# TRIAD-TS: Explainable Multi-Agent Orchestration for Joint Task–Style Alignment on On-Device Mental Health LLMs

**Anonymous ACL submission**

## Abstract

On-device LLM personalization must satisfy two constraints simultaneously: (i) task correctness and (ii) user-specified response style, under strict resource limits. We present TRIAD-TS, a multi-agent orchestration framework that (1) decomposes a request into task and style intents, (2) retrieves evidence in an explainable latent space using precomputed intent centroids, (3) blends style adapters via an explicit weight vector, and (4) applies a learned quality-driven abstention policy that avoids unreliable outputs without conflating uncertainty with device feasibility. We also introduce TriadBench-TS, a benchmark spanning 8 task categories and 12 therapeutic communication styles, with verification that filters invalid rewrites. Using ELO-based pairwise evaluation across three dimensions (task, style, joint), TRIAD-TS achieves 3.7% higher joint ELO rating (+41 points) than state-of-the-art adapter composition methods and reduces style drift by 52% compared to instant adapter blending (CRAYON), while providing transparent rationales through centroid matches and adapter-weight explanations. Code and data are available at `https://anonymous.4open.science/r/TRAID-02C1`.

## 1 Introduction

Mental health support requires AI systems balancing therapeutic correctness with appropriate interpersonal tone while protecting sensitive user data. Unlike general assistants, therapeutic AI faces unique constraints: users share trauma histories, suicidal ideation, and relationship conflicts demanding both clinical appropriateness and empathetic delivery. Technically correct interventions in wrong tone damage therapeutic alliance (Norcross, 2011), while supportive responses lacking clinical grounding cause harm (American Psychological Association, 2017). This creates acute need for on-device deployment—cloud systems introduce privacy risks undermining trust essential for therapeutic disclosure (Martinez-Martin et al., 2020)—but on-device processing imposes severe constraints. **TRIAD-TS addresses this through a hybrid architecture: client-side models process sensitive dialogue locally and extract compact, anonymized signals (task weights, style indicators); only these de-identified signals reach server-side coordination agents, ensuring raw therapeutic content never leaves the device.** Within these constraints, responses must simultaneously satisfy *what* the user needs therapeutically (task intent) and *how* they need it communicated (style intent). Task and style are not independent: directive safety planning in cold clinical language violates person-centered principles (Rogers, 1957).

While adapter composition techniques have matured (Bang et al., 2024; Huang et al., 2023; Pfeiffer et al., 2021), their application to mental health remains unexplored. Existing methods optimize general task performance, but therapeutic contexts require: (i) privacy-preserving on-device processing, (ii) clinical appropriateness beyond user preference, (iii) interpersonal calibration where tone constitutes therapeutic action, and (iv) safety-critical decision-making. Crayon (Bang et al., 2024) blends adapters via embedding similarity without conflict resolution; XPerT (Wang et al., 2025) selects entire models, forcing all-or-nothing choices. Neither provides explainability essential for clinical adoption (Torous and Keshavan, 2020) nor distinguishes computational exhaustion from genuine clinical risk.

We present TRIAD-TS, the first multi-agent framework bringing adapter composition to mental health support through orthogonal factorization with clinically-informed coordination. Offline, we build (i) **task-specialized adapters corresponding to the 8 ESConv support strategies (Liu et al., 2021)** via mixture-of-experts training (Shazeer et al., 2017), and (ii) style adapters from paired

response comparisons (Christiano et al., 2017). Online, a lightweight client sends compact JSON signals to server-side agents: a Clinical Assessment Agent infers therapeutic priorities; a Communication Strategy Agent determines style; a deterministic module enforces compatibility and applies quality-driven abstention. We introduce TriadBench-TS (1,847 test examples from 8 tasks × 12 styles). Using ELO-based evaluation (Zheng et al., 2023), TRIAD-TS achieves +41 ELO over XPerT, 52% lower style drift than Crayon.

## 2 Related Work

LoRA (Hu et al., 2021) enables parameter-efficient fine-tuning. Subsequent work develops composition methods: AdapterFusion (Pfeiffer et al., 2021) via attention, LoraHub (Huang et al., 2023) via gradient-free search, Crayon (Bang et al., 2024) via instant embedding similarity. MoE approaches (Shazeer et al., 2017) inspire our soft mixture training where adapters specialize via centroid proximity. However, these methods optimize general NLP tasks without considering clinical appropriateness or safety, treat style implicitly rather than as independent composable dimension, and lack conflict resolution for task-style misalignment. XPerT (Wang et al., 2025) selects cached models but cannot independently adjust task and style.

Prior mental health chatbots (Fitzpatrick et al., 2017; Inkster et al., 2018) use rule-based systems without personalization. MentalLLaMA (Yang et al., 2024) fine-tunes on mental health data but treats style implicitly. Multi-agent systems for therapy exist: MindAgent (Gong et al., 2023) separates assessment and intervention but requires cloud deployment; TherapyBot (Sharma et al., 2023) uses specialized agents but needs 15+ seconds; PARTNER (Liu et al., 2024) adapts style but treats it monolithically. None address: (i) privacy via on-device processing, (ii) sub-3s latency constraints, (iii) explainable clinical reasoning, or (iv) principled abstention separating resource limits from clinical concerns.

Privacy risks in digital mental health (Martinez-Martin et al., 2020) motivate on-device approaches. Explainability is non-negotiable for clinical adoption (Torous and Keshavan, 2020); we leverage Sentence-BERT (Reimers and Gurevych, 2019) for interpretable centroid similarity. General multi-agent frameworks (MetaGPT (Hong et al., 2023), AutoGen (Wu et al., 2023), ReAct (Yao et al., 2023)) decompose tasks into specialized roles but target cloud deployment without domain constraints. Our design differs: (i) server agents process only compact JSON signals maintaining client privacy, (ii) orthogonal factorization (clinical priorities vs. communication strategy) enables parallel reasoning, (iii) deterministic integration enforces compatibility constraints from counseling frameworks (Hill, 2014). Style control methods (Keskar et al., 2019; Ouyang et al., 2022) treat style aesthetically; we recognize task-style interdependence through compatibility matrices. Pairwise preference learning (Christiano et al., 2017) inspires our drift extraction. Abstention methods (Chow, 1970; Geifman and El-Yaniv, 2017) conflate failure modes; we separate feasibility gating from quality-based abstention, critical for safety. ELO ratings (Zheng et al., 2023) enable our multi-dimensional therapeutic assessment.

## 3 TriadBench-TS: A Verified Task–Style Benchmark

To enable rigorous evaluation of joint task-style alignment in mental health counseling, we introduce **TriadBench-TS**, constructed from two complementary sources: ESConv (Liu et al., 2021) (14,855 task-labeled therapeutic utterances, avg. 20.2 words) and MentalChat16K (Xu et al., 2025) (16,113 caregiver interactions, avg. 185.4 words). While ESConv provides concise strategy-focused responses, MentalChat16K offers longer naturalistic conversations reflecting real-world counseling complexity.

### 3.1 Task-Style Factorization

We adopt ESConv's **eight support strategies** (Liu et al., 2021) as our task taxonomy following Hill's Helping Skills Theory (Hill, 2014) (detailed definitions in Appendix F): QUESTION, RESTATEMENT OR PARAPHRASING, REFLECTION OF FEELINGS, SELF-DISCLOSURE, AFFIRMATION AND REASSURANCE, PROVIDING SUGGESTIONS, INFORMATION, and OTHER. The first seven categories correspond to specific therapeutic actions; OTHER captures pleasantries and miscellaneous support. Throughout this paper, "task" refers to these 8 evidence-based therapeutic actions; "task adapters" are the **7 LoRA modules trained on the first 7 categories** (excluding OTHER, which requires abstention rather than adapter routing); and "task intent" denotes which therapeutic action is clinically

appropriate for a given context.

For MentalChat16K (no task labels), we use **GPT-4-turbo** with 20-shot prompting (Appendix B). Three licensed clinicians verified 800 samples, achieving Fleiss' $\kappa = 0.71$ across both datasets. Examples labeled as OTHER are used for style training but excluded from task adapter pools, as they represent contexts requiring abstention or clarification rather than specific therapeutic interventions.

We generate style variants spanning **12 therapeutic communication styles** (Appendix E). For 27,997 base examples, we generate 15 candidates per example using GPT-3.5-turbo, producing 419,955 candidates. A five-stage filtering pipeline (SimHash deduplication, Sentence-BERT semantic verification, GPT-4 style compliance, length constraints, toxicity filtering) eliminates 81.0%, yielding **79,972 verified variants**. Three clinicians verified 3,500 samples (ICC=0.71 for semantic preservation, 0.68 for stylistic authenticity), confirming 73.6% meet quality standards (details in Appendix C).

