# Haven

## Technical Report & System Design

A fully local AI therapy companion: QLoRA fine-tuned Llama, a trained emotion classifier, retrieval-grounded answers, local voice, and a defense-in-depth safety system — every step of the machine-learning lifecycle on one 8 GB consumer GPU.

Siddartha Darisi · July 2026
Live artifact: siddarthadarisi.github.io/haven-ml-pipeline · Code: github.com/SiddarthaDarisi/haven-companion

# 1 · Executive Summary

Haven is a private mental-wellness companion whose every model runs on the author's own machine — an RTX 3070 laptop GPU with 8 GB of VRAM. The system combines a Llama 3.2 3B Instruct model fine-tuned with QLoRA on 10,752 real emotional-support conversations, a DistilRoBERTa classifier scoring 28 emotions per message, retrieval over 36 public-domain clinical publications, fully local speech-to-text and neural text-to-speech, and a four-layer safety system whose guarantees do not depend on model behavior. Headline results: held-out validation perplexity improved from 32.6 (base) to 9.79 (fine-tuned); the emotion classifier beats a TF-IDF baseline on both micro-F1 (0.563 vs 0.460) and macro-F1 (0.433 vs 0.411); and the application passes 100% of an 18-prompt crisis/control/boundary safety suite by construction, with the patched model itself beating the base model on crisis handling (75% vs 62.5%, deterministic decoding).

# 2 · System Architecture

Three local processes cooperate: a Next.js front end, a FastAPI model server owning all five models, and SQLite persistence. No conversation data — text or audio — leaves the machine.

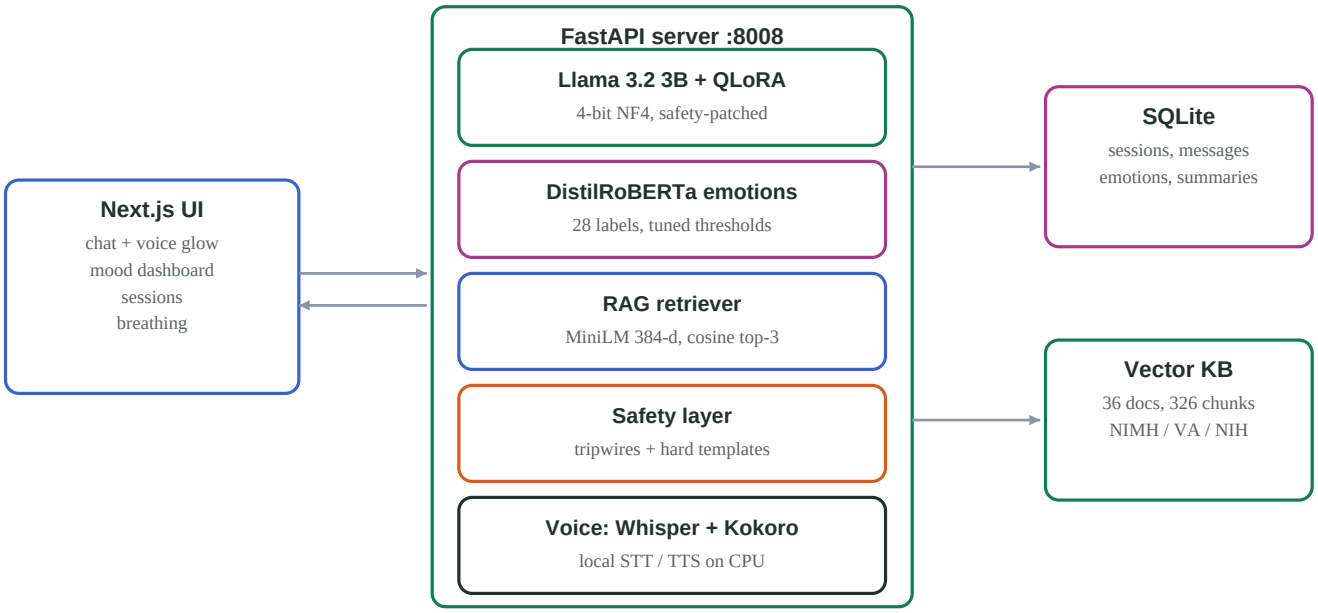


Figure 1 — system architecture. All components on one machine; zero network egress for user data.

