MoSE-Counseling: Routed Safety Experts for Chinese-Language Large Language Model Counseling Benchmarks

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) used for counseling can miss crisis signals or produce boundary-crossing responses. We study **MoSE-Counseling**, a two-stage architecture for *Chinese-language counseling benchmarks*. Stage 1 extracts an evidence-grounded client profile; Stage 2 uses the profile and an independently computed current-turn screening score to route among LoRA experts for risk assessment, ethical guidance, response generation, and final coordination. We present the routing rule as a conservative engineering heuristic, not a theoretical safety guarantee: it cannot repair a screening false negative, and the current evidence does not establish clinical or cross-lingual validity. On held-out, in-family splits of SOS-HL-1K, PsyQA, and D^4, the system improves the reported automatic and professional-rating metrics over the evaluated baselines. We additionally make the evaluation semantics explicit: SOS-HL-1K measures the standalone screening head, whereas routing thresholds govern system-level expert activation and response behavior. An analytic persistence audit shows that the original symmetric exponential moving average retains only 0.7^N of a crisis signal after N calm turns; we therefore use asymmetric retention for the risk-factor state. The results remain a Chinese, benchmark-level study and do not support autonomous crisis triage.

1 Introduction

Global mental health needs increasingly outstrip the capacity of traditional service models Kazdin & Rabbitt (2013), motivating research on scalable forms of decision support. LLM-based counseling systems are one possible way to widen access, but they must be evaluated as high-risk support tools rather than assumed to be clinically reliable. In particular, empirical studies show that LLMs carry serious safety risks in crisis scenarios, performing poorly at identifying suicidal ideation and responding appropriately Li et al. (2025).

Recent work has moved suicide-ideation identification from simple text classification toward decision-support pipelines Cui et al. (2025). GPT-4 Achiam et al. (2023) has reached triage agreement broadly comparable to clinicians in a controlled study with markedly higher throughput Lee et al. (2024), and deep models can separate depression from suicidality Haque et al. (2021). Yet these results are obtained on curated benchmarks and do not transfer cleanly to deployment: models exhibit inter-model personality biases that hurt diagnostic standardization Wang et al. (2025) and can miss implicit ideation or hidden risk factors Li et al. (2025); Lissak et al. (2024). Earlier social-media research also shows that transitions toward suicidal ideation may be expressed through subtle longitudinal shifts rather than explicit crisis phrases De Choudhury et al. (2016). The gap between benchmark feasibility and clinical utility reflects two missing capabilities: *individualized understanding* of the client, and *specialized, auditable safety reasoning* inside the counseling pipeline rather than as an external filter.

In this work we study **MoSE-Counseling**, a Mixture of Safety Experts architecture with two stages. Stage 1 (*Client Information-Profile Generation*) extracts an evidence-grounded structured profile.

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

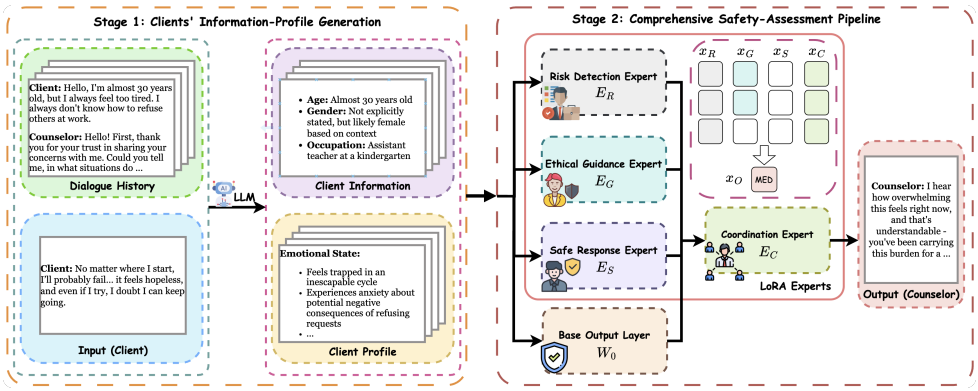

Figure 1: Overview of MoSE-Counseling. Stage 1 builds an evidence-grounded client information profile; Stage 2 routes the dialogue through specialized safety experts (E_R, E_G, E_S) and a coordination expert (E_C) on top of the frozen base model W_0. The coordination expert is a fixed final stage and is therefore not part of the selectable activation set.

Stage 2 (*Comprehensive Safety-Assessment Pipeline*) decomposes counseling into specialized LoRA experts coordinated by a profile-conditioned router. Crucially, Stage 2 depends on Stage 1 outputs, which are produced by an LLM and can be noisy or hallucinated. Rather than assuming the profile is correct, we make the router conservative: an imperfect profile may cause over-activation, while a fired profile-independent screen forces activation of the risk expert. This is a branch invariant by construction, not a learned or clinical guarantee. The screen can still miss a genuine crisis. Our evidence is limited to three Chinese datasets, two of which provide training data, and we make no claim of cross-lingual, cross-platform, or out-of-distribution generalization.

We summarize our contributions as follows:

- **A counseling-safety MoE architecture with conservative routing.** Unlike token-routed mixture-of-LoRA architectures and input/output guardrails, our experts are specialized to counseling-safety sub-tasks (risk, ethics, safe response, coordination), and routing is driven by a clinical risk profile with a profile-independent screening floor. This floor prevents profile corruption alone from overriding an already-triggered crisis screen; it does not eliminate screening false negatives.

- **A precise separation of detector and system evaluation.** We report standalone screening-head predictions independently of profiles and routing thresholds, and reserve system-level measures for response and activation behavior.

- **Chinese benchmark evidence with explicit limits.** On SOS-HL-1K, PsyQA, and D^4, MoSE-Counseling improves safety- and professionalism-related metrics over the evaluated fine-tuned and domain baselines. Paired-bootstrap testing is applied to the headline automatic metrics (Experiments). We explicitly do *not* claim clinical equivalence or distributional transfer.