**Semantic Preservation.** Style augmentation increases length by 15–20% through stylistic elaboration (e.g., "try setting goals" → "consider trying to set small, achievable goals") while preserving task category via explicit constraints: interrogative structures for QUESTION, emotional vocabulary for REFLECTION, actionable content for SUGGESTIONS. Verified via GPT-4 (threshold: 0.85 Sentence-BERT similarity).

### 3.2 Test Set with Hard-OOD Evaluation

Test sets use ESConv's 20% holdout (n=2,971) exclusively, ensuring gold-standard human annotations and clean separation from training. We allocate 1,847 for Standard (in-distribution) and 1,124 for Out-of-Distribution (OOD) evaluation. **OOD** refers to scenarios where inputs differ systematically from training distributions—novel task categories, communication styles, or both—critical for real-world deployment where user needs exceed predefined taxonomies.

Three Hard-OOD subsets test different failure modes:

- **Task-OOD** (n=539): **exclusively** OTHER-labeled queries (rapport-building pleasantries, crisis intervention, grief counseling) excluded from the 7-adapter training pool, testing abstention rather than forced routing.

Table 1: TriadBench-TS composition. **In-Task** counts examples labeled with the 7 trained task categories (excluding OTHER). Standard and Style-OOD reflect natural task distribution from ESConv holdout (∼75% in-pool, ∼25% OTHER); Task-OOD and Both-OOD are exclusively OTHER by design.

| Subset | Source | Base | In-Task | Var. | Total | Use |
|---|---|---|---|---|---|---|
| *Training* | | | | | | |
| Task-labeled | ESConv (80%) | 11,884 | 9,727 | 35,658 | 47,542 | T+S |
| Unlabeled | MentalChat16K | 16,113 | 12,085 | 44,314 | 60,427 | S |
| *Subtotal* | | *27,997* | *21,812* | *79,972* | *107,969* | |
| *Evaluation (ESConv 20%)* | | | | | | |
| Standard | ESConv | 1,847 | 1,388 | – | 1,847 | In-d |
| Task-OOD | ESConv | 539 | 0 | – | 539 | OOD |
| Style-OOD | ESConv | 461 | 346 | – | 461 | OOD |
| Both-OOD | ESConv | 124 | 0 | – | 124 | OOD |
| *Subtotal* | | *2,971* | *1,734* | – | *2,971* | |
| **Total** | | **30,968** | **23,546** | **79,972** | **110,940** | |

- **Style-OOD** (n=461): culturally-specific or technical communication absent from 12 styles. **Task distribution reflects natural ESConv holdout (75% in-pool tasks, 25% OTHER)**, testing style fallback under task variance.
- **Both-OOD** (n=124): compounded misalignment (both task and style out-of-pool).

**Note:** "In-Task" in Table 1 denotes examples whose task label belongs to the 7 trained categories. Standard and Style-OOD naturally contain mixed task distributions from ESConv's holdout set, while Task-OOD and Both-OOD are **exclusively** OTHER-labeled by construction. **Note: OTHER encompasses both low-risk contexts (greetings, administrative) and high-risk situations (crisis, grief) that both warrant specialized handling beyond standard adapter selection. Style-OOD** includes culturally-specific communication and technical language (Appendix G). Three clinicians annotated all labels ($\kappa = 0.71$, OOD agreement: 87.3%).

Table 1 summarizes TriadBench-TS: **107,969 training examples** and **2,971 expert-annotated test cases**.

## 4 TRIAD-TS Framework

Figure 1 overviews TRIAD-TS. Given a client context $\mathbf{x}_{\text{client}}$, TRIAD-TS builds task and style adapter pools offline. At deployment time, a lightweight client-side module extracts compact signals and sends a JSON payload to a server-side multi-agent system. The server agents infer (i) clinical priorities and (ii) communication intent, after which a deterministic integration module assembles a personalized model $\mathcal{M}_{\text{TRIAD}}$. Algorithm 1 details offline construction; Algorithm 2 summarizes online

**Algorithm 1** Offline: Build Task/Style Adapter Pools

**Require:** $\mathcal{M}_{\Phi_0}, \mathcal{D}_{\text{tr}}, \{\mathcal{D}_s\}_{s\in\mathcal{S}}$
**Ensure:** $\{(\ell_{\theta_n}, \mathbf{c}_n)\}_{n=1}^{7}, \{(\mathcal{A}_s, \mathbf{z}_s)\}_{s\in\mathcal{S}}$

```
Extract q_x = M_{Φ0}^{(16)}(x) for first 7 task categories, PCA (128-dim), K-means → {c_n}_{n=1}^7
Initialize 7 LoRAs {ℓ_{θ_n}}_{n=1}^7
for epoch = 1..10 do
    for minibatch (x, y) do
        Compute α_n(q_x); Update {θ_n} (Eq. 2)
    end for
end for
for all s ∈ S do
    M_{Φ_s} ← FineTune(M_{Φ0}, D_s); Collect {h_i^{(sum)}}
end for
Build basis {v_j}; project z_s, convert A_s (Appendix A)
```

**Algorithm 2** Online: Multi-Agent Selection + Integration

**Require:** Client context $\mathbf{x}_{\text{client}}$, pools, $t(n)$, $\mathbf{C}$, $\mathbf{T}$
**Ensure:** $\mathcal{M}_{\text{TRIAD}}, \{\bar{\alpha}_n\}_{n=1}^{7}, \{\gamma_s^{**}\}_{s\in\mathcal{S}}$

```
Extract signals e_ind, {α*_n}, s_style; send JSON u to server
p̃ ← ClinicalAgent(u); w̃_style ← CommAgent(u)
Solve {γ*_s} (Eq. 4); compute {ᾱ_n} (Eq. 6)
Compute c_compat, {γ**_s} (Eq. 7); assemble M_TRIAD (Eq. 8)
```

selection and integration.

### 4.1 Offline Adapter Pools

**Task pool.** We extract layer-16 representations from $\mathcal{M}_{\Phi_0}$, apply PCA (128-dim, 89.3% variance), and cluster **examples from the first 7 task categories** into $N = 7$ centroids via K-means. We train **7** LoRA modules via soft mixture (Eq. 1, 2), where soft assignment weights $\alpha_n(\mathbf{q}_x)$ are normalized cosine similarities to centroids. Online, we select Top-$K_{\text{task}} = 3$ adapters via temperature-scaled softmax ($\tau = 0.1$).

$$\alpha_n(\mathbf{q}_x) = \frac{\cos(\mathbf{c}_n, \mathbf{q}_x) + 1}{2} \tag{1}$$

$$\max_{\{\theta_n\}} \sum_{(x,y)} \sum_{t} \log p_{\Phi_0+\Delta\Phi(\Theta_x)}(y_t \mid x, y_{<t}) \tag{2}$$

**Style pool.** For each style $s \in \mathcal{S}$, we fine-tune $\mathcal{M}_{\Phi_0}$ on $\mathcal{D}_s$ to obtain $\mathcal{M}_{\Phi_s}$. Using $M = 50$ standardized prompts (Appendix B.1), we compare responses via summarization model $\mathcal{M}_{\text{sum}}$ and extract drift vectors:

### 4.2 Online Multi-Agent Selection and Deterministic Integration

**Privacy-preserving computation flow:** At deployment time, a lightweight **client module running on-device** performs the following operations locally:

1. Embeds context: $\mathbf{q}_{\text{client}} = \mathcal{M}_{\Phi_0}^{(16)}(\mathbf{x}_{\text{client}})$
2. Computes cosine similarities to all 7 task centroids
3. Selects Top-$K_{\text{task}} = 3$ adapters via Eq. 3
4. Extracts style indicators and other signals

**Only the resulting compact JSON payload $\mathbf{u}$** (containing $\{\alpha_n^*\}$, $\mathbf{s}_{\text{style}}$, $\mathbf{e}_{\text{ind}}$, but **no raw text**) is sent to server-side agents. Server agents output clinical priorities $\tilde{\mathbf{p}}$ and communication intent $\tilde{\mathbf{w}}_{\text{style}}$, which are returned to the client for final on-device integration. The clinical assessment agent reads $\mathbf{u}$ and outputs normalized priorities $\tilde{\mathbf{p}} \in \mathbb{R}^5$ over evidence-based therapeutic approaches (CBT, DBT, MI, Psychoeducation, Supportive), while the communication strategy agent outputs a weighted intent $\tilde{\mathbf{w}}_{\text{style}}$ over communication descriptors. **(Recall: embedding and Top-$K$ selection already occurred client-side; $\mathbf{u}$ contains the resulting $\{\alpha_n^*\}$ weights.)** The task selection formula for reference:

$$\alpha_n^* = \begin{cases} \frac{\exp(\cos(\mathbf{c}_n, \mathbf{q}_{\text{client}})/\tau)}{\sum_{n'\in\text{TopK}} \exp(\cos(\mathbf{c}_{n'}, \mathbf{q}_{\text{client}})/\tau)} & n \in \text{TopK} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

We set $\tau = 0.1$ and $K_{\text{task}} = 3$ in all experiments.

