm $c$ |

**Table 1: Notation used throughout the paper.**

Each node receives the original input and produces a label from the shared action space $\mathcal{A} = \{\texttt{safe}, \texttt{unsafe}, \texttt{escalate}\}$, $K = 3$. A `safe` or `unsafe` label terminates routing; `escalate` passes the input to the next node. If all three nodes return `escalate`, the input goes to human review. Notation is summarized in Table 1.

## Clinical Grounding of the Action Space

All three agent nodes are grounded in the C-SSRS [15], a validated clinical instrument for suicide risk assessment widely adopted by hospitals and crisis programs.

• **Safe** — content is appropriate and supportive, consistent with safe messaging guidelines from SPRC and AFSP.

• **Unsafe** — content poses direct patient safety risk meeting C-SSRS criteria, including suicidal ideation, self-harm intent, or any content a clinician would flag for immediate intervention.

• **Escalate** — the node cannot make a confident determination; the input requires review by the next specialized agent. Reserved for genuine uncertainty, not content sensitivity.

## Adaptive Sampling

Prior work [11] uses fixed-sample majority voting: draw $n$ independent LLM samples and return the plurality winner. This applies the same computational budget to every input regardless of clinical

**Algorithm 1** Adaptive Sampling for Node $s$, Input $x$

**Require:** Budget $B \in \mathbb{N}$, confidence $\delta \in (0,1)$
1: $\mathcal{A}_{\text{act}} \leftarrow \mathcal{A}$; $T_c \leftarrow 0\ \forall c \in \mathcal{A}$
2: **while** $|\mathcal{A}_{\text{act}}| > 1$ **and** $\sum_c T_c < B$ **do**
3: **for** each $c \in \mathcal{A}_{\text{act}}$ **do**
4: Draw one LLM sample; update $\hat{p}_s(c)$, $T_c \leftarrow T_c + 1$
5: **end for**
6: $w_c \leftarrow \sqrt{\dfrac{\ln(4KT_c^2/\delta)}{2T_c}}$ for each $c \in \mathcal{A}_{\text{act}}$
7: $c_{\max} \leftarrow \arg\max_c \hat{p}_s(c)$
8: Remove $c$ from $\mathcal{A}_{\text{act}}$ if $\hat{p}_s(c_{\max}) - w_{c_{\max}} > \hat{p}_s(c) + w_c$
9: **end while**
10: **if** $|\mathcal{A}_{\text{act}}| = 1$ **then**
11: **return** surviving arm
12: **else**
13: **return** `escalate`
14: **end if**

complexity—if the first three samples all return `unsafe`, the system still draws $n-3$ additional samples that add no information while incurring API cost and latency.

We replace this with the adaptive sampling procedure in Algorithm 1, following Even-Dar et al. [4]. Each node on each input defines a $K$-armed bandit problem: arm $c \in \mathcal{A}$ has unknown mean $p_s(c;x)$ (the true probability node $s$ returns label $c$ on input $x$), and one pull is one LLM API call. The goal is to identify the best arm $c^* = \arg\max_c p_s(c;x)$ within budget $B$.

The algorithm initializes all $K = 3$ labels as active arms and eliminates clearly suboptimal options, stopping as soon as one label survives. In each round, it pulls every active arm and updates empirical estimates (lines 3–5), then computes a shrinking confidence interval for each arm (line 6). An arm is eliminated when its most optimistic estimate falls below the most pessimistic estimate of the leading arm (lines 7–8). When one arm survives, it is returned as the routing decision (lines 9–10).

The key clinical contribution is the termination condition. Standard adaptive sampling [4] forces a label when the budget is exhausted. Our variant instead returns `escalate` (lines 11–12), preserving the safety guarantee that genuinely uncertain inputs always reach a more specialized reviewer rather than receiving a potentially wrong label. This identify-or-escalate condition ensures the pipeline degrades gracefully under uncertainty—a property essential for safe clinical deployment.