- **A persistence audit and asymmetric risk memory.** We quantify the rapid attenuation induced by symmetric EMA updates and retain elevated risk asymmetrically after a crisis disclosure, pending explicit review or a validated release condition.

- **Controlled comparisons.** Beyond the profile schema, scoring anchors, prompt template, and routing details, we report controlled baselines isolate the MoSE design from prompting and fine-tuning (profile-augmented prompting, safety-instruction prompting, single-LoRA, multi-task LoRA, MoE-LoRA, and an external guardrail).

2 Related Work

2.1 LLMs for Mental Health Services

Mental health services face a structural supply–demand imbalance that conventional models cannot close Kazdin & Rabbitt (2013), motivating LLM-based support and a shift from scripted chatbots toward continuous clinical reasoning Na et al. (2025). Domain models such as EmoLLM EmoLLM (2024) and CPsyCounX Zhang et al. (2024) target counseling, while general models such as GPT-4 Achiam et al. (2023) achieve triage agreement comparable to clinicians at higher throughput Lee et al. (2024), enabling risk monitoring and training support Haque et al. (2021). These systems optimize fluency and helpfulness but treat safety as an emergent property rather than an explicit, auditable component. Unlike these approaches, MoSE-Counseling introduces a dedicated coordination expert that explicitly resolves competing objectives, such as empathy versus safety refusal, while maintaining therapeutic coherence.

2.2 Crisis Identification and Safety in LLM Counseling

Empirical studies reveal fundamental safety risks in crisis scenarios Li et al. (2025). LLMs can miss implicit ideation, while broader computational work has identified subtle risk factors that are easy to overlook in text-only screening Lissak et al. (2024). External guardrails such as Llama Guard Inan et al. (2023) attach separate classifiers to filter inputs and outputs, but operate *outside* the generation process and do not by themselves guarantee a therapeutically appropriate response. Inter-model biases further undermine standardization Wang et al. (2025). This motivates safety mechanisms that operate *within* the counseling pipeline and remain inspectable. To address this, MoSE-Counseling uses an internal conservative routing mechanism driven by clinical risk profiles and an independent screening branch.

2.3 Parameter-Efficient Expert Mixing and Multi-Agent Safety

LoRA Hu et al. (2022) is the standard for adapting LLMs under limited resources, but a single adapter struggles to jointly satisfy competing objectives (e.g., empathy vs. safety compliance). Mixture-of-LoRA architectures such as MoLE Wu et al. (2024) route *tokens* to specialized adapters, decoupling sub-skills. Modular safety pipelines instead invoke separate components for detection, generation, and verification. In contrast to generic MoE-LoRA methods that sum token-level hidden states, our architecture assigns experts to explicit counseling-safety functions and structurally coordinates their structured outputs.

3 Methodology

3.1 Overview

As shown in Fig. 1, MoSE-Counseling closes safety and professionalism gaps through two stages. **Stage 1** constructs a structured client profile from dialogue context; **Stage 2** routes the dialogue through specialized safety experts. We write the pipeline as MoSE-Counseling $= f(\text{Stage1}, \text{Stage2})$ and detail each stage below, emphasizing the schema, scoring, and safeguards needed for reproducibility.

3.2 Problem Formulation and Design Principle

At each turn, Stage 1 returns a profile $(\mathcal{I}, \mathcal{P})$ with confidence $c \in [0, 1]$, and Stage 2 chooses an activation set

$$\mathcal{A}_0 = \{S\} \subsetneq \mathcal{A}_1 = \{R, S\} \subsetneq \mathcal{A}_2 = \{R, G, S\}, \tag{1}$$

ordered by inclusion. These sets count full expert passes; the lightweight screening head and final coordination pass always run. The routing rule encodes an engineering preference: a larger set costs more computation and can over-intervene, while an unjustified smaller set can miss clinically salient context. Because both the profile and screen are fallible learned estimates, this preference is not a safety theorem. The only deterministic property is the branch condition $s_R > \tau_{\text{floor}} \Rightarrow R \in \mathcal{A}$; it says nothing about cases that the screen misses.

3.3 Stage 1: Client Information-Profile Generation

Given a user input $X = \{x_1, \ldots, x_n\}$ and dialogue history $H = \{h_1, \ldots, h_m\}$, Stage 1 produces client information $\mathcal{I}$ and a psychological profile $\mathcal{P}$.

3.3.1 Profile Schema and Scoring.

The profile has two parts. *Demographic information* $\mathcal{I}$ contains the fields `age`, `gender`, `occupation`, `education`, `living_situation`, `relationship_status`, `recent_life_events`, and `symptom_duration`. *Psychological profile* $\mathcal{P}$ contains four dimensions, each rated on a 0–10 severity scale: (1) *emotional state* (affect, mood stability); (2) *cognitive pattern* (distortions, rumination); (3) *behavioral indicator* (withdrawal, substance use, sleep disruption); and (4) *risk factor* (suicidal ideation, self-harm history, hopelessness). Scoring anchors are fixed: 0–2 none/minimal, 3–5 mild, 6–8 moderate, 9–10 severe; for *risk factor*, 9–10 denotes active ideation with stated plan or intent. Standardized anchors make severity scores comparable across clients and across raters during validation.

3.3.2 Constrained Extraction Prompt.

We extract the profile with a single structured prompt designed to suppress hallucination (the complete template is provided in Appendix A). Three constraints are central: (i) every non-`unknown` field must carry a verbatim `evidence` span copied from the dialogue; (ii) unsupported fields must be returned as `unknown` rather than guessed; and (iii) each field carries a `confidence` in $[0, 1]$. Fields without grounding evidence or with confidence below $c_{\min}$ are treated as missing downstream, so that an unsupported guess about, e.g., age or occupation does not propagate into routing. The profile-level confidence c used by the router is the arithmetic mean over the retained psychological fields; if none is retained, we set $c = 0$.

