Model Card — Haven Conversation Model

haven-3b-qlora (safety-patched) · Siddartha Darisi · July 2026

Model Details

Field	Value
Base model	Llama 3.2 3B Instruct (4-bit NF4 quantized)
Adaptation	QLoRA — LoRA r=16, alpha=32, all attention + MLP projections (24.3M trainable, 0.75%)
Continued training	Two safety-patch rounds on the same adapter (targeted crisis/boundary set + replay buffer)
Companion models	DistilRoBERTa 28-emotion classifier (tuned thresholds); all-MiniLM-L6-v2 retrieval embedder
Training hardware	Single RTX 3070 Laptop GPU, 8 GB VRAM (chunked cross-entropy to fit)
License / lineage	Llama 3.2 Community License; adapters trained on research-released public datasets

Intended Use

Supportive, empathetic conversation in a private, fully local wellness companion: reflective listening, gentle questioning, evidence-based coping suggestions grounded in retrieved public-health passages, mood awareness, and reliable referral to crisis resources. Deployed only inside the Haven application, whose server-side safety layer is considered part of the model system.

Out of Scope

Diagnosis of any condition; medication advice of any kind; crisis counseling as a replacement for human services; use by or marketing to individuals in acute crisis; any deployment without the accompanying safety layer. The model states plainly that it is an AI companion, not a licensed therapist.

Training Data

Source	Size (train)	Contribution
EmpatheticDialogues (Rashkin et al. 2019)	8,000 convs (capped)	conversational warmth across 32 emotions
ESConv (Liu et al. 2021)	910 convs	support-strategy structure (questions, reflection, affirmation)
CounselChat-style corpus	1,809 pairs (deduped)	therapist-written vocabulary and boundaries
Hand-written safety set	26 unique examples	crisis + boundary gold responses (oversampled)
GoEmotions (classifier only)	43,410 texts	28-label emotion supervision

Evaluation & Metrics

Evaluation	Result
Held-out validation perplexity	base 32.59 → fine-tuned 9.79 (695 unseen conversations)
Crisis suite (8 prompts, greedy)	model 75% vs base 62.5%; application 100% (guaranteed)
False-positive controls (6 prompts)	100% — figures of speech never trigger the crisis script
Boundary suite (4 prompts)	model 75% (base parity); application 100% via referral footer
Emotion classifier (GoEmotions test)	micro-F1 0.563 / macro-F1 0.433 — beats TF-IDF baseline on both

Known Failure Modes & Mitigations

Documented honestly because they shaped the system: (1) fine-tuning drift — the first adapter regressed crisis handling to 25% before targeted patching; (2) short-reply habit — the chatty training corpus occasionally produces terse responses to nuanced messages; (3) comprehension limits at 3B scale — conditional phrasing ("I would feel better if…") can be misread; (4) sampling variance — worst-case sampled outputs can omit resources the greedy output includes. Mitigations: safety-patch training, attentive-listening system prompt, deterministic evaluation, and above all the server-side guarantee layer, which makes application-level safety independent of model behavior.

Ethical Considerations

Mental-health conversation is the most sensitive data a person can produce; Haven’s answer is architectural: no network egress for any user content, storage in one local SQLite file the user can inspect and erase in-app, and voice processed entirely on-device. The system is explicit about being an AI, refuses clinical roles, and surfaces real crisis resources (988, Crisis Text Line 741741, 911, NDVH 1-800-799-7233, Childhelp 1-800-422-4453). It has not been clinically validated and must not substitute for professional care.