## Theoretical Results

*Node-level correctness.* The shared action space $\mathcal{A}$ across all three nodes is the key structural choice enabling a unified theoretical treatment. With it, the same guarantee applies uniformly: with probability at least $1-\delta$, Algorithm 1 never eliminates the true best arm $c^*$—it either returns $c^*$ or returns `escalate`. This rests on a Good Event $\mathcal{G}$ requiring all confidence intervals to contain the true probability at every round:

$$\mathcal{G} = \left\{ \forall c \in \mathcal{A},\ \forall m \geq 1 : |\hat{p}_s^{(m)}(c) - p_s(c)| \leq \sqrt{\frac{\ln(4KT_c^2/\delta)}{2m}} \right\}.$$

$\mathcal{G}$ holds with probability $\geq 1-\delta$ by a union bound over $K$ arms using $\sum_{m=1}^{\infty} 1/m^2 < 2$ (proof in Appendix B), derived from the DKW inequality applied to the categorical distribution over $\mathcal{A}$ (Appendix A).

*Sample complexity.* The minimum pulls needed to identify $c^*$ is $n^* = 2\ln(2/\delta)/\Delta_s(x)^2$, where $\Delta_s(x)$ is the probability gap between the best and second-best label. Clinically clear inputs resolve with few samples; ambiguous cases exhaust the budget and escalate to human review. This provides theoretical grounding for the empirical inflection between $B = 75$ and $B = 100$ in the results.

*System-level reliability over time.* The adaptive policy achieves $O(\log T)$ cumulative regret over $T$ deployment episodes, compared to $O(T)$ under fixed majority voting. For a hospital deploying this system, doubling patient volume increases total errors by only $\ln 2 \approx 0.69$ rather than proportionally (proof in Appendix C).

## Experimental Setup

We evaluate on two behavioral health datasets spanning different source domains.

**AEGIS 2.0** [6] is a content safety benchmark comprising human-annotated LLM responses. We filter to suicide and self-harm content across all three splits, yielding N=163 unique examples after deduplication (44 safe, 117 unsafe).[1] Ground truth labels are from the `response_label` field following Joshi and Rudow [11]. This is our primary benchmark.

**SWMH** [10] is a Reddit mental health dataset. We use subreddit membership as a proxy for clinical ground truth: SuicideWatch posts are labeled `unsafe` and posts from four general mental health communities (depression, anxiety, bipolar, general support) are labeled `safe`. We sample 50 posts per subreddit for a stratified evaluation set of N=250 (50 unsafe, 200 safe), used as a held-out generalization test.

We evaluate ten conditions: **Single-agent** (one LLM call to Worker, no DAG routing); **Majority vote** MV $n \in \{1,3,5\}$ (full DAG with $n$ calls per node, plurality winner); and **Adaptive Sampling** with budgets $B \in \{10, 50, 75, 100, 124, 150\}$. $B = 100$ is the primary adaptive condition: the sample complexity bound predicts roughly 90 total pulls for inputs with $\Delta_s(x) \approx 0.5$, and $B = 100$ provides a buffer above this threshold.

We report accuracy, FPR, FNR, escalation rate, and average pulls per input, all with 95% Wilson score confidence intervals [18]. FPR and FNR are computed over non-escalated examples. For SWMH, SuicideWatch FNR (SW FNR) is the primary clinical metric, as it captures the rate of missed acute crisis cases. Full implementation details are in Appendix D.

## Results

## Experiment 1: AEGIS 2.0

*False positive rate.* $B = 100$ achieves FPR = 0.095 (95% CI [0.037, 0.221]), compared to 0.159 (95% CI [0.079, 0.294]) for single-agent—a 40% reduction in incorrect flagging of safe content. This is the primary clinical finding: in behavioral health settings, false positives translate directly to unnecessary escalations, contributing to alert

[1] Two examples excluded due to Unicode encoding errors. Results reported over N=161.

| AEGIS 2.0 (N=161) | | | | |
|---|---|---|---|---|
| **Condition** | **Acc.↑** | **FPR↓** | **FNR★↓** | **Esc.↓** |
| Single-agent | 0.752 [0.679, 0.813] | 0.159 [0.079, 0.294] | 0.283 [0.208, 0.372] | 0.025 |
| MV $n = 1$ | 0.760 [0.687, 0.820] | 0.114 [0.050, 0.240] | 0.290 [0.214, 0.379] | 0.019 |
| MV $n = 3$ | 0.760 [0.687, 0.820] | 0.114 [0.050, 0.240] | 0.290 [0.214, 0.379] | 0.019 |
| MV $n = 5$ | 0.753 [0.681, 0.814] | 0.159 [0.079, 0.294] | 0.281 [0.206, 0.369] | 0.019 |
| $B = 10$ | — | — | — | 1.000 [0.977, 1.000] |
| $B = 50$ | — | — | — | 1.000 [0.977, 1.000] |
| $B = 75$ | 0.753 [0.680, 0.815] | 0.100 [0.040, 0.231] | 0.298 [0.222, 0.388] | 0.044 |
| $B = 100$ | **0.768 [0.695, 0.827]** | **0.095 [0.037, 0.221]** | **0.283 [0.208, 0.372]** | 0.037 |
| $B = 124$ | *0.761 [0.688, 0.822]* | *0.116 [0.051, 0.245]* | *0.286 [0.210, 0.375]* | *0.037* |
| $B = 150$ | *0.761 [0.688, 0.822]* | *0.116 [0.051, 0.245]* | *0.286 [0.210, 0.375]* | *0.037* |
| ★ *higher due to annotation noise—see Discussion* | | | | |