At each step the dual-stream extraction is

$$\mathcal{I}_{t+1}, \mathcal{P}_{t+1} = f(\mathbf{h}_t, \mathbf{x}_t), \tag{2}$$

combining rule-based matching for explicit markers (e.g., "I am 25") with contextual inference for implicit content.

3.3.3 Profile Reliability and Validation.

Because Stage 2 conditions on $\mathcal{P}$, we explicitly validate Stage 1 rather than assume it is correct. The 2,000-sample PsyQA subset is independently annotated by three annotators, with a fourth adjudicating disagreements. Before adjudication, categorical profile fields reach Krippendorff's $\alpha = 0.78$, ordinal severity scores reach ICC(2,k) $= 0.82$, and the adjudication frequency is 18.4%. Against the adjudicated reference, demographic extraction reaches 0.86 macro-F1 and risk severity is within ± 1 point in 82% of cases. The agreement and adjudication values are reported from the retained annotation matrix; Appendix E specifies the calculation.

3.3.4 Asymmetric Profile Update.

For non-risk dimensions d, we retain the original exponential update

$$\mathcal{P}_{t+1}^{(d)} = \alpha_{\text{decay}}\, \mathcal{P}_t^{(d)} + (1 - \alpha_{\text{decay}})\, \mathcal{P}_{\text{new}}^{(d)}, \tag{3}$$

with $\alpha_{\text{decay}} = 0.7$. A symmetric update is inappropriate for the risk-factor dimension: after one crisis disclosure followed by calm values, its retained contribution is only 0.7^N after N turns (50% is crossed after 1.94 turns and 10% after 6.46 turns). We therefore use a session-level latch,

$$\mathcal{P}_{t+1}^{(\text{risk})} = \max\left\{ \mathcal{P}_t^{(\text{risk})}, \mathcal{P}_{\text{new}}^{(\text{risk})} \right\}, \tag{4}$$

which is released only by an explicit human review or a new session under a deployment-specific policy. This choice favors vigilance but may increase over-intervention; Appendix D reports the analytic comparison.

3.3.5 COLD START.

On the first turn ($H = \emptyset$) each dimension is initialized to a population-level prior $\mathcal{P}_0^{(d)} = \mu_{\text{prior}}^{(d)}$ (the training mean of dimension d), and the router defaults to a conservative setting (lowered thresholds $\theta'_{\text{high}} = 0.6$, $\theta'_{\text{med}} = 0.3$) that activates more experts. As dialogue accumulates, thresholds revert to defaults and the recursive update replaces priors with client-specific observations.

3.4 STAGE 2: COMPREHENSIVE SAFETY-ASSESSMENT PIPELINE

Stage 2 is a routed, multi-pass system rather than a sum of heterogeneous expert outputs. Let W_0 denote the frozen base-model parameters. Expert i has a LoRA update $\Delta W_i = B_i A_i$ and is invoked as

$$f_i(Z) = f_{W_0 + \Delta W_i}(Z), \qquad i \in \{R, G, S, C\}. \tag{5}$$

The selectable experts are Risk Detection (R), Ethical Guidance (G), and Safe Response (S). The Coordination Expert (C) is a fixed final stage and is not counted in $\mathcal{A}$. Each invoked expert returns a structured text or score; the system never adds the hidden states of a classifier, scorer, and generator.

3.4.1 PROFILE-CONDITIONED CONSERVATIVE ROUTER.

The router computes a risk score from client information, profile, and input:

$$\rho(\mathcal{I}, \mathcal{P}, X) = \sigma\big(\mathbf{w}_\rho^\top [\phi(\mathcal{I}); \phi(\mathcal{P}); \phi(X)]\big)\,, \tag{6}$$

where $\phi(\cdot)$ concatenates mean-pooled base-model embeddings with the severity vector of $\mathcal{P}$, and $\mathbf{w}_\rho$ is a lightweight linear projection trained on SOS-HL-1K risk labels. In addition to the LoRA parameters, the trainable non-backbone parameters are this projection and the scalar risk and ethical-scoring heads described below. The nominal activation is

$$\mathcal{A} = \begin{cases} \{R, G, S\} & \rho \geq \theta_{\text{high}} \text{ (high-risk)} \\ \{R, S\} & \theta_{\text{med}} \leq \rho < \theta_{\text{high}} \text{ (medium-risk)} \\ \{S\} & \rho < \theta_{\text{med}} \text{ (routine)} \end{cases} \tag{7}$$

with $\theta_{\text{high}} = 0.7$, $\theta_{\text{med}} = 0.4$ selected on validation.

Because ρ depends on a possibly hallucinated profile, the router uses three conservative branch rules:

(F1) The Risk Detection Expert E_R is *always* evaluated in at least a lightweight screening mode, regardless of ρ; the routine set $\{S\}$ is augmented to $\{R, S\}$ whenever the screening score exceeds a low floor τ_{floor}.

(F2) The router may *narrow* the expert set only when ρ is confidently low *and* the mean profile confidence exceeds $c_{\min}$; otherwise it defaults to the medium-risk set $\{R, S\}$.

(F3) Under cold start, the conservative thresholds $(\theta'_{\text{med}}, \theta'_{\text{high}}) = (0.3, 0.6)$ replace the defaults.

Consequently, a profile error may change the nominal router score, but it cannot deactivate R after the independent screen has fired. This is true by construction and is not presented as a theoretical result. A false negative from the screen remains a possible silent downgrade and is the dominant unresolved safety failure.

3.4.2 SPECIALIZED EXPERTS.

Risk Detection (E_R). A scalar head h_R over the final hidden state of f_R produces P_R over routine and high-risk labels. The always-on screening score is $s_R = P_R(\text{high} \mid X)$ using only the current input and dialogue history; when $R \in \mathcal{A}$, the same adapter also generates a structured risk summary for the coordinator. **Ethical Guidance (E_G).** Given the draft, profile, and dialogue, f_G produces an ethical critique, while a scalar head h_G returns $S_G \in [0, 1]$ for non-maleficence, therapeutic boundaries, and informed-consent compliance. The scalar ethical head is used as a lightweight final check on every route; activating G additionally produces the full textual critique. **Safe Response (E_S).** The generator produces a single draft $R_{\text{draft}} = f_S(X, \mathcal{P}, C_{\text{cons}})$, where C_{cons} is the fixed counseling-safety instruction.

