Vibe: Hiring Round Task

Problem Specification

Objective

Fine–tune an open–weights LLM to behave as an *empathetic, best–friend* chatbot and demonstrate a measurable improvement over the base model on an empathy–focused benchmark (EQ–Bench 3). You will implement a multi–objective supervised fine–tuning (SFT) loss with auxiliary heads for emotion and support–strategy prediction, then (optionally) apply a lightweight preference–alignment stage (DPO). Report quantitative gains and concise qualitative analysis.

Setup (Compute & Base)

Two tracks are acceptable:
- **Low–compute (recommended)**: 20B GPT-OSS/30B Qwen3 model with QLoRA/PEFT.
- **High–compute (optional)**: 70B Llama–class/ 120 with LoRA (no full–precision SFT).

Data

Curate a multi–task mixture (train/dev splits):
- **EmpatheticDialogues**: user–assistant turns with empathy signals.
- **ESConv**: emotional support conversations with *strategy* labels.
- **GoEmotions**: single–utterance emotion labels for an auxiliary classifier head.

Use temperature–based sampling to mix datasets without dominance:

$$p_i \;=\; \frac{n_i^{\alpha}}{\sum_j n_j^{\alpha}}, \qquad \alpha \in (0,1],$$

where n_i is the number of training examples from dataset i.

Model & Supervision

Let the base decoder produce token logits for assistant turns (LM head). Add (i) an **emotion head** that classifies the user's latest turn, (ii) a **strategy head** that classifies the intended support strategy for the next assistant turn, and (iii) an optional **valence/arousal** regression head.

Primary LM loss. Given input context x and target assistant tokens $y_{1:T}$,

$$\mathcal{L}_{\text{NLL}} \;=\; -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, x).$$

Auxiliary classification/regression. For emotion labels $e \in \{1,\dots,K\}$ and strategy labels $s \in \{1,\dots,S\}$,

$$\mathcal{L}_{\text{emo}} \;=\; -\sum_{k=1}^{K} \mathbf{1}[e=k]\;\log q_\theta^{\text{emo}}(k \mid x), \quad \mathcal{L}_{\text{strat}} \;=\; -\sum_{j=1}^{S} \mathbf{1}[s=j]\;\log q_\theta^{\text{strat}}(j \mid x),$$

and for (optional) continuous targets (v, a) (valence, arousal),

$$\mathcal{L}_{\text{val}} \;=\; \|\hat{v}_\theta(x) - v\|_2^2 \;+\; \|\hat{a}_\theta(x) - a\|_2^2.$$

Safety regularization (teacher KL). Distill from a rules–prompted *safety teacher* with logits z_T; let z_θ be student logits. Using temperature $\tau > 0$ and softmax $\sigma(\cdot)$,

$$\mathcal{L}_{\text{safe}} \;=\; \mathrm{KL}\Big(\sigma\big(\tfrac{z_T}{\tau}\big) \;\Big\|\; \sigma\big(\tfrac{z_\theta}{\tau}\big)\Big)\,.$$

Total multi–objective SFT loss.

$$\boxed{\mathcal{L}_{\text{SFT}} \;=\; \lambda_{\text{LM}}\,\mathcal{L}_{\text{NLL}} \;+\; \lambda_{\text{emo}}\,\mathcal{L}_{\text{emo}} \;+\; \lambda_{\text{strat}}\,\mathcal{L}_{\text{strat}} \;+\; \lambda_{\text{val}}\,\mathcal{L}_{\text{val}} \;+\; \lambda_{\text{safe}}\,\mathcal{L}_{\text{safe}}}$$

Tune λ weights on the dev set; provide chosen values in your report.

Preference Alignment (Optional but Rewarded)

Apply Direct Preference Optimization (DPO) on pairs $\big((x, y^+), (x, y^-)\big)$ where y^+ is judged more empathetic than y^-. With inverse–temperature β and a fixed reference policy π_{ref},

$$\mathcal{L}_{\text{DPO}} \;=\; -\log \sigma\Big(\beta\big[\log \pi_\theta(y^+ \mid x) - \log \pi_\theta(y^- \mid x) - \big(\log \pi_{\text{ref}}(y^+ \mid x) - \log \pi_{\text{ref}}(y^- \mid x)\big)\big]\Big).$$

If you add safety constraints, include a penalty $\lambda_{\text{viol}} \cdot \mathbf{1}[\text{violation}]$.

Decoding Policy (Inference)

Use short *style tokens* (e.g., `<tone:warm><persona:best_friend>`). Implement a two–step controller: internally generate a one–line reflection (not shown to users) to ensure the reply includes (i) acknowledgment, (ii) feeling–naming, (iii) a gentle follow–up question; if a safety rule is triggered, re–decode with stronger penalties on directive/advice tokens.

Evaluation & Deliverables

1. **Primary metric**: EQ–Bench 3 score (report raw and normalized/Elo). Compare *Base* vs *SFT* vs *SFT+DPO*.
2. **Ablations** (at least two): remove emotion head; remove strategy head; no safety KL; no DPO.
3. **Qualitative**: 3–5 side–by–side conversations showing wins/failures; brief error taxonomy.
4. **Safety sheet**: 3 red–team prompts with expected safe behavior and your model's outputs.
5. **Reproducibility**: config, hyperparameters, data cards, and training logs (loss curves).

Constraints & Scoring

Constraints: PEFT/QLoRA required; context window $\leq$ base model default; no training on proprietary/private data.
Scoring (guideline): 30% modeling & losses, 25% evaluation quality, 20% safety, 15% engineering hygiene, 10% write–up clarity.

Stretch ideas (optional): persona contrastive loss; emotion rebalancing; memory policy (no history on turn 1; 20–token feeling summary from turn 2); lightweight judge consistency check (GoEmotions F1).