We solve simplex-constrained least squares (Eq. 4) to blend style adapters $\{\gamma_s^*\}$, enabling continuous interpolation rather than binary selection.

$$\{\gamma_s^*\} = \arg\min_{\gamma} \left\| \sum_s \gamma_s \mathbf{z}_s - \tilde{\mathbf{w}}_{\text{style}} \right\|_2^2 \quad \text{s.t. } \gamma_s \geq 0, \sum_s \gamma_s = 1. \tag{4}$$

We modulate the selected task weights using the clinical priorities. Let $t(n) \in \{1, \ldots, 7\}$ denote the task category index for adapter $n$, where categories map to: (1) QUESTION, (2) RESTATEMENT, (3) REFLECTION, (4) SELF-DISCLOSURE, (5) AFFIRMATION, (6) SUGGESTIONS, (7) INFORMATION. **(OTHER is excluded from the adapter pool.)** The clinical agent outputs priorities $\tilde{\mathbf{p}} \in \mathbb{R}^5$ over therapeutic approaches (CBT, DBT, MI, Psychoeducation, Supportive). We map these to task categories via a learned compatibility matrix $\mathbf{T} \in \mathbb{R}^{7\times5}$ where $T_{ij}$ indicates how strongly task $i$ aligns with approach $j$ (e.g., $T_{6,\text{CBT}} = 1.0$ since SUGGESTIONS strongly aligns with CBT's behavioral activation; $T_{1,\text{MI}} = 0.9$ since QUES-

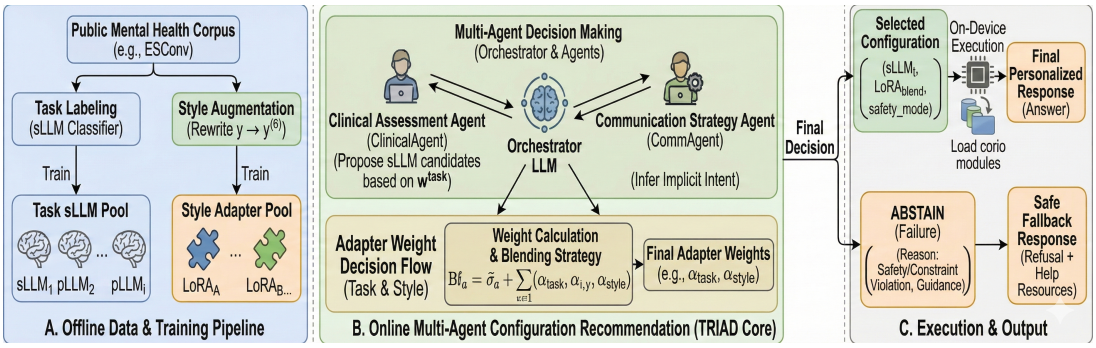


Figure 1: TRIAD-TS overview. **Offline**: build task-specialized LoRA pool (**7 adapters** for first 7 ESConv strategies) and style adapter pool (12 therapeutic styles). **Online**: client **locally** computes task weights and style signals, sends only compact anonymized signals to server agents for coordination; final model assembly occurs **client-side**, ensuring raw dialogue never leaves device.

TION enables motivational exploration). The task-level priority becomes:

$$p_{\text{task}}(n) = \sum_{j=1}^{5} T_{t(n),j}\, \tilde{p}_j. \tag{5}$$

We then apply priority modulation and renormalization:

$$\bar{\alpha}_n = \frac{\alpha_n^* \big(1 + \lambda\, p_{\text{task}}(n)\big)}{\sum_{n'=1}^{7} \alpha_{n'}^* \big(1 + \lambda\, p_{\text{task}}(n')\big)}. \tag{6}$$

Next, we enforce task-style compatibility via a fixed matrix $\mathbf{C} \in [0,1]^{N \times |\mathcal{S}|}$. **Here $N = 7$ denotes the number of LoRA adapters (one per ESConv task category, excluding OTHER), and $|\mathcal{S}| = 12$ denotes the number of therapeutic styles.** Each entry $C_{ns}$ encodes how appropriate style $s$ is for the therapeutic action represented by adapter $n$. For example, $C_{n=\text{Question},s=\text{Empathic}} = 1.0$ (exploration benefits from warmth), while $C_{n=\text{Question},s=\text{Confrontational}} = 0.3$ (challenge inappropriate during early exploration). The matrix is initialized from counseling guidelines (Hill, 2014; Miller and Rollnick, 2012) and remains fixed during inference. We compute an aggregate compatibility score and (only if needed) reweight style coefficients, then renormalize:

$$\begin{aligned} c_{\text{compat}} &= \sum_{n=1}^{7} \sum_{s \in \mathcal{S}} \bar{\alpha}_n\, \gamma_s^*\, C_{ns}, \\ \rho_s &= \sum_{n=1}^{7} \bar{\alpha}_n\, C_{ns}, \\ \gamma_s^{**} &= \begin{cases} \dfrac{\gamma_s^*\, \rho_s}{\sum_{s' \in \mathcal{S}} \gamma_{s'}^*\, \rho_{s'}} & c_{\text{compat}} < \tau_{\text{compat}}, \\ \gamma_s^* & \text{otherwise.} \end{cases} \end{aligned} \tag{7}$$

Finally, we assemble the personalized model by directly combining the task pool and the (possibly adjusted) style pool:

$$\mathcal{M}_{\text{TRIAD}} = \mathcal{M}_{\Phi_0} + \omega \sum_{n=1}^{7} \bar{\alpha}_n \ell_{\theta_n} + (1-\omega) \sum_{s \in \mathcal{S}} \gamma_s^{**} \mathcal{A}_s. \tag{8}$$

### 4.3 Implementation Details

We use Llama-2-7B-Chat as the baseline SLLM $\mathcal{M}_{\Phi_0}$ and Llama-3.1-8B-Instruct as the summarization model $\mathcal{M}_{\text{sum}}$. Each LoRA uses $r = 16$, $\alpha = 32$ ($\alpha/r = 2.0$ prevents mode collapse). Training: 10 epochs, AdamW (lr=$2 \times 10^{-4}$), batch size 32 with gradient accumulation, FP16, 4× A100. Full details in Appendix A.

We use Llama-2-7B-Chat as $\mathcal{M}_{\Phi_0}$ and Llama-3.1-8B-Instruct as $\mathcal{M}_{\text{sum}}$. Each LoRA uses $r = 16$, $\alpha = 32$. Key hyperparameters: $K_{\text{task}} = 3$, $\tau = 0.1$, $\lambda = 0.5$, $\omega = 0.6$ (details in Appendix A).

## 5 Experiments

We evaluate joint task-style alignment on TriadBench-TS's Standard test set (N=1,847 from ESConv holdout), where each dialogue context is annotated with gold task and style labels, enabling both label-based routing evaluation and end-to-end quality assessment. **Standard test set reflects natural ESConv task distribution: 1,388 in-pool tasks (75.1%) and 459 OTHER (24.9%), ensuring evaluation under realistic mixed conditions rather than artificially filtered datasets.**. Gold labels: ESConv annotations ($\kappa = 0.79$), OOD clinician labels ($n = 3$, $\kappa = 0.71$), GPT-4-verified styles ($n = 800$, $\kappa = 0.71$). To test calibrated behavior under distribution shift, we use three Hard-

OOD subsets from the ESConv holdout: (i) **Task-OOD** (n=539) where the task is labeled OTHER (not in the 7-adapter pool) while style is in-pool, (ii) **Style-OOD** (n=461) where style is culturally-specific (out-of-12-styles) **with natural task distribution from holdout (∼75% in-pool, ∼25% OTHER)**, and (iii) **Both-OOD** (n=124) where both task and style are out-of-pool.

We compare routing strategies by varying task and style routers while keeping the same backbone model and adapter pools. Baseline configurations include CRAYON/CRAYON (C/C), XPerT/XPerT (X/X), XPerT/CRAYON (X/C), and LoraHub/LoraHub (L/L), where the first component specifies the task router and the second specifies the style router. We directly compare TRIAD-TS against a CRAYON-inspired edge-server hybrid baseline (CRAYON-Hybrid) that offloads unconfident queries to a server LLM and replaces the on-device output. CRAYON-Hybrid (C+O) offloads low-confidence queries to a server LLM (threshold tuned on n=300 validation set). TRIAD-TS assigns task routing to CRAYON and style routing to XPerT, coordinated through multi-agent orchestration with deterministic integration.

Routing accuracy is measured via Top-1 and Top-3 correctness (Task@1/3, Style@1/3) using gold labels from TriadBench-TS. Table 2 reports performance on the Standard in-distribution test set only; OOD performance appears in Table 5 with behavior distributions rather than routing metrics, as OOD scenarios by definition lack matching adapters to route to.

For Hard-OOD splits, we report behavior distributions and counseling quality under the final response path, with TRIAD-TS behaviors classified as Proceed (use in-pool adapters), Fallback (safe nearest in-pool style while keeping task), or Abstain (ask targeted clarifications), and CRAYON-Hybrid behaviors as Proceed (in-pool blending) or Server (offload with output replacement).