# 3 · The Process We Followed

The build deliberately walked all ten steps of the machine-learning lifecycle, in order, with the artifact of each step feeding the next:

| # | Step | Key output |
|---|---|---|
| 1 | Problem definition | Success criteria fixed before any data: beat base perplexity; beat TF-IDF macro-F1; beat base on crisis suite with zero control regressions |
| 2 | Data collection | 4 corpora: EmpatheticDialogues, ESConv (strategy-annotated), CounselChat, hand-written safety set; GoEmotions for the classifier |
| 3 | Cleaning & preparation | 10,752 train / 695 val / 696 test conversations; 1,474 exact duplicates removed BEFORE splitting (leakage prevention) |

| # | Step | Key output |
|---|---|---|
| 4 | Exploratory analysis | Class imbalance (neutral 14,219 : grief 77) dictated macro-F1; strategy counts showed listening moves outnumber advice |
| 5 | Feature engineering | Assistant-only loss masks; conversation windowing; greedy packing into fixed blocks; 28-dim multi-hot targets |
| 6 | Model selection | Baseline-first discipline; iteration speed treated as a selection criterion (3B now, 8B background) |
| 7 | Training | Three tracks (classifier, 3B QLoRA + safety patches, 8B QLoRA) — detailed in §5 |
| 8 | Evaluation | Perplexity, classifier F1 vs baseline, 3-level safety suite with deterministic decoding |
| 9 | Deployment | FastAPI + Next.js + SQLite app with voice, RAG citations, mood tracking, session memory |
| 10 | Ethics & limits | Defense-in-depth safety; explicit non-clinical boundaries; honest limitation statements |

# 4 · How We Got Here — the Build Journey

The path was not linear, and the detours produced the report's most valuable engineering lessons. Chronology of the major events:

| Event | What happened | Lesson / fix |
|---|---|---|
| Data pipeline | Four raw formats unified into chat-message form; 42% of the counseling corpus turned out to be exact duplicates | Dedupe before splitting or leak test answers into training |
| Classifier v1 | DistilRoBERTa won micro-F1 (.596) but LOST macro-F1 (.395 vs .411) to the class-weighted baseline: rare emotions scored F1 = 0 at a flat 0.30 threshold | Imbalanced multi-label problems are won at the decision threshold: per-class tuning on validation lifted macro-F1 to .433 |
| 8B attempt #1-3 | Llama 3.1 8B QLoRA thrashed at seq 1024, 768, then 512: 100% GPU util at idle power, zero steps — VRAM silently spilling to system RAM | The 128k-token vocabulary makes fp32 loss logits a ~1.5 GB spike; wrote a chunked cross-entropy trainer (128-token slices under recomputation) so the tensor never materializes |
| 8B economics | With chunked CE the 8B finally trained — at 112 s/step, a ~40-hour run | Iteration speed is a model-selection criterion: pivoted to 3B for same-day results, queued 8B as the background quality ceiling |
| Grad-accum bug | First 3B run logged loss ~41 instead of ~2.5 | Losses summed across 16 accumulation micro-batches without normalization; every update hit the gradient clip. Fixed via num_items_in_batch normalization |
| Hardware reality | Two mid-run shutdowns: one kernel-power crash, one Windows sleep killing CUDA | Power-plan management is part of ML ops; 88°C thermal throttling stretched 35 s steps to ~65 s |
| Safety drift | The fine-tune tripled fluency but REGRESSED crisis handling 75% → 25%: 8,000 chatty conversations swamped 33 safety examples ("goodbye letters" → "oh wow thats pretty deep.") | Evaluation earned its keep by failing. Targeted safety-patch training (26 examples, heavy weight, replay buffer) + deterministic re-eval: patched model beats base 75% vs 62.5% |
| Sampling noise | Three sampled eval runs gave contradictory pass rates on 8-prompt suites | Small-suite metrics need deterministic (greedy) decoding to be reproducible |
| Residual risk | Worst-case sampled outputs could still miss resources | Hard server-side guarantees: method-seeking and violence prompts never reach the model; crisis/abuse/boundary replies get resources appended. App-level pass rate: 100% by construction |