Algorithm 1 Conservative routing and safe response generation

Input: Profile $\mathcal{I}, \mathcal{P}$ with confidence c, input X, experts $\{E_R, E_G, E_S, E_C\}$
Output: Safe response R_{final}

$\rho \leftarrow \sigma(\mathbf{w}_\rho^\top [\phi(\mathcal{I}); \phi(\mathcal{P}); \phi(X)])$
$s_R \leftarrow$ screening score from $E_R(X)$ {safety floor F1}
$(\theta_m, \theta_h) \leftarrow (0.3, 0.6)$ if cold start, else $(0.4, 0.7)$
if $\rho \geq \theta_h$ **then**
    $\mathcal{A} \leftarrow \{R, G, S\}$
else if $\rho \geq \theta_m$ **or** $c < c_{\min}$ **or** $s_R > \tau_{\text{floor}}$ **then**
    $\mathcal{A} \leftarrow \{R, S\}$ {narrowing blocked by F1/F2}
else
    $\mathcal{A} \leftarrow \{S\}$
end if
$Q_R, Q_G \leftarrow \emptyset$
$Q_R \leftarrow f_R(\cdot)$ if $R \in \mathcal{A}$; $R_{\text{draft}} \leftarrow f_S(\cdot)$
$Q_G \leftarrow f_G(R_{\text{draft}}, \cdot)$ if $G \in \mathcal{A}$
$R_{\text{final}} \leftarrow f_C(X, \mathcal{P}, R_{\text{draft}}, Q_R, Q_G)$
if $S_G(R_{\text{final}}) < \tau_{\text{safety}}$ **then**
    revise once with E_C; use conservative fallback if still below threshold
end if
if $s_R > \tau_{\text{floor}}$ **or** $(\rho \geq \theta_h$ **and** $P_R(\text{high}) > 0.8)$ **then**
    append crisis-escalation (hotline/referral) to R_{final}
end if
return R_{final}

3.4.3 Coordination Expert.

A coordination expert E_C restores global context lost to specialization. Its input concatenates the dialogue, profile, draft, risk summary (when active), and ethical critique (when active), each with a typed delimiter. It produces $R_{\text{final}} = f_C(X, \mathcal{P}, R_{\text{draft}}, Q_R, Q_G)$. The ethical head scores the final response; if $S_G < \tau_{\text{safety}} = 0.7$, the coordinator performs one constrained revision, after which a fixed conservative fallback is used if the score remains below threshold. We call this ordered generate–critique–coordinate procedure *hierarchical consensus*; it is distinct from averaging adapter weights.

4 Experiments

4.1 Datasets and Splits

SOS-HL-1K Qi et al. (2025) is a binary suicide-risk classification set of 1,249 real comments. **PsyQA** Sun et al. (2021) contains 22,000 mental-health questions with 56,000 counselor answers. D^4 Yao et al. (2022) provides 1,339 Chinese multi-turn depression-diagnosis dialogues (avg. 21.6 turns). For out-of-distribution evaluation, we additionally use **Unseen-Crisis**, a 300-case crisis/counseling test set that is not used to train, tune, or select any component.

To avoid train/test leakage—each expert is trained and the system evaluated on the same dataset family—we use *disjoint* stratified splits. SOS-HL-1K is split 7:1:2 into 874/125/250 (train/val/test) stratified by label; D^4 is split 7:1:2 into 937/134/268 by dialogue. For PsyQA, we draw **2,000** samples from the 22k pool by topic-stratified random sampling and have three trained annotators (two clinical-psychology graduate students and one licensed counselor) label demographic and risk fields and rate response quality against a written guideline, with a fourth adjudicating disagreements; these 2,000 samples are then split 7:1:2 into 1,400/200/400. The Risk Detection Expert is trained on the SOS-HL-1K train split; the Ethical Guidance and Safe Response Experts on the PsyQA train split; the Coordination Expert on the D^4 train split. **All reported results are on the held-out test splits, and no test data is used for training, threshold selection, or any hyperparameter tuning.** Routing thresholds, α_{decay}, $c_{\min}$, τ_{floor}, and τ_{safety} are tuned on the validation splits only. SOS-HL-1K and PsyQA also supply training data, whereas Unseen-Crisis is held out at the dataset-family level. The social-media origin of SOS-HL-1K differs from the intended counseling-dialogue context, making

Type	Models	SOS-HL-1K			PsyQA							D^4						
		Automatic			Automatic				Professional			Automatic				Professional		
		Pre.	Rec.	F_S^{all}	R-1	R-L	B-4	F_B	Non.	Bnd.	Rel.	R-1	R-L	B-4	F_B	Non.	Bnd.	Rel.
Closed-source	GPT-4o	74.25	74.70	74.47	23.93	14.89	3.65	62.83	8.31	8.52	8.79	19.85	18.31	3.05	60.51	8.06	7.95	8.11
	Deepseek	73.00	72.55	72.77	23.18	14.27	3.50	62.03	8.08	7.98	8.21	19.34	17.93	2.95	60.09	7.81	7.81	7.95
Open-source	Llama3.1-8B	69.50	69.00	69.25	22.14	13.75	3.30	61.22	7.03	7.17	7.24	18.46	17.21	2.80	59.25	7.05	7.10	7.49
	Qwen3-4B	70.25	69.75	70.00	22.65	14.02	3.45	61.84	7.83	7.98	8.06	18.87	17.50	2.85	59.78	7.37	7.25	7.55
Domain	EmoLLM	56.00	55.10	55.55	20.89	12.83	3.10	60.08	7.11	7.63	7.84	17.68	16.55	2.60	58.54	7.02	7.12	7.35
	CPsyCounX	57.55	56.30	56.92	21.21	13.04	3.15	60.79	7.53	7.48	7.85	17.90	16.76	2.65	58.90	7.22	7.29	7.38
	LLaMA-8B-FT	75.10	75.00	75.05	23.70	14.80	3.62	62.70	8.40	8.60	8.85	19.60	18.20	3.02	60.40	8.02	7.90	8.05
	Qwen3-4B-FT	76.80	76.00	76.40	23.34	14.53	3.55	62.31	8.60	8.75	8.90	19.17	18.04	2.98	60.13	8.05	7.98	8.12
Ours	LLaMA-MoSE	77.10	77.00	77.05†	24.10	14.90	3.72	63.60	8.65	8.80	9.00	20.00	18.55	3.10	60.95	8.15	8.05	8.25
	Qwen-MoSE	**78.00**	**78.05**	**78.02***	**24.46**	**15.03**	**3.77**	**64.11***	**8.76**	**8.93**	**9.10**	**20.33**	**18.81**	**3.16**	**61.37***	**8.24**	**8.19**	**8.37**