| SWMH (N=250) | | | | |
|---|---|---|---|---|
| **Condition** | **Acc.↑** | **FPR↓** | **SW FNR↓** | **Esc.↓** |
| Single-agent | 0.790 [0.735, 0.836] | 0.237 [0.184, 0.301] | 0.100 [0.044, 0.214] | 0.008 |
| MV $n = 1$ | 0.790 [0.735, 0.836] | 0.237 [0.184, 0.301] | 0.100 [0.044, 0.214] | 0.008 |
| *MV $n = 3$* | *0.795 [0.741, 0.841]* | *0.231 [0.178, 0.295]* | *0.100 [0.044, 0.214]* | *0.004* |
| *MV $n = 5$* | *0.791 [0.736, 0.837]* | *0.236 [0.183, 0.300]* | *0.100 [0.044, 0.214]* | *0.004* |
| $B = 10$ | — | — | — | 1.000 [0.985, 1.000] |
| $B = 50$ | — | — | — | 1.000 [0.985, 1.000] |
| $B = 75$ | *0.799 [0.745, 0.845]* | *0.227 [0.174, 0.291]* | *0.100 [0.044, 0.214]* | *0.024* |
| $B = 100$ | **0.798 [0.744, 0.844]** | **0.227 [0.174, 0.291]** | 0.100 [0.044, 0.214] | 0.008 |
| $B = 124$ | *0.794 [0.740, 0.840]* | *0.232 [0.179, 0.296]* | *0.100 [0.044, 0.214]* | *0.008* |
| $B = 150$ | *0.795 [0.741, 0.841]* | *0.231 [0.178, 0.295]* | *0.100 [0.044, 0.214]* | *0.004* |

**Table 2: Results on AEGIS 2.0 (top) and SWMH (bottom) with 95% Wilson CIs. Bold: best result. *Italics*: FPR CIs overlap with $B = 100$. $B = 10$ and $B = 50$ escalate all inputs and produce no classified outputs.**

fatigue and reduced vigilance toward genuinely high-risk cases [1]. All multi-agent conditions reduce FPR relative to single-agent, with $B = 100$ achieving the largest reduction.

*False negative rate.* FNR is flat across all valid conditions, ranging from 0.281 to 0.290 regardless of budget. The same inputs are missed consistently, suggesting the ceiling reflects annotation quality rather than sampling strategy—we examine these examples in the Discussion.

*Accuracy.* $B = 100$ reaches 0.768 (95% CI [0.695, 0.827]), compared to 0.752 for single-agent. Confidence intervals overlap substantially; accuracy alone does not differentiate conditions on this dataset.

*Budget sweep.* $B = 10$ and $B = 50$ escalate all inputs: with $K = 3$ active arms, a budget of 50 yields roughly 16 pulls per arm, too few for elimination to fire. $B = 75$ produces valid classifications (FPR = 0.100) but falls short of $B = 100$. $B = 124$ and $B = 150$ offer no further gain. The inflection between $B = 75$ and $B = 100$ provides empirical grounding for the sample complexity bound: inputs in this dataset require roughly 80–90 total pulls per arm for reliable elimination. Figure 2 shows the accuracy-versus-compute Pareto scatter; Figure 3 shows total token usage.