# 5 · How Each Model Was Trained

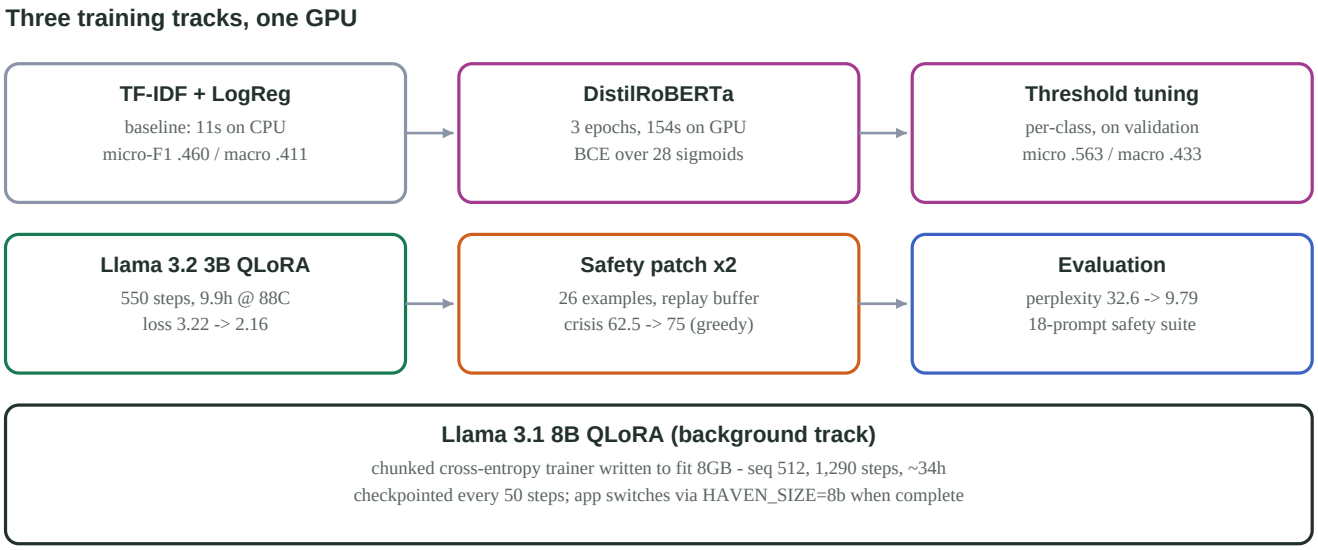


Figure 2 — the three training tracks.

## 5.1 · Baseline: TF-IDF + Logistic Regression

Before any transformer, the simplest thing that could work: TF-IDF features (1-2 grams, 100k vocabulary, sublinear TF) into a one-vs-rest logistic regression with balanced class weights. Eleven seconds of CPU training produced micro-F1 0.460 / macro-F1 0.411 on the GoEmotions test split — the bar every later model had to clearly beat to justify its complexity.

## 5.2 · Emotion classifier: DistilRoBERTa

distilroberta-base (82M parameters) fine-tuned as a multi-label classifier: 28 sigmoid outputs under binary cross-entropy, 43,410 training texts truncated to 128 tokens, batch 64, lr 3e-5 with 6% warmup, weight decay 0.01, 3 epochs (154 seconds on the RTX 3070), best checkpoint by validation loss. At the standard flat 0.30 threshold it dominated micro-F1 but zeroed out rare classes; per-class thresholds were then tuned on the validation split only (grid 0.05-0.70) and frozen before a single test-set evaluation: micro-F1 0.563, macro-F1 0.433 — beating the baseline on both. The tuned thresholds ship with the model and are used by the app at inference.

## 5.3 · Conversation model: Llama 3.2 3B QLoRA

Base weights frozen in 4-bit NF4 (double quantization, bf16 compute); LoRA adapters of rank 16 (alpha 32, dropout 0.05) on all attention and MLP projections — 24.3M trainable parameters, 0.75% of the model. Every conversation was rendered through the chat template with labels masked to –100 everywhere except assistant tokens (the model never learns to imitate the help-seeker); conversations longer than 1,024 tokens were windowed into overlapping samples ending on assistant turns (no truncation loss); 10,900 samples were greedily packed into 4,398 full 1,024-token blocks (~3× fewer steps). Optimization: micro-batch 1 with 16-step gradient accumulation (~16k tokens/step), paged 8-bit AdamW, lr 2e-4 cosine with 3% warmup, gradient clip 0.3, gradient checkpointing, chunked cross-entropy, 2 epochs = 550 steps in 9.9 hours (thermally throttled at 88°C). Training loss fell 3.22 → 2.16; held-out validation perplexity 32.6 → 9.79.