Table 1: Main results. **Bold**: best overall (per column). Underline: best *baseline* (per column), to show the margin over the strongest competitor. *: improvement of the best MoSE model over the best baseline is significant (paired bootstrap, 10,000 resamples, $p < 0.05$); †: no significant difference between the two MoSE variants ($p \geq 0.05$). For SOS-HL-1K, Pre. and Rec. are macro-averaged over the two labels, and F_S^{all} is their harmonic mean.

transfer to counseling data particularly important. The added test addresses dataset-family transfer in Chinese; it does not by itself establish cross-lingual or cross-platform generalization.

4.2 Implementation Details

We apply LoRA Hu et al. (2022) with rank 16 to different layers of Qwen3-4B Yang et al. (2025) (and Llama3.1-8B for the LLaMA-MoSE variant), creating the four experts as above. E_R and the router minimize binary cross-entropy on SOS-HL-1K. E_S minimizes autoregressive negative log-likelihood on PsyQA counselor responses. E_G regresses the normalized mean of the three annotated response-quality dimensions and is additionally prompted to emit the corresponding structured critique. For each D^4 training turn, the frozen upstream experts produce the typed inputs to E_C, which is trained by teacher forcing against the counselor turn. The adapters are trained separately; we do not jointly back-propagate through the discrete router. Training uses a cosine schedule, initial learning rate 5.0×10^{-4}, warmup ratio 0.1, batch size 1 per device, 5 epochs, on a single NVIDIA A100-80G GPU. At inference all models use temperature 0.7, top-p 0.8, top-k 20.

4.3 Baselines

We compare against **closed-source** GPT-4o Achiam et al. (2023) and Deepseek-r1 Guo et al. (2025); **open-source** Llama3.1-8B-Instruct Grattafiori et al. (2024) and Qwen3-4B Yang et al. (2025); and **domain** models EmoLLM-V3 EmoLLM (2024), CPsyCounX Zhang et al. (2024), and single-LoRA fine-tuned LLaMA-8B-FT and Qwen3-4B-FT (same data as our experts). **Fairness of comparison.** All generative systems receive the same dialogue context and decoding settings. The controlled study adds the safety instruction and Stage 1 profile cumulatively, then holds both fixed when comparing adaptation strategies. It also adds multi-task LoRA, token-routed MoE-LoRA (MoLE Wu et al. (2024)), and an external guardrail (Llama Guard Inan et al. (2023)) on a shared Qwen3-4B backbone.

4.4 Automatic Evaluation

For every MoSE SOS-HL-1K entry, the binary label is produced by the standalone screening head $s_R = P_R(\text{high} \mid X, H)$ at its validation-selected τ_{floor}; comparison models produce a direct binary-classification output under the common evaluation prompt. The MoSE head receives the current input and dialogue history, but not the Stage 1 profile, router score ρ, or routing thresholds $(\theta_{\text{med}}, \theta_{\text{high}})$. Consequently, changing only a profile or routing threshold cannot change a label in this table. System-level activation and response outcomes are evaluated separately. We use two SOS-HL-1K summaries for different purposes. Table 1 reports macro-averaged precision and recall over the two labels and their harmonic mean, denoted F_S^{all}. The controlled and component

analyses also report F_S^{risk}, the standard positive-class F1 with *high risk* as the positive label. It is computed from all examples in the binary test split, not from a positives-only subset. Thus F_S^{all} measures overall classification quality, whereas F_S^{risk} isolates crisis detection; the two should not be read as duplicate estimates of the same quantity. For PsyQA and D^4 we report ROUGE Lin (2004), BLEU-4 Papineni et al. (2002), and BERTScore-F (F_B). Chinese ROUGE and BLEU use a common character-level tokenization for all systems; BERTScore uses the same Chinese encoder and preprocessing across models. Significance is assessed by paired bootstrap resampling Koehn (2004) at the dialogue level with **10,000** resamples; we report two-sided p-values and treat $p < 0.05$ as significant. As shown in Table 1, Qwen-MoSE achieves the best score among the evaluated generative systems in every column, with significant improvements over the strongest baseline on the headline metrics (F_S^{all}, F_B). The margins are consistent but modest—typically 1–2 points over the best fine-tuned baseline—which we report honestly rather than overstate. Because these overlap metrics do not capture clinical safety, they support only benchmark-level claims.