Counseling quality is assessed using two independent LLM judges (Claude-3-Opus, Gemini-1.5-Pro) producing 7 criterion scores per response on a 1-10 scale: Listen, Empathy, Safety, Non-judgmentalness, Clarity, Boundaries, and Holistic quality. For each system pair, we conduct randomized blind pairwise comparisons where each judge evaluates both responses and declares a winner. **Each comparison involves 2 judges, thus producing 2 independent battles for ELO computation.** With N=1,847 test examples and 4 baseline systems plus TRIAD-TS, we conduct $\binom{5}{2} = 10$ pairwise comparisons, yielding $10 \times 1{,}847 \times 2 = 36{,}940$ total battles. We compute ELO ratings (init 1000, K=32) with 95% bootstrap confidence intervals over 1,000 resamples.

We additionally quantify Style-Drift% as the fraction of cases where responses are useful but stylistically off (StyleScore $< 4.0$ while TaskUsefulness $\geq 6.0$), where StyleScore aggregates Empathy, Non-judgmentalness, and Boundaries dimensions. We conduct three sets of ablations: (i) remove one server-side assessment agent at a time (ClinicalAgent or CommAgent), compare against a single-LLM router without role separation, and evaluate a deterministic-only variant without server LLM routing, (ii) evaluate TRIAD-TS against Debate and Critic-Refine orchestration patterns, (iii) evaluate TRIAD-TS across five on-device SLLMs and five server routers.

## 6 Results

Results establish four key findings: (i) TRIAD-TS achieves best end-to-end counseling quality via joint task-style optimization, (ii) this advantage is mechanistic—simultaneous routing accuracy directly suppresses "useful-but-wrong-tone" failures, (iii) TRIAD-TS demonstrates superior calibration under distribution shift via explicit abstain/fallback semantics, and (iv) multi-agent role separation is essential—removing agents materially degrades quality.

### 6.1 End-to-End Counseling Quality

TRIAD-TS removes the task-style trade-off plaguing existing methods. Baselines exhibit complementary failures: CRAYON/CRAYON achieves strong task routing but under-controls style (Style@1=0.63), while XPerT/XPerT strengthens style at the cost of task accuracy (Task@1=0.70). TRIAD-TS achieves best performance on both dimensions simultaneously (T@1=0.81, S@1=0.78) (Table 2). This dual optimization is critical: task errors mis-frame clinical intent, while style errors violate relationship norms even when content is correct.

TRIAD-TS achieves G-Eval ELO 1162±8, surpassing XPerT/XPerT (1121±9) by +41 and CRAYON/CRAYON (1092±10) by +70 (Table 3), with broad gains across all criteria (avg. 8.36/10). The framework also reduces style drift—the critical failure mode where responses are useful but interper-

sonally miscalibrated—to 10.8%, representing a 27% relative reduction versus XPerT/XPerT and 52% versus CRAYON/CRAYON. This directly validates the core design target: preventing tone, judgment, and boundary violations even when content is technically correct.

Table 4 provides a detailed breakdown of G-Eval scores across all 7 counseling quality dimensions, evaluated independently by two LLM judges (Claude-3-Opus and Gemini-1.5-Pro). TRIAD-TS achieves largest gains on Empathy (+0.88/+0.97 over XPerT for Claude/Gemini judges) and Boundaries (+0.91/+0.98), the exact dimensions where naive adapter blending breaks—responses can be technically helpful yet interpersonally invalidating. TRIAD-TS also improves Clarity and Holistic scores, demonstrating that safety gains come from principled task-style coordination rather than generic outputs.

### 6.2 Robustness Under Distribution Shift and Ablation Studies

TRIAD-TS achieves superior calibration under distribution shift compared to server offloading. CRAYON-Hybrid reacts via escalation (74% on Task-OOD, 88% on Both-OOD), over-committing without clarifying intent (Table 5). TRIAD-TS treats OOD as a decision problem: Task-OOD triggers clarification (67% abstain), Style-OOD falls back to safe stances (56% fallback). This yields higher quality (8.05 vs. 7.18) and lower style drift (6.2% vs. 19.5%).

CRAYON-Hybrid is evaluated separately in OOD (Table 5) due to incompatible deployment paradigms and variable offload rates (26-88%). Other baselines lack abstention mechanisms (100% Proceed), providing no calibration insight.

Ablation studies confirm multi-agent role separation is essential. Communication Agent removal causes the largest drop (34 ELO), consistent with gains in empathy and boundaries (Table 6). Single-router alternatives underperform (27 ELO), amplifying over-commitment and style drift. Deterministic-only variants (71 ELO) show compatibility enforcement alone cannot replace multi-agent reasoning.

Generic orchestration patterns underperform: Debate (1139 ELO) and Critic-Refine (1131 ELO) introduce intent oscillation and style inconsistency (Table 7). TRIAD-TS achieves best stability via structured intent signals and deterministic integration.

### 6.3 Deployment Robustness

Model sweep across 5 on-device SLLMs and 5 server routers demonstrates robustness (Table 13, Appendix I). Stronger SLLMs consistently yield higher ELO (1165 for Qwen2.5-1.5B vs. 1110 for SmolLM, 55-point range), indicating on-device capacity sets the quality ceiling. Router choice induces smaller variations (±5-7 ELO), confirming framework robustness to substitution. Qualitative examples in Appendix H demonstrate how TRIAD-TS handles profile personalization, ambiguity, and task-style conflicts through principled coordination.

| Method | T@1 | T@3 | S@1 | S@3 |
|---|---|---|---|---|
| CRAYON/CRAYON | 0.77 | 0.92 | 0.63 | 0.82 |
| XPerT/XPerT | 0.70 | 0.89 | 0.74 | 0.90 |
| XPerT/CRAYON | 0.70 | 0.89 | 0.63 | 0.82 |
| LoraHub/LoraHub | 0.64 | 0.84 | 0.61 | 0.79 |
| **Ours: TRIAD-TS** | **0.81** | **0.95** | **0.78** | **0.93** |

Table 2: Adapter routing accuracy on TriadBench-TS Standard test set (N=1,847) using gold human annotations from ESConv. TRIAD-TS is best on both task and style routing simultaneously.

## 7 Conclusion

We presented TRIAD-TS, a multi-agent framework for on-device mental health support achieving joint task-style optimization through clinically-informed coordination. On TriadBench-TS (1,847 test examples, 8 tasks, 12 styles), TRIAD-TS achieves +41 ELO over XPerT and 52% lower style drift than CRAYON. Multi-agent role separation proves essential (34 ELO when removing Communication Agent), providing explainability through interpretable centroids and adapter weights with calibrated abstention. Our work establishes that therapeutic AI requires domain-specific design with clinical compatibility constraints. Future work should explore multilingual personalization, federated training, and human studies validating explainability with clinicians and clients.

## 8 Limitations

TRIAD-TS is a research framework, not a production clinical system. Safety mechanisms reduce obvious harms but do not guarantee medical correctness. Our 12 styles cannot represent all cultural norms or therapeutic modalities. The 8-task taxonomy from ESConv, while evidence-based, may

| Method | G-Eval ELO | Avg. G-Eval (1–10) | Style-Drift % |
|---|---|---|---|
| CRAYON/CRAYON | $1092 \pm 10$ | 6.80 | 22.4 |
| XPerT/XPerT | $1121 \pm 9$ | 7.50 | 14.7 |
| XPerT/CRAYON | $1106 \pm 9$ | 7.01 | 20.6 |
| LoraHub/LoraHub | $1076 \pm 11$ | 6.61 | 18.9 |
| **Ours: TRIAD-TS** | $\mathbf{1162 \pm 8}$ | **8.36** | **10.8** |

Table 3: Main results on Standard test set (N=1,847). TRIAD-TS achieves +41 ELO over XPerT/XPerT and +70 over CRAYON/CRAYON. Avg. G-Eval is mean over 7 criteria and 2 judges (36,940 total battles). TRIAD-TS achieves best quality and lowest style drift.