## Experiment 2: SWMH Generalization

Results on the stratified SWMH sample (N=250) are consistent with AEGIS 2.0. $B = 100$ again achieves the lowest FPR at 0.227, and all valid conditions achieve SW FNR = 0.100 (95% CI [0.044, 0.214])—the same 5 SuicideWatch posts missed consistently across every condition. The higher FPR on SWMH relative to AEGIS 2.0 reflects label noise in the safe subreddits: posts from depression, anxiety, and bipolar communities contain distressing content that the C-SSRS grounded prompt flags as potentially unsafe even when subreddit-level labels classify them as safe. Budget conditions converge at 77–78 average pulls on SWMH compared to 86–89 on AEGIS 2.0,

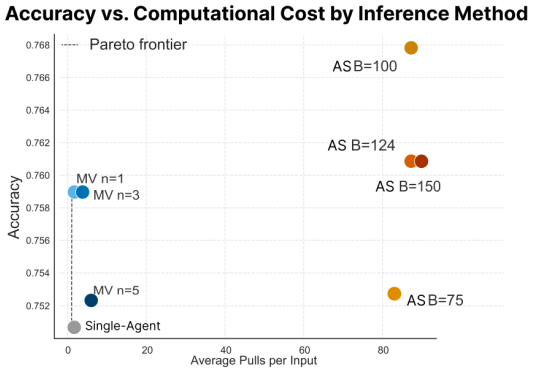


**Figure 2: Accuracy versus average pulls per input on AEGIS 2.0. MV $n = 1$ and MV $n = 3$ lie on the Pareto frontier. Adaptive sampling conditions achieve higher accuracy than single-agent at substantially greater compute cost—the cost of the reliability guarantee.**

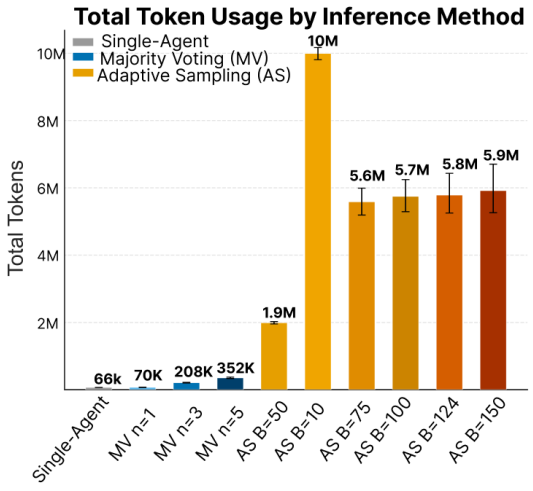


**Figure 3: Total token usage by condition on AEGIS 2.0.** $B = 50$ **consumes more tokens than any other condition, including** $B = 100$**, while producing no classified outputs—a consequence of nodes exhausting their full budget before escalating.**

suggesting SWMH posts are on average more clearly safe or unsafe under the C-SSRS prompt.

## Discussion

Adaptive sampling reduces false positives by 40% relative to single-agent on AEGIS 2.0 without affecting false negative rates, and the pattern holds on SWMH, suggesting the finding generalizes across source domains. In clinical terms, this reduction in unnecessary escalation directly addresses alert fatigue—a known problem in behavioral health decision support that can reduce clinician responsiveness to genuine risk [1].

The false negative rate does not improve with additional sampling on either dataset. Inspection of consistently misclassified examples on AEGIS 2.0 reveals annotation noise: several posts are appropriate chatbot refusals labeled unsafe, and others contain content outside C-SSRS criteria entirely. The same inputs are missed from single-agent through $B = 150$, confirming the ceiling is a property of the labels rather than the sampling strategy. The 5 missed SuicideWatch posts on SWMH tell a similar story: two explicitly deny suicidal intent despite expressing distress (correctly classified as safe under C-SSRS criteria), one is written by a caregiver rather than a person in crisis, and two express distress indirectly without meeting explicit C-SSRS thresholds. These cases point to a mismatch between criteria encoded in the prompt and the broader clinical signals a human clinician would recognize. Improving ground truth annotation in behavioral health benchmarks, ideally grounded in clinician-assigned C-SSRS labels, remains an open problem.

The annotation noise observed across both datasets points to a broader challenge in benchmark development for mental health AI. AEGIS 2.0 labels reflect LLM safety annotations rather than clinician-assigned C-SSRS severity ratings, and SWMH relies on subreddit membership as a proxy for clinical ground truth. Both introduce systematic label noise that sets a ceiling on false negative rates regardless of sampling strategy. Constructing behavioral health benchmarks with direct clinician annotation, grounded in validated instruments, is the most important methodological step for meaningful progress on the missed detection problem and a gap the community should prioritize.