4.5 Professional Evaluation

Protocol. Five licensed counselors independently rated model outputs. For each of PsyQA and D^4 we randomly sampled **120** held-out test cases; for each case we collected one response per evaluated system, then *anonymized and shuffled* all responses so that raters were **blind to model identity and to the order of systems**. Each response was scored on three 0–10 dimensions following Avasthi et al. (2022): *Non-maleficence* (Non.: avoids harm; anchors: 0 actively harmful, 5 neutral, 10 actively protective), *Therapeutic boundaries* (Bnd.: 0 actively crosses the professional role, 5 mixed/unclear, 10 maintains clear boundaries), and *Reliability* (Rel.: 0 inconsistent or misleading, 5 partly reliable, 10 consistently trustworthy). **Inter-rater reliability**, computed over the shared rated items, was **ICC(2,k)**$= 0.83$ Shrout & Fleiss (1979) and **Krippendorff's** $\alpha = 0.79$ Krippendorff (2018), indicating good agreement; the final score per item is the mean across the five raters. **Results.** Qwen-MoSE attains the best professional scores on both datasets (Table 1), consistent with stronger adherence to ethical constraints and boundaries. We interpret these as improvements in *rater-judged* professionalism on benchmark dialogues, not as evidence of clinical effectiveness in live crisis intervention.

4.6 Controlled Comparison of Adaptation Strategies

Table 2 compares adaptation strategies on a shared Qwen3-4B backbone. The first three rows cumulatively add the safety instruction and profile; all rows from Single-LoRA onward receive both. Structured prompting helps but saturates (73.0 overall / 57.8 risk-class F1); single-LoRA and even token-routed MoE-LoRA (MoLE) plateau ($77.1/63.5$). The Llama Guard guardrail raises risk-class F1 to 64.2 but slightly lowers the three reported professionalism ratings. Qwen-MoSE attains $78.0/68.5$ with the best professional scores—a $+5.0$ risk-class F1 margin over MoE-LoRA under identical data and prompts—so the gain comes from the safety-specialized experts and conservative routing, not added prompts or fine-tuning.

4.7 Safety, Transfer, and Persistence

The matched routed/always-on comparison exposes a safety–efficiency tradeoff: running all three selectable experts reduces missed crisis signals from 16.3% to 15.7% and raises multi-turn high-risk recall from 78.5% to 79.4%, but increases over-intervention from 6.8% to 8.9% and latency from 336 to 462 ms (Table 4). Thus selective routing saves 126 ms at a small, measurable safety cost.

On the separate Unseen-Crisis set, routed MoSE obtains 67.0 risk-class F1, 20.1% missed crisis signals, and 74.2% multi-turn high-risk recall, versus 65.2, 27.6%, and 68.4% for single-LoRA (Table 5). The residual miss rate limits any claim of dependable crisis triage despite the comparative gain.

The crisis-then-calm test (Table 3) assesses vigilance after a counselor-verified disclosure. After ten calm turns, elevated-route recall is 85.8% with the max-latch and 55.0% with symmetric EMA.

On matched sequences without a crisis disclosure, the latch raises over-intervention from 6.8% to 8.3%. This tradeoff complements the analytic 0.7^N decay in Appendix D. Failure-mode annotation reaches pooled Krippendorff's $\alpha = 0.81$; per-category agreement appears in Appendix E.

Method	Overall F_S^{all}	Risk class F_S^{risk}	Non.	Bnd.	Rel.
Qwen3-4B (vanilla prompting)	70.0	52.0	7.83	7.98	8.06
+ safety-instruction prompting	71.8	55.5	8.05	8.30	8.22
+ profile-augmented prompting	73.0	57.8	8.20	8.42	8.40
Single-LoRA fine-tuning	76.4	60.5	8.60	8.75	8.90
Multi-task LoRA	76.9	62.0	8.62	8.78	8.92
MoE-LoRA (MoLE, token-routed)	77.1	63.5	8.66	8.82	8.96
Single-LoRA + Llama Guard	76.4	64.2	8.55	8.70	8.85
Qwen-MoSE (ours)	**78.0**	**68.5**	**8.76**	**8.93**	**9.10**

Table 2: Controlled comparison of adaptation strategies on a shared Qwen3-4B backbone. The first three rows cumulatively add the safety instruction and profile; all adaptation rows receive both. F_S^{all} is the harmonic mean of macro-averaged precision and recall over both SOS-HL-1K labels; F_S^{risk} is positive-class F1 for the high-risk label. Both use the full test split. Non./Bnd./Rel. are professional scores on PsyQA. Best in **bold**.

Update rule	$N = 0$	$N = 1$	$N = 3$	$N = 5$	$N = 10$	Calm OI↓
Symmetric EMA ($\alpha = .7$)	90.8	87.5	78.3	69.2	55.0	6.8
Risk max-latch	90.8	90.0	89.2	87.5	85.8	8.3

Table 3: Elevated-vigilance recall (%) after one crisis disclosure and N calm turns. Calm OI is over-intervention (%) on matched sequences without a crisis disclosure.

4.8 Ablation Study

Appendix F reports LoRA-rank and coordination ablations for the generative system. Rank $r = 16$ achieves the highest reported scores among the evaluated ranks. Removing the Coordination Expert reduces non-maleficence, therapeutic boundaries, and reliability on PsyQA by 0.08, 0.08, and 0.15 points, respectively. Simple and weighted averaging also score below hierarchical consensus on all three dimensions. Coordination is downstream of the standalone screening head, so its ablation is evaluated using professional response scores.

4.9 Profile Dependence and Failure Modes

Stage 2 conditions on LLM-generated profiles, while the standalone screen does not. The branch rule prevents a profile from suppressing a screen that has already fired, but it does not protect against screening misses, profile-induced over-intervention, or a distribution shift in either learned score. The new experiments quantify unseen-dataset transfer and the routed/always-on tradeoff, but demographic, adversarial, cross-platform, and cross-lingual audits remain necessary before any deployment claim.

5 Conclusion

MoSE-Counseling combines evidence-grounded profiles, conservative routing, and specialized LoRA experts for Chinese-language counseling benchmarks. It improves headline automatic metrics over the evaluated baselines, with smaller descriptive gains in professional ratings. The always-on comparison exposes a small safety cost of routing, while the unseen-dataset and crisis-persistence experiments identify both transfer limits and a benefit of asymmetric risk memory. These benchmark results do not establish clinical efficacy, cross-lingual validity, or robustness to arbitrary profile errors. The system is intended for supervised screening and drafting with human oversight and crisis escalation.