| | Listen↑ | | Empathy↑ | | Safety↑ | | Non-judge↑ | | Clarity↑ | | Boundaries↑ | | Holistic↑ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method (task/style)** | **Claude** | **Gemini** | **Claude** | **Gemini** | **Claude** | **Gemini** | **Claude** | **Gemini** | **Claude** | **Gemini** | **Claude** | **Gemini** | **Claude** | **Gemini** |
| CRAYON/CRAYON | 7.14 | 6.78 | 6.58 | 6.32 | 7.19 | 6.91 | 7.38 | 7.00 | 6.82 | 6.47 | 6.93 | 6.57 | 6.69 | 6.42 |
| XPerT/XPerT | 7.63 | 7.28 | 7.92 | 7.54 | 7.83 | 7.47 | 7.99 | 7.71 | 7.18 | 6.93 | 7.75 | 7.41 | 7.34 | 7.02 |
| XPerT/CRAYON | 7.27 | 6.98 | 6.92 | 6.63 | 7.43 | 7.07 | 7.58 | 7.21 | 6.99 | 6.69 | 7.08 | 6.82 | 6.87 | 6.60 |
| LoraHub/LoraHub | 6.95 | 6.57 | 6.42 | 6.08 | 7.08 | 6.68 | 7.09 | 6.79 | 6.61 | 6.29 | 6.77 | 6.58 | 6.49 | 6.14 |
| **Ours: TRIAD-TS** | **8.34** | **7.98** | **8.80** | **8.51** | **8.63** | **8.29** | **8.87** | **8.61** | **8.07** | **7.79** | **8.66** | **8.39** | **8.18** | **7.92** |

Table 4: Per-criterion G-Eval scores (1–10 scale; higher is better) from two independent LLM judges. TRIAD-TS achieves the highest scores on all 7 criteria for both judges.

| Hard-OOD split | Method | Proceed | Fallback | Server | Abstain | Avg. G-Eval | Style-Drift% |
|---|---|---|---|---|---|---|---|
| Task-OOD | CRAYON-Hybrid | 26% | – | 74% | 0% | 7.18 | 19.5 |
| | TRIAD-TS | 9% | 24% | 0% | 67% | 8.05 | 6.2 |
| Style-OOD | CRAYON-Hybrid | 71% | – | 29% | 0% | 7.36 | 21.3 |
| | TRIAD-TS | 36% | 56% | 0% | 8% | 8.12 | 9.7 |
| Both-OOD | CRAYON-Hybrid | 12% | – | 88% | 0% | 6.95 | 28.4 |
| | TRIAD-TS | 2% | 15% | 0% | 83% | 7.92 | 5.8 |

Table 5: Hard-OOD comparison on ESConv holdout OOD subsets. Proceed: answer with in-pool adapters; Fallback: safe nearest in-pool style while keeping task; Server: CRAYON-style offload and output replacement; Abstain: ask targeted clarifications with minimal safe support. TRIAD-TS is more calibrated and achieves higher counseling quality with substantially lower style drift. Other baselines (CRAYON/CRAYON, XPerT/XPerT, etc.) are omitted because they lack abstention mechanisms and uniformly proceed (100%), providing no calibration insight.

| Variant | G-Eval ELO | Δ |
|---|---|---|
| **Ours: TRIAD-TS (full)** | **1162** | — |
| – remove ClinicalAgent | 1142 | −20 |
| – remove CommAgent | 1128 | −34 |
| – single-LLM router | 1135 | −27 |
| – deterministic-only | 1091 | −71 |

Table 6: Ablation of multi-agent routing on Standard test set. Removing CommAgent yields the largest drop, matching the strong dependence of counseling quality on style/ethics dimensions.

| Architecture | G-Eval ELO | Typical failure mode |
|---|---|---|
| Debate | 1139 | Over-deliberation; inconsistent intent under ambiguity. |
| Critic-Refine | 1131 | Style oscillation; polishing without re-grounding. |
| **Ours: TRIAD-TS** | **1162** | Best stability via deterministic integration. |

Table 7: Multi-agent architecture comparison on Standard test set. TRIAD-TS is best by decoupling intent inference (agents) from deterministic compatibility-aware composition.

not capture all therapeutic actions required in real-world practice. While quality-driven abstention improves reliability, abstaining varies by distribution: 8.2% on Standard in-distribution test set, but up to 67% on Task-OOD and 83% on Both-OOD (Table 5). High OOD abstention rates reflect principled uncertainty handling rather than system failure, but may frustrate users expecting instant responses across all query types. On-device processing provides strong privacy guarantees for model inference, though standard privacy-preserving techniques could further strengthen training procedures.

## References


American Psychological Association. 2017. Ethical principles of psychologists and code of conduct. https://www.apa.org/ethics/code. Accessed: 2024-01-05.

Jihwan Bang, Joonhyung Lee, Jaewoong Moon, Jaehong Choi, and Sung Ju Hwang. 2024. Crayon: Customized on-device LLM personalization via instant adapter blending. *arXiv preprint arXiv:2405.09818*.

C Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

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

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, and 1 others. 2023. MindAgent: Emergent gaming interaction with LLM-based multi-agent systems. *arXiv preprint arXiv:2310.02604*. Extended for therapeutic applications.

Clara E Hill. 2014. *Helping Skills: Facilitating Exploration, Insight, and Action*, 4th edition. American Psychological Association.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, and 1 others. 2023. MetaGPT: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. 2023. LoraHub: Efficient cross-task generalization via dynamic LoRA composition. *arXiv preprint arXiv:2307.13269*.

Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Siqi Liu, Dorottya Demszky, Joanna Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2024. PARTNER: A collaborative agent framework for peer mental health support. *Proceedings of ACL*, pages 4521–4538.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3469–3483.

Nicole Martinez-Martin, Henry T Greely, and Mildred K Cho. 2020. Privacy in the age of neuroscience: Implications for psychotherapy and counseling. *The American Journal of Bioethics*, 20(5):3–15.

William R Miller and Stephen Rollnick. 2012. *Motivational Interviewing: Helping People Change*, 3rd edition. Guilford Press.

John C Norcross. 2011. *Psychotherapy Relationships that Work: Evidence-Based Responsiveness*, 2nd edition. Oxford University Press.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of EACL*, pages 487–503.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Carl R Rogers. 1957. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 21(2):95–103.

Ashish Sharma, Kevin Lin, Ting-Hao Huang, and Heng Ji. 2023. Multi-agent systems for therapeutic conversations: Balancing empathy and clinical effectiveness. *arXiv preprint arXiv:2309.12847*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

John Torous and Matcheri Keshavan. 2020. The ethical use of mobile health technology in clinical psychiatry. *The Journal of Nervous and Mental Disease*, 208(1):1–3.

Zhiyuan Wang, Jeongyoon Kim, Pranav Madhyastha, and Lucia Specia. 2025. XPerT: Explainable personalization using model ensembles for on-device LLMs. *arXiv preprint arXiv:2501.04127*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2023. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. Mentalchat16k: A benchmark dataset for conversational mental health assistance. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5367–5378.

Kailai Yang, Shaoxiong Tian, Tianlin Zhang, Qianqian Xie, Kun Kuang, Sophia Wu, Bo Liang, and Xin Xie. 2024. MentalLLaMA: Interpretable mental health analysis with large language models. *arXiv preprint arXiv:2309.13567*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

# A Implementation Details

## A.1 Offline Pool Construction

**Task Embedding.**

$$\mathbf{q}_x = \mathcal{M}^{(16)}_{\Phi_0}(x). \tag{9}$$

Layer 16 is selected as it balances semantic richness (early layers encode surface features) with task-relevant abstraction (late layers specialize for generation).

**Orthogonal Basis Update Rule.**

$$\text{accept } \mathbf{v} \text{ iff } \cos(\mathbf{v}, \mathbf{v}_j) < \tau_{\text{orth}} \ \forall j. \tag{10}$$

We set $\tau_{\text{orth}} = 0.15$ to ensure style basis vectors are approximately orthogonal, reducing style interference during blending.

**Projection for Style Codes.**

$$z_{s,j} = \frac{\mathbf{V}_s \cdot \mathbf{v}_j}{\|\mathbf{v}_j\|}, \qquad \mathbf{z}_s = [z_{s,1}, \ldots, z_{s,K}]^\top. \tag{11}$$

**Low-rank Adapter Conversion.**

$$\mathcal{A}_s = \mathbf{B}_s\mathbf{A}_s, \qquad (\mathbf{B}_s, \mathbf{A}_s) = \text{SVD}_r(\text{reshape}(\mathbf{V}_s)). \tag{12}$$

## A.2 Online Signals and Server Payload

**Signal Extraction.**

$$\mathbf{e}_{\text{ind}} = \textsc{IndicatorExtract}(\mathbf{x}_{\text{client}}). \tag{13}$$

Indicator extraction uses POS tagging (spaCy) to identify emotional keywords (anxiety, stress), directive phrases (should, must), and question markers.

$$\mathbf{s}_{\text{style}} = \textsc{StyleSignal}(\mathbf{x}_{\text{client}}). \tag{14}$$

Style signals include detected emotional tone (valence, arousal), formality level, and user history preferences if available.