The pipeline's current routing behavior has a practical implication worth noting for deployment. At $B = 50$, each node exhausts its budget without converging and passes the input downstream, resulting in up to 150 total pulls per input while producing no classified outputs and consuming more tokens than $B = 100$. Early termination on budget exhaustion, escalating directly rather than continuing downstream, would eliminate this cost and is a practical improvement for clinical deployment.

## Limitations

Adaptive sampling uses approximately 80–90 average pulls per input compared to 1–5 for majority vote, a 20–30× increase in API calls. For organizations with limited compute budgets, MV $n = 3$ offers comparable accuracy at substantially lower cost, and the FPR benefit must be weighed against operational cost in deployment planning. The pipeline currently evaluates three agent nodes; whether the theoretical guarantees and empirical gains hold with larger agent counts is an open question. SWMH relies on subreddit membership as a proxy for clinical ground truth rather than direct clinician annotation, which limits the interpretability of FNR on that dataset. Finally, all experiments use a single foundation model; performance under different model families or with task-specific clinical models may differ.

## Conclusion

Self-harm risk screening is a setting where AI system errors have direct consequences for patient safety. This work shows that replacing fixed-sample majority voting with adaptive sampling reduces false positives by 40% on a behavioral health benchmark, without affecting the rate of missed detections. When the system cannot reach a confident decision within budget, it escalates to human clinician review rather than guessing—a formal safety guarantee absent from existing heuristic approaches. The same pattern holds on a held-out Reddit dataset, suggesting the finding is not specific to one benchmark. The false negative rate does not improve with more sampling on either dataset, pointing to annotation quality in current behavioral health benchmarks as the limiting factor. Building clinician-annotated benchmarks grounded in validated instruments like C-SSRS is the most important next step for meaningful progress on the false negative problem. In safety-critical clinical settings, knowing when not to decide is as important as knowing how to decide.

## References

[1] Jessica S Ancker, Alison Edwards, Stephanie Nosal, David Hauser, Elizabeth Mauer, and Rainu Kaushal. 2017. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Medical Informatics and Decision Making* 17, 1 (2017), 1–9.

[2] Mohammad Ghavamzadeh Azar, Ian Osband, and Remi Munos. 2017. Minimax Regret Bounds for Reinforcement Learning. In *International Conference on Machine Learning*. 263–272.

[3] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. 1956. Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *The Annals of Mathematical Statistics* 27, 3 (1956), 642–669.

[4] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. 2006. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of Machine Learning Research* 7, 1079–1105.

[5] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Amit Sheth, et al. 2019. Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention. In *The World Wide Web Conference*. 514–525.

[6] Shaona Ghosh, Prasoon Varshney, Thomas Westfechtel, Ryota Uehara, and Romain Paulus. 2025. AEGIS2.0: A Diverse AI Safety Dataset and Risks Taxonomy for Alignment of LLM Guardrails. *arXiv preprint arXiv:2501.09004* (2025).

[7] Jia-Wen Guo, Julianne Kimmel, and Lauri A Linder. 2024. Text Analysis of Suicide Risk in Adolescents and Young Adults. *Journal of the American Psychiatric Nurses Association* 30, 1 (2024), 169–173. doi:10.1177/10783903221077292

[8] Xiaolei Huang, Lei Zhang, Tianli Liu, David Chiu, Tingshao Zhu, and Xin Li. 2014. Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons. *arXiv preprint arXiv:1411.0778* (2014).

[9] Thomas Jaksch, Ronald Ortner, and Peter Auer. 2010. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research* 11 (2010), 1563–1600.

[10] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. *arXiv preprint arXiv:2110.15621* (2021).

[11] Ananya Joshi and Michael Rudow. 2026. Constrained Process Maps for Multi-Agent Generative AI Workflows. arXiv:2602.02034 [cs.AI]

[12] Rumeng Li, Xun Wang, Dan Berlowitz, Jesse Mez, Honghuang Lin, and Hong Yu. 2025. CARE-AD: A Multi-Agent Large Language Model Framework for Alzheimer's Disease Prediction Using Longitudinal Clinical Notes. *npj Digital Medicine* 8, 1 (2025), 541. doi:10.1038/s41746-025-01940-4

[13] Pascal Massart. 1990. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability* 18, 3 (1990), 1269–1283.