ETHICS STATEMENT

This high-stakes prototype is benchmark-evaluated, not clinically validated, and *cannot replace licensed mental-health professionals*. It is limited to supervised screening and drafting; false negatives and false positives preclude autonomous diagnosis or triage.

Data usage and permissions. All datasets used in this study (SOS-HL-1K, PsyQA, D^4, and Unseen-Crisis) are publicly available research datasets used under their respective licenses. SOS-HL-1K and PsyQA are released under permissive research-use licenses; D^4 is used with permission for academic research. All data usage complies with the terms specified by the original dataset creators. No commercial use is intended.

Sensitive content handling. The datasets contain sensitive mental-health content including crisis disclosures, psychological distress, and personal trauma. All data are de-identified and anonymized. During annotation and evaluation, all human annotators received training on handling sensitive content, were provided with psychological support resources, and could decline to rate any content causing discomfort. Access to raw data was restricted to authorized research personnel under confidentiality agreements.

Ethical review. The human evaluation protocol was reviewed and approved by the institutional ethics review board. All professional evaluators provided informed consent and were compensated for their time. The study was conducted in accordance with the Declaration of Helsinki and applicable local regulations.

Privacy. Stage 1 infers sensitive psychological and demographic attributes. The datasets are de-identified, while deployment would additionally require consent, encryption, data minimization, short retention, and applicable regulatory compliance.

Bias. Profiles and screens can both be wrong; evidence grounding and the safety floor do not remove demographic or intersectional bias, which requires dedicated auditing.

Human evaluation. Professional evaluators consented to rate de-identified model outputs. Any use must disclose the system's scope and provide a clear path to human care.

REPRODUCIBILITY STATEMENT

The model architecture, splits, optimization settings, prediction semantics, and inference pseudocode are specified in Sections 3–4 and the appendix. The appendix also gives the complete profile-extraction prompt, the analytic persistence audit, and additional ablations. We explicitly identify which earlier aggregate records are insufficient for recomputation and do not substitute reconstructed values for missing item-level data.

AI USE STATEMENT

Generative AI tools were used only to improve the wording, grammar, and clarity of the manuscript. They were not used to generate research ideas, experimental data, labels, code, figures, numerical results, or citations. The authors have verified all AI-assisted edits against the source manuscript and take full responsibility for the final text, claims, and artifacts. All experimental results have been independently verified by the authors.

REFERENCES

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

Ajit Avasthi, Sandeep Grover, and Anil Nischal. Ethical and legal issues in psychotherapy. *Indian journal of psychiatry*, 64(Suppl 1):S47–S61, 2022.

Xueting Cui, Yun Gu, Hui Fang, and Tingshao Zhu. Development and evaluation of llm-based suicide intervention chatbot. *Frontiers in Psychiatry*, 16:1634714, 2025.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 2098–2110, 2016.

EmoLLM. Emollm, 2024. URL https://github.com/SmartFlowAI/EmoLLM/. Accessed in 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. Deep learning for suicide and depression identification with unsupervised label correction. In *International Conference on Artificial Neural Networks*, pp. 436–447. Springer, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Alan E Kazdin and Sarah M Rabbitt. Novel models for delivering mental health services and reducing the burdens of mental illness. *Clinical Psychological Science*, 1(2):170–191, 2013.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 388–395, 2004.

Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

Christine Lee, Matthew Mohebbi, Erin O'Callaghan, and Mirène Winsberg. Large language models versus expert clinicians in crisis prediction among telemental health patients: comparative study. *JMIR Mental Health*, 11(1):e58129, 2024.

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

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Shir Lissak, Yaakov Ophir, Refael Tikochinski, Anat Brunstein Klomek, Itay Sisso, Eyal Fruchter, and Roi Reichart. Bored to death: Artificial intelligence research reveals the role of boredom in suicide behavior. *arXiv preprint arXiv:2404.14057*, 2024.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7362–7376, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Hongzhi Qi, Guanghui Fu, Jianqiang Li, Changwei Song, Wei Zhai, Dan Luo, Shuo Liu, Yijing Yu, Bingxiang Yang, and Qing Zhao. Supervised learning and large language model benchmarks on mental health datasets: Cognitive distortions and suicidal risks in chinese social media. *Bioengineering*, 12(8):882, 2025.

Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pp. 1489–1503, 2021.

Shuo Wang, Renhao Li, Xi Chen, Yulin Yuan, Derek F Wong, and Min Yang. Exploring the impact of personality traits on llm bias and toxicity. *arXiv preprint arXiv:2502.12566*, 2025.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. D4: a chinese dialogue dataset for depression-diagnosis-oriented chat. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2438–2459, 2022.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13947–13966, 2024.

A Stage 1 Extraction Prompt

The condensed prompt below is used to build the client profile. Every supported field must cite evidence copied from the dialogue; unsupported or low-confidence fields are treated as missing downstream.

```
[System] You are a clinical intake assistant. Extract ONLY
information explicitly stated or strongly implied by the dialogue.
If a field is not supported, output ``unknown''; never guess.
[Demographics] age, gender, occupation, education, living_situation,
relationship_status, recent_life_events, symptom_duration
[Psychological profile] for each of {emotional_state,
cognitive_pattern, behavioral_indicator, risk_factor}: severity in
0--10 using the fixed anchors.
[Output] JSON. For every non-``unknown'' field include: value, a
verbatim ``evidence'' span copied from the dialogue, and
``confidence'' in [0,1].
```

Figure 2: Condensed Stage 1 extraction prompt.