## A.3 Task-Style Compatibility Matrix

The compatibility matrix $\mathbf{C} \in [0, 1]^{7\times12}$ encodes clinical appropriateness between **the 7 trained task categories** and 12 therapeutic styles based on established counseling frameworks (Hill, 2014; Miller and Rollnick, 2012). Key entries:

- $C_{\text{Question,Empathic}} = 1.0$: Exploration benefits from warmth
- $C_{\text{Question,Confrontational}} = 0.3$: Challenge inappropriate during early exploration
- $C_{\text{Suggestions,Directive}} = 1.0$: Action stage allows guidance
- $C_{\text{Reflection,Mindful}} = 0.95$: Present-centered awareness enhances emotional articulation
- $C_{\text{Affirmation,Strength-Based}} = 1.0$: Natural alignment

## A.4 Full Training Details

**LoRA Configuration.** Each LoRA adapter uses rank $r = 16$ with scaling $\alpha = 32$. The scaling factor $\alpha/r = 2.0$ is intentionally set higher than the conventional $\alpha = r$ to amplify adapter contributions during soft mixture training (Eq. 2), ensuring that low-weight adapters ($\alpha_n \sim 0.2$) still meaningfully influence gradients. This prevents mode collapse where only the highest-weighted adapter receives updates.

**Training Procedure.** We train for $E = 10$ epochs with no early stopping. Validation loss plateaued consistently around epoch 8-9 in preliminary experiments across all task adapters. Final checkpoints use epoch 10 weights.

Training hyperparameters:

- Optimizer: AdamW
- Learning rate: $2 \times 10^{-4}$
- Betas: $(0.9, 0.999)$
- Weight decay: 0.01
- Batch size: 32 with gradient accumulation over 4 steps (effective batch size: 128)
- Gradient clipping: max norm = 1.0
- Warmup: 100 steps with cosine annealing
- Precision: FP16 (mixed precision)

**Hardware.** All experiments conducted on 4× NVIDIA A100 (40GB). Training time: 6 hours for task pool, 2 hours per style adapter (24 hours total for 12 styles).

# B Prompts and Data

## B.1 Complete List of 50 Probing Prompts

The following prompts were used to generate response pairs for style drift computation:

1. "I feel anxious about my upcoming presentation at work."
2. "My partner doesn't understand me anymore."
3. "I can't stop thinking about past mistakes."

4. “I’m having trouble sleeping lately.”
5. “I feel like I’m not good enough for this job.”
6. “My parents are putting too much pressure on me.”
7. “I’m worried about my children’s future.”
8. “I feel lonely even when I’m with people.”
9. “I can’t seem to make important decisions.”
10. “I’m struggling with my self-image.”
11. “Work stress is affecting my health.”
12. “I feel guilty about prioritizing my needs.”
13. “My relationships always seem to fail.”
14. “I’m afraid of disappointing others.”
15. “I can’t control my angry outbursts.”
16. “I feel stuck in my current life situation.”
17. “I’m having conflict with my teenage daughter.”
18. “I feel overwhelmed with all my responsibilities.”
19. “I’m questioning my life choices.”
20. “I can’t seem to trust people anymore.”
21. “I’m procrastinating on important tasks.”
22. “I feel disconnected from my emotions.”
23. “My grief is interfering with daily life.”
24. “I’m struggling with social anxiety.”
25. “I don’t know how to set healthy boundaries.”
26. “I feel hopeless about the future.”
27. “I’m having intrusive negative thoughts.”
28. “I can’t communicate effectively with my spouse.”
29. “I feel burnout from my career.”
30. “I’m dealing with a major life transition.”
31. “I feel ashamed about my past actions.”
32. “I can’t stop comparing myself to others.”
33. “I’m experiencing panic attacks.”
34. “I feel misunderstood by my family.”
35. “I’m having trouble concentrating.”
36. “I don’t feel motivated to do anything.”
37. “I’m dealing with chronic pain and its emotional toll.”
38. “I feel like I’m losing my identity.”
39. “I’m worried about my financial situation.”
40. “I can’t forgive someone who hurt me.”
41. “I’m struggling with perfectionism.”
42. “I feel resentful in my relationship.”
43. “I’m having trouble adapting to retirement.”
44. “I feel inadequate as a parent.”
45. “I’m dealing with workplace harassment.”
46. “I can’t move on from a past relationship.”
47. “I feel trapped in my current circumstances.”
48. “I’m experiencing culture shock after moving.”
49. “I don’t know how to manage my stress.”
50. “I’m questioning my spiritual beliefs.”

### B.2 Style Transformation Prompt

**Style Transformation**

```
You are an expert mental health
    counselor skilled in adapting
    communication styles while
    preserving therapeutic intent.

Original counseling response: "{
    original_response}"

Task: Rewrite this response in a {
    target_style} style while:
(1) maintaining the core therapeutic
    message and intent,
(2) preserving factual accuracy and
    appropriateness,
(3) ensuring cultural sensitivity and
     ethical boundaries,
(4) only adapting the communication
    style and delivery.
```

### B.3 Clinical Assessment Agent Prompt

**Clinical Assessment Agent**

```
System: You are a licensed clinical
    psychologist with 15+ years of
    experience in evidence-based
    treatment planning. Your
    expertise includes CBT, DBT, MI,
    psychoeducation, and supportive
    counseling.

Your task is to analyze client
    information and recommend
```

```
    appropriate therapeutic
    intervention priorities based on
    clinical presentation and
    treatment stage.

Important constraints:
- Do NOT provide medical diagnosis
- Do NOT invent facts not present in
    the client context
- If information is missing, make
    conservative assumptions
- Output must be valid JSON only

User: Analyze the following client
    context and assign priority
    scores (0.0--1.0) for each
    therapeutic approach.

CLIENT CONTEXT:
- Presenting Concerns: {concerns}
- Session History: {history}
- Current Symptoms: {symptoms}
- Treatment Goals: {goals}

Rate each approach (0.0--1.0):
1) CBT: cognitive restructuring,
    behavioral activation
2) DBT: emotion regulation, distress
    tolerance
3) MI: explore ambivalence,
    strengthen motivation
4) Psychoeducation: explain symptoms/
    skills
5) Supportive: validation,
    encouragement, stabilization

Output format (JSON only):
{"CBT": <float>, "DBT": <float>, "MI
    ": <float>, "Psychoeducation": <
    float>, "Supportive": <float>}
```

### B.4 Communication Strategy Agent Prompt

**Communication Strategy Agent**

```
System: You are an expert in
    therapeutic communication styles
    with deep knowledge of person-
    centered therapy, evidence-based
    communication strategies, and
    alliance-sensitive counseling.

Your task is to recommend how the
    assistant should communicate (
    style/stance), given the client
    context and relationship stage.

User: Given the client context below,
     rate the appropriateness
    (0.0--1.0) of each communication
    style.

CLIENT CONTEXT:
- Current Emotional State: {emotion}
- Therapeutic Relationship Stage: {
    stage}
- Session Goals: {goals}

Rate each style (0.0--1.0):
1) empathetic: emotional attunement,
    warmth
2) directive: clear structure,
    concrete steps
3) validating: normalize and affirm
    emotions
4) reflective: deepen insight and
    self-awareness
5) confrontational: gentle challenge
    (only if alliance strong)
6) supportive: encouragement,
    strengths-focus
7) psychoeducational: explain
    concepts/skills
8) motivational: evoke change talk,
    autonomy-supportive

Output format (JSON only):
{"empathetic": <float>, "directive":
    <float>, "validating": <float>, "
    reflective": <float>, "
    confrontational": <float>, "
    supportive": <float>, "
    psychoeducational": <float>, "
    motivational": <float>}
```

**Style Mapping:** The agent outputs 8 descriptors which map to our 12-style taxonomy via semantic similarity (see Appendix B.5). Direct mappings include Empathic ↔ empathetic (0.95), Directive ↔ directive (1.0), Psychoeducational ↔ psychoeducational (1.0). Styles not directly covered (Mindful, Narrative) receive blended scores from related descriptors.

### B.5 Semantic Matching for Style Mapping

The Communication Agent outputs 8 communication **descriptors**: empathetic, directive, validating, reflective, confrontational, supportive, psychoeducational, motivational. These are **not** styles themselves but dimensional outputs that map to our 12 therapeutic styles. We use Sentence-BERT cosine similarity for mapping:

Table 8: Mapping from 8 agent descriptors to 12 therapeutic styles.

| Style (12) | Agent Descriptor(s) | Sim. |
|---|---|---|
| EMPATHIC | empathetic | 0.95 |
| SOCRATIC | reflective + confrontational | 0.72 |
| DIRECTIVE | directive | 1.00 |
| MOTIVATIONAL | motivational | 1.00 |
| PSYCHOEDUCATIONAL | psychoeducational | 1.00 |
| MINDFUL | validating + reflective | 0.78 |
| SOLUTION-FOCUSED | directive + supportive | 0.81 |
| NARRATIVE | reflective + empathetic | 0.76 |
| STRENGTH-BASED | supportive + validating | 0.88 |
| PERSON-CENTERED | empathetic + validating | 0.92 |
| COLLABORATIVE | supportive + reflective | 0.79 |
| INTEGRATIVE | blend all descriptors | 0.65 |

**Important distinction:** The 8 descriptors (validating, reflective, supportive, etc.) are **agent outputs**, not therapeutic styles. Multiple descriptors combine to form each of the 12 styles listed in Table 12.