[14] Adrianos Pavlopoulos, Theodoros Rachiotis, and Ilias Maglogiannis. 2024. An overview of tools and technologies for anxiety and depression management using AI. *Applied Sciences* 14, 19 (2024), 9068.

[15] Kelly Posner, Gregory K Brown, Barbara Stanley, David A Brent, Kseniya V Yershova, Maria A Oquendo, Glenn W Currier, Glenn A Melvin, Laurence Greenhill, Sa Shen, et al. 2011. The Columbia–Suicide Severity Rating Scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American Journal of Psychiatry* 168, 12 (2011), 1266–1277.

[16] Jie Sun, Tangsheng Lu, Xuexiao Shao, Ying Han, Yu Xia, Yongbo Zheng, Yongxiang Wang, Xinmin Li, Arun Ravindran, Lizhou Fan, et al. 2025. Practical AI application in psychiatry: historical review and future directions. *Molecular Psychiatry* 30, 9 (2025), 4399–4408.

[17] Samantha Weber, Barbora Provaznikova, Nikolas Psathakis, Andrea Casanova, Lucian Amthor, Lara Kirchhofer, José Antonio García Macías, Tobias Welt, Golo Kronenberg, Tobias Kowatsch, et al. 2026. Can Conversational AI Reliably Assess Depressive Symptoms? A Feasibility, Acceptability and Bias-Aware Validation Study of MADRS-Chat. *A Feasibility, Acceptability and Bias-Aware Validation Study of MADRS-Chat* (2026).

[18] Edwin B Wilson. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *J. Amer. Statist. Assoc.* 22, 158 (1927), 209–212.

[19] Yingxuan Yang et al. 2025. AgentNet: Decentralized Evolutionary Coordination for LLM-based Multi-Agent Systems. *arXiv preprint arXiv:2504.00587* (2025).

[20] Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. 24–33.

## Appendix A: DKW Inequality and Node-Level Bound

The Dvoretzky-Kiefer-Wolfowitz (DKW) inequality provides simultaneous error bounds across all categories of a categorical distribution without the $\ln K$ penalty that arises from applying Hoeffding's inequality to each category separately and taking a union bound.

*Theorem (Node-Level Categorical Bound).* Under i.i.d. sampling at node $s$, for any $\delta \in (0,1)$, with probability at least $1-\delta$ simultaneously over all $c \in \mathcal{A}$:

$$|\hat{p}_s(c\,;\,x) - p_s(c\,;\,x)| \le \epsilon_n(\delta) := \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

*Proof.* Impose a canonical ordering $\mathcal{A} = \{c_1, c_2, c_3\}$ and encode each sample as $Z_k = \phi(a^{(k)}) \in \{1,2,3\}$. Each category probability is a difference of CDF values: $p_s(c_j) = F(j) - F(j-1)$, where $F$ is the CDF of $Z$. Therefore:

$$|\hat{p}_s(c_j) - p_s(c_j)| \le |\hat{F}_n(j) - F(j)| + |\hat{F}_n(j-1) - F(j-1)| \le 2 \sup_t |\hat{F}_n(t) - F(t)|.$$

Applying the DKW inequality:

$$\mathbb{P}\left( \sup_t |\hat{F}_n(t) - F(t)| > \epsilon \right) \le 2e^{-2n\epsilon^2}.$$

Setting $2e^{-2n\epsilon^2} = \delta$ and solving gives $\epsilon = \sqrt{\ln(2/\delta)/(2n)} = \epsilon_n(\delta)$, which holds simultaneously over all $c \in \mathcal{A}$ with no $K$ factor. The union-bounded Hoeffding approach would require $n \ge \frac{\ln(2/\delta)+\ln K}{2\epsilon^2}$ samples; DKW requires only $\frac{\ln(2/\delta)}{2\epsilon^2}$, saving $\frac{\ln K}{2\epsilon^2}$ samples per node. For $K=3$, $\epsilon = 0.05$: approximately 220 fewer samples per node.

*Corollary (Correct Routing Guarantee).* Under the unique best action and i.i.d. sampling assumptions, if $n \ge n^*(\delta, \Delta_s(x)) := \frac{2\ln(2/\delta)}{\Delta_s(x)^2}$, then with probability $\ge 1-\delta$, the empirical argmax equals the true best action: $\hat{c}^* = c^*(s,x)$.