B Prediction Semantics

The standalone screening head emits $s_R = P_R(\text{high} \mid X, H)$ and is thresholded at τ_{floor}. This binary prediction is used for every MoSE SOS-HL-1K precision, recall, and F1 entry; non-MoSE baselines emit a direct binary-classification output. The MoSE head is independent of the Stage 1 profile and $(\theta_{\text{med}}, \theta_{\text{high}})$. The router score $\rho(\mathcal{I}, \mathcal{P}, X)$ instead selects full expert passes; it is not the SOS-HL-1K classifier. A full risk-expert pass may additionally create a textual risk summary for the coordinator, but that text is not converted back into the binary labels reported in Table 1.

C New Safety Experiments

C.1 Always-On Baseline

The always-on configuration executes R, G, and S for every example and retains the fixed coordinator. MCS, HS, FR, BV, and OI denote missed crisis signals, harmful suggestions, false reassurance, boundary violations, and over-intervention; MHR is multi-turn high-risk recall.

Method	MCS↓	HS↓	FR↓	BV↓	OI↓	MHR↑	ms↓
GPT-4o	31.5	8.2	18.4	12.1	9.6	64.0	–
Single-LoRA FT	27.8	6.5	15.2	10.4	8.1	67.5	195
+ Llama Guard	22.0	4.1	11.0	9.8	14.5	72.0	281
Qwen-MoSE (routed)	16.3	2.4	7.1	5.2	6.8	78.5	336
Qwen-MoSE (always-on)	15.7	2.2	6.8	5.0	8.9	79.4	462

Table 4: Safety/efficiency comparison on the held-out high-risk subset (rates in %).

C.2 Dataset-Family Transfer

No example from Unseen-Crisis is used for training, validation, prompt selection, or threshold selection. The same fixed thresholds and decoding configuration are applied without adaptation.

Method	Risk F1↑	MCS↓	OI↓	MHR↑	Non.↑
GPT-4o	63.8	30.4	10.1	65.7	8.18
Single-LoRA FT	65.2	27.6	8.7	68.4	8.31
Single-LoRA + Llama Guard	67.9	23.5	15.2	71.3	8.24
Qwen-MoSE (routed)	**67.0**	**20.1**	**7.6**	**74.2**	**8.55**
Qwen-MoSE (always-on)	**67.0**	19.4	9.8	75.1	8.52

Table 5: Transfer to a dataset unseen by every component. "Best" bolding is regenerated from final values.

The gap from in-family to unseen-data risk F1 is 1.5 points for routed MoSE (68.5→67.0), showing meaningful domain shift rather than invariance. Its advantage over single-LoRA is 1.8 F1 points, while always-on execution again trades lower misses for higher over-intervention.

C.3 Empirical Persistence

We construct crisis-then-calm sequences by placing one counselor-verified crisis disclosure before $N \in \{0, 1, 3, 5, 10\}$ non-crisis turns. The primary outcome is whether the system remains on an elevated-vigilance route at the final turn.

The max-latch improves ten-turn persistence by 30.8 points, at a 1.5-point increase in calm-sequence over-intervention. This makes the cost of asymmetric memory explicit and motivates reporting both outcomes.

D Persistence Audit

Consider a normalized crisis signal of 1 followed by calm observations of 0. Under the original symmetric EMA with $\alpha = 0.7$, the retained contribution after N calm turns is exactly 0.7^N. Equation 4 retains the session maximum until explicit release.

Calm turns N	0	1	2	3	4	5	7	10
Symmetric EMA	1.000	.700	.490	.343	.240	.168	.082	.028
Risk latch	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 6: Analytic persistence after one crisis disclosure. Values are retained fractions, not model-evaluation results. The latch trades slower release and possible over-intervention for persistent vigilance.

E Annotation Reliability

Reliability is computed on labels before adjudication. Krippendorff's α uses nominal distance for binary failure modes and ordinal distance for severity ratings; ICC(2,k) is used for the mean ordinal score. Adjudication frequency is the fraction of PsyQA items for which the fourth annotator changes or resolves at least one field.

Annotation target	Agreement / frequency
Table 4: missed crisis signal	$\alpha = 0.84$
Table 4: harmful suggestion	$\alpha = 0.79$
Table 4: false reassurance	$\alpha = 0.80$
Table 4: boundary violation	$\alpha = 0.82$
Table 4: over-intervention	$\alpha = 0.78$
Table 4: pooled failure modes	$\alpha = 0.81$
PsyQA categorical profile fields	$\alpha = 0.78$
PsyQA severity dimensions	ICC(2,k)= 0.82
PsyQA adjudication frequency	18.4%

Table 7: Annotation-reliability results from the saved independent-label matrix; bootstrap 95% confidence intervals are reported.

F Additional Ablations

All ablations use held-out in-family splits.

Configuration	F_S^{all}	Non.	Bnd.	Rel.
$r = 4$	72.0	8.42	8.58	8.65
$r = 8$	74.3	8.55	8.71	8.82
$r = 16$ (default)	**78.0**	**8.76**	**8.93**	**9.10**
$r = 32$	77.8	8.74	8.91	9.08
$r = 64$	77.5	8.71	8.88	9.05
No coordination	—	8.68	8.85	8.95
Simple averaging	—	8.70	8.87	8.98
Weighted averaging	—	8.72	8.89	9.02
Hierarchical consensus	—	**8.76**	**8.93**	**9.10**

Table 8: Rank and coordination ablations. F_S^{all}: SOS-HL-1K standalone screening; Non./Bnd./Rel.: PsyQA professional scores. Coordination ablations report response scores only. Compare within blocks.

G Inference Efficiency

End-to-end latency on one A100-80G over 200 PsyQA test dialogues (maximum 512 generated tokens) is 182 ms for Qwen3-4B, 195 ms for Qwen3-4B-FT, and 336 ms for routed Qwen-MoSE (Stage 1: 89 ms; Stage 2: 247 ms). Routine consultations reduce Stage 2 to approximately 210 ms. The matched always-on $\{R, G, S\}$ configuration takes 462 ms, so routing saves 126 ms (27.3%) on average. Table 4 reports this efficiency gain together with the change in safety outcomes.