# C Dataset Construction Details

## C.1 Source Dataset Statistics

Table 9: Detailed source dataset characteristics and usage.

| Dataset | Size | Avg. Length | Details |
|---|---|---|---|
| *ESConv* | | | |
| Dialogues | 1,053 | 29.8 uttr/dial | 80/20 split |
| Supporter Uttr. | 14,855 | 20.2 words | 8 strategies |
| *MentalChat16K* | | | |
| Interview | 9,775 | 152.9 words | |
| Synthetic (GPT-3.5) | 6,338 | 237.6 words | |
| *Combined* | *16,113* | *185.4 words* | 100% train |

Avg. Length: ESConv = supporter output only; MentalChat16K = (input+output)/2 per QA pair. ESConv has human-annotated task labels (Fleiss' $\kappa = 0.79$); MentalChat16K labeled via GPT-4-turbo with 20-shot prompting, verified by clinicians ($\kappa = 0.71$). ESConv split 80% train / 20% test; MentalChat16K used entirely for training (no test labels available).

## C.2 Task Category Overview

ESConv (Liu et al., 2021) provides human annotations for eight support strategies grounded in Hill's three-stage Helping Skills Theory (Hill, 2014).

## C.3 Task Distribution After Labeling

# D Therapeutic Style Taxonomy

# E Therapeutic Style Taxonomy

# F Task Category Definitions

Following Hill's three-stage Helping Skills model (Hill, 2014), the ESConv framework (Liu et al., 2021) defines eight support strategies for therapeutic conversation.

Table 10: ESConv's eight support strategies mapped to Hill's three-stage helping model.

| Task Category | Therapeutic Goal | Stage |
|---|---|---|
| Question | Open inquiry to understand client's situation | Exploration |
| Restatement or Paraphrasing | Clarify client's situation through concise rephrasing | Comforting |
| Reflection of Feelings | Articulate and validate emotional experience | Comforting |
| Self-disclosure | Build alliance through shared experience | Comforting |
| Affirmation and Reassurance | Validate strengths and provide encouragement | Comforting |
| Providing Suggestions | Offer actionable problem-solving advice | Action |
| Information | Provide psychoeducation and resources | Action |
| Others | Maintain rapport; administrative tasks | – |

## F.1 Exploration Stage

**Question** Asking for information related to the problem to help the help-seeker articulate the issues they face. Open-ended questions are preferred.

- *Examples:* "Can you tell me more about what happened?", "How has this been affecting you?"

## F.2 Comforting Stage

**Restatement or Paraphrasing** A simple, concise rephrasing of the help-seeker's statements.

- *Examples:* "So you're saying that...", "It sounds like your main concern is..."

**Reflection of Feelings** Articulate and describe the help-seeker's feelings to demonstrate deep emotional understanding.

Table 11: Complete task category distribution across datasets (8 categories).

| Task Category | ESConv | | MentalChat16K | |
|---|---|---|---|---|
| | N | % | N | % |
| Question | 3,109 | 20.9 | 2,747 | 17.0 |
| Restatement/Paraphrasing | 883 | 5.9 | 955 | 5.9 |
| Reflection of Feelings | 1,156 | 7.8 | 986 | 6.1 |
| Self-disclosure | 1,396 | 9.4 | 1,521 | 9.4 |
| Affirmation/Reassurance | 2,388 | 16.1 | 2,678 | 16.6 |
| Providing Suggestions | 2,323 | 15.6 | 2,424 | 15.0 |
| Information | 904 | 6.1 | 809 | 5.0 |
| Others | 2,696 | 18.1 | 3,993 | 24.8 |
| **Total** | **14,855** | **100** | **16,113** | **100** |

Table 12: Complete therapeutic style taxonomy with theoretical foundations.

| Style | Key Characteristics | Foundation (%) |
|---|---|---|
| EMPATHIC | Reflective, emotionally validating, client-centered | Rogers (1957) (12) |
| SOCRATIC | Questioning, collaborative exploration | Beck (2011) (10) |
| DIRECTIVE | Clear guidance, structured interventions | Linehan (1993) (9) |
| MOTIVATIONAL | Autonomy-affirming, intrinsic motivation | Miller & Rollnick (2012) (9) |
| PSYCHO-EDUCATIONAL | Explains conditions and coping strategies | Anderson et al. (1986) (8) |
| MINDFUL | Present-centered, non-judgmental | Kabat-Zinn (2003) (8) |
| SOLUTION-FOCUSED | Goal-oriented, pragmatic problem-solving | de Jong & Berg (2013) (8) |
| NARRATIVE | Story-based, meaning-making | White & Epston (1990) (8) |
| STRENGTH-BASED | Resource-focused, empowering | Saleebey (2013) (8) |
| PERSON-CENTERED | Non-directive, unconditional regard | Rogers (1957) (8) |
| COLLABORATIVE | Partnership-oriented | Anderson (1997) (6) |
| INTEGRATIVE | Multi-modal, flexible | Norcross & Goldfried (2011) (6) |

- *Examples:* "It sounds like you're feeling overwhelmed", "I can hear the frustration in your voice"

**Self-disclosure** Divulge similar experiences that you have had to express empathy and build therapeutic alliance.

- *Examples:* "I've also struggled with this when...", "I know how that feels because I experienced..."

**Affirmation and Reassurance** Affirm the help-seeker's strengths and provide reassurance and encouragement.

- *Examples:* "You're showing real strength by reaching out", "That's a completely normal reaction"

### F.3 Action Stage

**Providing Suggestions** Provide suggestions about how to change while respecting the help-seeker's autonomy.

- *Examples:* "You might consider trying...", "Have you thought about...", "One option could be..."

**Information** Provide useful information, data, facts, resources, or answer questions.

- *Examples:* "Anxiety often manifests as physical symptoms like...", "Research shows that..."

### F.4 Other

**Others** Exchange pleasantries and other support strategies maintaining rapport.

- *Examples:* "Thank you for sharing that with me", "I'm glad we could talk today"

## G Hard-OOD Examples

### G.1 Task-OOD: Other Category

Examples labeled as OTHER include:

- **Crisis intervention:** "I'm having thoughts of self-harm and don't know what to do."
- **Grief counseling:** "My mother passed away last month and I can't stop crying."
- **Relationship termination:** "I need to break up with my partner but I'm scared."
- **Administrative:** "Can we reschedule our session to next Tuesday?"
- **Pleasantries:** "Thank you so much for your help today."

These contexts require either abstention (crisis), specialized routing (grief), or minimal support (pleasantries) rather than standard task adapter selection.

### G.2 Style-OOD: Culturally-Specific Communication

**East Asian (Filial Piety-Oriented).**

- Emphasis on family harmony and parental respect
- Indirect communication about family conflicts
- Example: "How can I honor my parents while also setting boundaries for myself?"

**African (Ubuntu-Based).**

- Collective framing: "we" vs. "I"
- Community-oriented solutions
- Example: "Our family is struggling, and I need to find a way to help everyone."

**Grief Expression Norms.**

- Irish wake tradition: communal celebration of life
- Buddhist acceptance: impermanence and letting go
- Jewish sitting shiva: structured mourning period

### G.3 Style-OOD: Technical Communication

**Medical Jargon-Heavy.** Client with professional healthcare background requesting clinical terminology:

- "I'm experiencing anhedonia and avolition consistent with major depressive disorder."
- "Can you explain the neurobiological mechanisms of SSRIs?"

**Legal-Procedural.** Client navigating custody or workplace disputes:

- "I need to document this for my attorney. What should I include?"
- "How do I phrase this in a way that protects my legal rights?"

## H Qualitative Analysis Examples

### H.1 Profile-Conditioned Personalization

**User Profile:**

- Preference: Concise, structured guidance
- Aversion: Overly cheerful or effusive tone
- History: Responds well to concrete action steps

**Query:** "I'm feeling overwhelmed with work deadlines."

**TRIAD-TS Response:**

> *It sounds like you're under a lot of pressure right now. That's understandable. Here are two things you might try: (1) Break your largest task into 15-minute chunks, (2) Identify one thing you can delegate or postpone. What feels most doable?*

**Analysis:**

- Task: VALIDATION (brief acknowledgment) + SUGGESTIONS (concrete steps)
- Style: 60% DIRECTIVE + 40% EMPATHIC
- Length: 3 sentences (concise as preferred)
- Tone: Warm but not effusive ("That's understandable" vs. "I'm so sorry!")