*Proof.* By the node-level bound, with probability $\ge 1-\delta$, $|\hat{p}_s(c) - p_s(c)| \le \epsilon_n(\delta)$ for all $c$. For any $c \ne c^*$:

$$\hat{p}_s(c^*) - \hat{p}_s(c) \ge [p_s(c^*) - \epsilon_n] - [p_s(c) + \epsilon_n] = \Delta_s(x) - 2\epsilon_n(\delta).$$

This is strictly positive when $\epsilon_n(\delta) < \Delta_s(x)/2$, i.e. when $n \ge 2\ln(2/\delta)/\Delta_s(x)^2 = n^*$.

*Remark.* The minimum sample count $n^* = 2\ln(2/\delta)/\Delta_s(x)^2$ depends on input $x$ through the probability gap $\Delta_s(x)$. Easy inputs with large gaps need few samples; ambiguous inputs with small gaps require many. Fixed majority voting ignores this dependence entirely, applying the same $n$ regardless of input difficulty.

## Appendix B: Good Event Proof

*Lemma (Good Event).* Define the good event as the event that all confidence intervals contain the true probability at every round simultaneously:

$$\mathcal{G} = \left\{ \forall c \in \mathcal{A},\ \forall m \ge 1 : |\hat{p}_s^{(m)}(c) - p_s(c)| \le \sqrt{\frac{\ln(4KT_c^2/\delta)}{2m}} \right\}.$$

Then $\mathbb{P}(\mathcal{G}) \ge 1-\delta$.

*Proof.* Fix one arm $c$ and one round where it has been pulled $m$ times. By Hoeffding's inequality:

$$\mathbb{P}\left( |\hat{p}_s^{(m)}(c) - p_s(c)| > \sqrt{\frac{\ln(4KT_c^2/\delta)}{2m}} \right) \le 2\exp\left( -\ln\left(\frac{4KT_c^2}{\delta}\right)\right) = \frac{\delta}{2KT_c^2}.$$

Taking a union bound over $K$ arms and summing over all possible rounds $m \in \{1,2,\ldots\}$ using the standard series bound $\sum_{m=1}^{\infty} \frac{1}{m^2} = \frac{\pi^2}{6} < 2$:

$$\mathbb{P}(\mathcal{G}^c) \le \sum_{c=1}^{K} \sum_{m=1}^{\infty} \frac{\delta}{2KT_c^2} \le K \cdot \frac{\delta}{2K} \cdot \frac{\pi^2}{6} < \delta.$$

Therefore $\mathbb{P}(\mathcal{G}) \ge 1-\delta$.

*Proposition (Correctness of Algorithm 1).* Under the unique best action assumption, with probability $\ge 1-\delta$, Algorithm 1 never eliminates $c^*$. It either returns $c^*$ or returns `escalate`.

*Proof.* Under $\mathcal{G}$, arm $c^*$ is eliminated only if $\hat{p}_s(c) - w_c > \hat{p}_s(c^*) + w_{c^*}$ for some $c \ne c^*$. Under $\mathcal{G}$, this would require $p_s(c) > p_s(c^*)$, contradicting the assumption that $c^*$ is the unique best arm. Therefore $c^*$ is never eliminated. If the budget is exhausted before a single arm survives, the algorithm returns `escalate` rather than forcing a potentially wrong label.

## Appendix C: MDP Formulation and Regret Proofs

*MDP Formulation.* We model the evaluation workflow as a bounded-horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \tau_{\max})$ structured by the DAG $G$. The state space $\mathcal{S}$ corresponds to the three agent nodes plus terminal states $\mathcal{S}_{\text{term}} = \{\texttt{safe}, \texttt{unsafe}, \texttt{human-review}\}$. Transitions $P(s' \mid s, a)$ are restricted to DAG edges: terminal labels route to the corresponding terminal state, while `escalate` routes to the next agent. The reward function $R(s,a)$ captures evaluation quality at each step with $|R(s,a)| \le R_{\max}$, and episodes have maximum length $\tau_{\max} \le 3$. Cumulative regret over $T$ episodes is:

$$\text{Reg}(T) = \sum_{t=1}^{T} \left[ V^{\pi^*}(s_0, x_t) - V^{\pi}(s_0, x_t) \right],$$

where $\pi^*$ is the oracle policy knowing all true label distributions.

*Theorem (Majority-Vote Regret).* Under fixed majority voting with $n$ samples per node:

$$\text{Reg}_{\text{MV}}(T) \le 2\tau_{\max} R_{\max} |\mathcal{S}| \cdot T \cdot e^{-n\Delta_{\min}^2/2}.$$

This grows as $\Theta(T)$: the error probability is constant across all episodes because the policy never adapts.

*Proof.* By the node-level bound with $\epsilon = \Delta_s(x)/2$, the routing error probability at node $s$ is at most $2e^{-n\Delta_{\min}^2/2}$. Each error contributes at most $\tau_{\max} R_{\max}$ to regret. A union bound over $|\mathcal{S}|$ nodes and summing over $T$ episodes gives the stated bound. Achieving sublinear regret requires $n = \Omega(\ln T / \Delta_{\min}^2)$, which requires knowing $\Delta_{\min}$ in advance.

*Theorem (Adaptive Regret).* The adaptive policy $\pi_{\text{UCB}}$ using Algorithm 1 achieves:

$$\text{Reg}_{\text{UCB}}(T) \le C \cdot \frac{\tau_{\max} R_{\max} \cdot K|\mathcal{S}| \ln T}{\Delta_{\min}^2},$$

for a universal constant $C > 0$, without knowledge of $\Delta_{\min}$. This grows as $O(\log T)$: doubling the number of patient interactions increases total mistakes by only $\ln 2 \approx 0.69$.

*Proof.* Set $\delta = 1/T$. By the sample complexity result, each node identifies $c^*$ after $O(K \ln(KT)/\Delta_{\min}^2)$ episodes (learning phase). Each learning-phase episode contributes at most $\tau_{\max} R_{\max}$ to regret; exploitation-phase episodes contribute $O(R_{\max}/T)$ each. Summing over $|\mathcal{S}|$ nodes gives the $O(\log T)$ bound.

*Theorem (System-Level Regret).* Under the adaptive policy:

$$\mathrm{Reg}_{\mathrm{UCB,sys}}(T) \leq C \cdot \frac{K \cdot |\mathcal{S}| \cdot \tau_{\max} \cdot R_{\max} \cdot \ln T}{\Delta_{\min}^2}.$$

The ratio of adaptive to fixed-budget regret satisfies:

$$\frac{\mathrm{Reg}_{\mathrm{UCB}}(T)}{\mathrm{Reg}_{\mathrm{MV}}(T)} = O\left(\frac{\ln T}{T \cdot e^{-n\Delta_{\min}^2/2}}\right) \xrightarrow{T\to\infty} 0.$$

The adaptive policy is asymptotically dominant.

*Proof.* At most $\tau_{\max}$ nodes are visited per episode, and $|\mathcal{S}|$ nodes contribute independently to learning-phase regret, giving the $K|\mathcal{S}|\tau_{\max}$ factor. The ratio follows from $\mathrm{Reg}_{\mathrm{UCB}}(T) = O(K|\mathcal{S}|\tau_{\max} \ln T/\Delta_{\min}^2)$ and $\mathrm{Reg}_{\mathrm{MV}}(T) = \Omega(\tau_{\max} T e^{-n\Delta_{\min}^2/2})$.

## Appendix D: Reproducibility

*Model configuration.* All experiments use `claude-sonnet-4-6` at temperature 0.7 with a maximum of 10 output tokens per call. The 10-token limit is sufficient for the label-only output format and reduces cost substantially relative to unconstrained generation.

*System prompt structure.* Each node receives a role-specific system prompt that defines its clinical function (Worker: frontline screening; Risk: secondary clinical review; Legal: institutional compliance) and instructs the model to respond with exactly one of `safe`, `unsafe`, or `escalate`, grounded in C-SSRS criteria. The prompt includes definitions of each label consistent with the action space defined in the Methods.

*Dataset preprocessing.* AEGIS 2.0 examples are filtered to rows where `violated_categories` contains suicide or self-harm keywords across all three splits (train, validation, test), deduplicated by example ID, yielding N=163 unique examples. Two examples are excluded due to Unicode encoding errors. SWMH examples are sampled with `random_state=42` using stratified sampling of 50 examples per subreddit.

*Confidence intervals.* All proportions are reported with 95% Wilson score confidence intervals, computed using the standard formula:

$$\frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}},$$

where $z = 1.96$ for 95% confidence and $n$ is the number of non-escalated examples for FPR and FNR, or total examples for escalation rate.

*Code.* Code and data preprocessing scripts can be found on: https://github.com/meghanakaranam1/AI4Mental_workshop