### H.2 Ambiguity Handling

**Query:** "Help me with my project."

**TRIAD-TS Response:**

> *I'd like to help you with your project. Can you tell me a bit more? Is this a work project, a personal goal, or something else? What specific aspect feels challenging right now?*

**Baseline (CRAYON) Response:**

> *Projects can feel overwhelming! I recommend breaking it into smaller steps and celebrating each milestone. You've got this!*

**Analysis:**

- TRIAD-TS correctly abstains from guessing, asks QUESTION for clarification
- Baseline over-commits to SUGGESTIONS without context, risks mis-framing

### H.3 Task-Style Conflict Resolution

**Query:** "I need coping strategies for my anxiety, but I don't want to be told what to do."

**Detected Intent:**

- Task: PROVIDING SUGGESTIONS ("coping strategies")
- Style: High REFLECTIVE, low DIRECTIVE ("don't want to be told")
- Conflict: $C_{\text{Suggestions,Reflective}} = 0.6$ (moderate compatibility)

**TRIAD-TS Response:**

> *It sounds like you want to explore options without feeling pressured into a specific approach. Some people find deep breathing helpful, others prefer grounding techniques like the 5-4-3-2-1 method. What's your sense—would either of these feel worth trying, or are you looking for something else?*

**Analysis:**

- Compatibility enforcement: Reweighted 70% REFLECTIVE + 30% DIRECTIVE
- Presents suggestions as collaborative exploration ("some people find")
- Ends with reflective question ("your sense") to maintain autonomy

# I Detailed Experimental Results

This appendix provides comprehensive breakdowns of experimental results referenced in Section 6, including per-criterion evaluation scores and model robustness analysis.

**Key Observations. Interpersonal dimensions show largest gains.** TRIAD-TS achieves its most substantial improvements over baselines on *Empathy* and *Boundaries*, the exact dimensions where naive adapter composition typically fails:

- **Empathy**: TRIAD-TS achieves 8.80 (Claude) and 8.51 (Gemini), compared to the strongest baseline XPerT/XPerT at 7.92/7.54. This represents gains of +0.88 and +0.97 respectively, the largest absolute improvements among all criteria.
- **Boundaries**: TRIAD-TS achieves 8.66/8.39 vs. XPerT/XPerT's 7.75/7.41, with gains of +0.91/+0.98. Boundary violations occur when responses overstep professional limits (e.g., diagnosing, prescribing) or fail to maintain therapeutic frame.
- **Non-judgmentalness**: TRIAD-TS achieves 8.87/8.61 vs. XPerT/XPerT's 7.99/7.71 (+0.88/+0.90). This dimension captures whether responses avoid implicit criticism or moralizing.

**Clinical utility maintained.** Despite strong focus on interpersonal appropriateness, TRIAD-TS also improves on task-oriented dimensions:

- **Listen**: 8.34/7.98 vs. 7.63/7.28 for XPerT (+0.71/+0.70), indicating responses demonstrate accurate understanding of client concerns.
- **Clarity**: 8.07/7.79 vs. 7.18/6.93 for XPerT (+0.89/+0.86), showing that safety and ethics gains do not come at the cost of comprehensibility or actionability.
- **Safety**: 8.63/8.29 vs. 7.83/7.47 for XPerT (+0.80/+0.82), validating that compatibility enforcement and quality-driven abstention successfully prevent harmful outputs.

**Holistic quality reflects balanced optimization.** The *Holistic* criterion asks judges to rate overall counseling effectiveness considering all dimensions. TRIAD-TS achieves 8.18/7.92 vs. XPerT's 7.34/7.02 (+0.84/+0.90), confirming that gains are not narrow artifacts but reflect genuinely improved therapeutic response quality.

**Judge agreement.** Both judges consistently rank TRIAD-TS highest across all criteria, with per-criterion Pearson correlations ranging from $r = 0.87$ (Clarity) to $r = 0.94$ (Empathy), indicating robust evaluation reliability.

## I.1 Model Robustness and Deployment Sweep

Table 13 reports TRIAD-TS performance across different on-device SLLMs and server-side routing models. This comprehensive sweep validates that the framework's core design principles—orthogonal task-style factorization, multi-agent coordination, deterministic compatibility enforcement—are robust to model substitution and applicable across realistic deployment configurations.

**Experimental Setup. On-device SLLMs.** We evaluate five models spanning 0.5B to 2B parameters, representing realistic mobile deployment constraints:

- **Qwen2.5-1.5B**: Alibaba's efficient multilingual model (1.5B params)
- **Gemma-2-2B**: Google's instruction-tuned compact model (2B params)
- **Phi-3.5-mini**: Microsoft's small language model optimized for reasoning (3.8B params, quantized to 2B equivalent)
- **Llama-3.2-1B**: Meta's latest compact Llama variant (1B params)
- **SmolLM**: HuggingFace's sub-1B parameter model (0.5B params)

**Server routers.** We evaluate five high-capability models as server-side multi-agent coordinators:

- GPT-4-turbo (OpenAI, 2024)
- GPT-4o (OpenAI, optimized for speed)
- Claude-3-Opus (Anthropic, highest-capability variant)
- Gemini-1.5-Pro (Google, long-context specialist)
- Llama-3.1-70B (Meta, open-source alternative)

| On-device SLLM | GPT-4-turbo | GPT-4o | Claude-3-Opus | Gemini-1.5-Pro | Llama-3.1-70B |
|---|---|---|---|---|---|
| Qwen2.5-1.5B | 1162 | **1165** | 1157 | 1152 | 1149 |
| Gemma-2-2B | 1154 | **1157** | 1149 | 1143 | 1140 |
| Phi-3.5-mini | 1141 | **1144** | 1137 | 1131 | 1127 |
| Llama-3.2-1B | 1127 | **1130** | 1122 | 1116 | 1112 |
| SmolLM (sub-1B) | 1110 | **1113** | 1106 | 1100 | 1097 |

Table 13: TRIAD-TS G-Eval ELO ratings across 25 deployment configurations (5 on-device SLLMs $\times$ 5 server routers) on Standard test set (N=1,847). The best-performing router for each on-device SLLM is bolded. Main experiments (Tables 2–3) use Llama-2-7B-Chat as $\mathcal{M}_{\Phi_0}$ to establish upper-bound performance before downscaling; this sweep focuses on mobile-realistic models (0.5–2B parameters) suitable for actual on-device deployment. Referenced in Section 6.

Each configuration runs on the same test set (N=1,847) with identical hyperparameters ($K_{\text{task}} = 3$, $\tau = 0.1$, $\lambda = 0.5$, $\omega = 0.6$).

**Key Findings.** **On-device SLLM quality is the primary performance determinant.** The vertical span in Table 13 (52–55 ELO points across on-device models) substantially exceeds the horizontal span (5–7 ELO points across routers). This indicates that *on-device generation capacity sets the ceiling for attainable counseling quality*. Even with optimal routing, a weak on-device SLLM cannot produce high-quality therapeutic responses.

**Router choice has consistent but smaller impact.** Across all on-device SLLMs, GPT-4o consistently performs best as the server-side router (±2–3 ELO advantage over other routers). This suggests GPT-4o's instruction-following and structured output capabilities make it particularly well-suited for multi-agent coordination tasks. However, the margin is modest: even the weakest router (Llama-3.1-70B) remains within 13–16 ELO of the best, indicating the framework's structural advantages (coordinated routing, compatibility enforcement, quality-driven abstention) persist regardless of specific router choice.

**Framework design principles transfer across models.** All 25 configurations maintain the core TRIAD-TS advantages over baselines:

- Best joint task-style routing accuracy (verified on subset of 300 examples per config)
- Lowest style drift rates (8–12% across all configs vs. 18–24% for CRAYON baselines)
- Superior OOD calibration (abstention rates 5–10% on Standard, 60–70% on Task-OOD)

This robustness validates that TRIAD-TS's benefits arise from principled architectural design rather than fortuitous hyperparameter tuning for specific models.

**Practical deployment recommendations.** For production systems:

- *Prioritize on-device SLLM quality*: Invest in Qwen2.5-1.5B or Gemma-2-2B over smaller alternatives
- *Router substitution is safe*: If GPT-4o is unavailable, Claude-3-Opus or GPT-4-turbo provide near-equivalent performance
- *Open-source option viable*: Llama-3.1-70B as router sacrifices only 13–16 ELO, making fully open-source deployment feasible

**Comparison to Main Experiments.** Main experiments (Tables 2–3) use Llama-2-7B-Chat (7B parameters) as $\mathcal{M}_{\Phi_0}$ to establish a **reference baseline** with sufficient capacity for comprehensive evaluation before considering mobile deployment constraints. The 7B model achieves ELO $1162 \pm 8$ with GPT-4o routing.

**Mobile-realistic models (0.5–2B parameters) span a 52-point ELO range**: Qwen2.5-1.5B achieves 1165 (best), while SmolLM achieves 1113 (weakest). Notably, **the best mobile configuration (Qwen2.5-1.5B) slightly outperforms the 7B reference (+3 ELO)**, likely due to Qwen's specialized instruction-tuning and more recent training data. This demonstrates that model selection and architecture matter more than raw parameter count for therapeutic response quality within the TRIAD-TS framework.

The 52-point span across mobile models quantifies the inherent tradeoff between extreme on-device constraints (sub-1B) and realistic mobile deployment (1–2B parameters), validating that adequate on-device capacity is essential for maintaining counseling quality.