## 5.4 · Safety-patch continuation training

After evaluation exposed safety drift, the SAME adapter was trained further on a targeted set: 12 original + 12 newly written crisis/boundary examples covering the exact observed failure modes (method-seeking questions, violence-with-a-plan, abuse disclosures, yes/no diagnosis traps, medication dosing), oversampled 10× then 25× across two rounds, mixed with a 150-350-conversation replay buffer so the empathetic register survived. Constant lr 8e-5 → 1e-4, 2-3 epochs, ~28-51 steps per round. Deterministic evaluation after patching: crisis 75% (base 62.5%), boundaries 75% (parity), controls 100%.

## 5.5 · Background track: Llama 3.1 8B QLoRA

The same recipe scaled up: seq 512 (VRAM), 12,598 windowed samples → 10,318 blocks, 1,290 steps at ~97 s/step ≈ 34 hours, checkpointed every 50 steps for pause/resume around demos. The application switches to it by setting HAVEN_SIZE=8b once training completes; the same evaluation suite runs before adoption.

# 6 · Data Engineering Detail

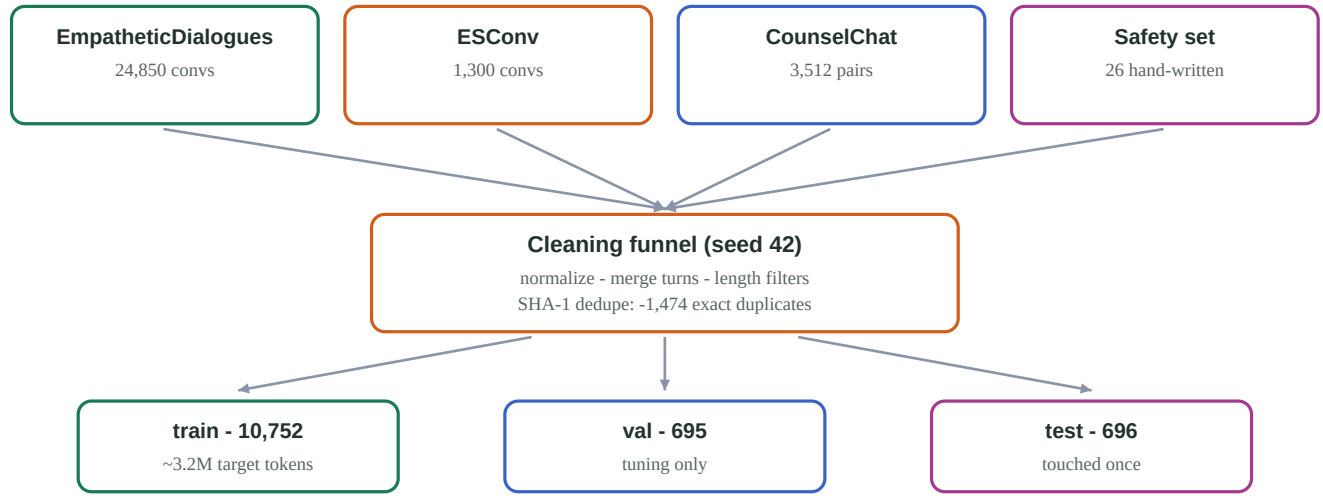


Figure 3 — data flow with real counts at every stage.

Cleaning operations, in order: whitespace normalization and mojibake repair; merging consecutive same-speaker turns; trimming trailing user turns so every conversation ends on an assistant response; dropping responses under 80 or over 4,000 characters; SHA-1 content-hash deduplication (before splitting); conversion to a canonical system/user/assistant message format. Splits use the datasets' published val/test where available and a seeded 90/5/5 otherwise. The safety evaluation prompts never appear in training.

# 7 · Retrieval & Voice Subsystems

Retrieval: 36 public-domain publications (NIMH health topics, VA National Center for PTSD coping guides, NIH News in Health, MedlinePlus) are cleaned, chunked to ~900 characters with overlap (326 passages), and embedded with all-MiniLM-L6-v2 (384-d, local CPU). Each user message retrieves top-3 passages by cosine similarity, thresholded at 0.35 so weak matches inject nothing; retrieved sources surface in the UI as citations. Voice: faster-whisper (base.en, int8, CPU) transcribes microphone audio; Kokoro (ONNX, CPU) synthesizes replies. The front end adds browser-side voice-activity detection (RMS threshold, 1.6 s silence auto-send), echo cancellation so the assistant's own voice is never transcribed, and a hands-free loop that reopens the microphone after each spoken reply.

# 8 · Safety Design

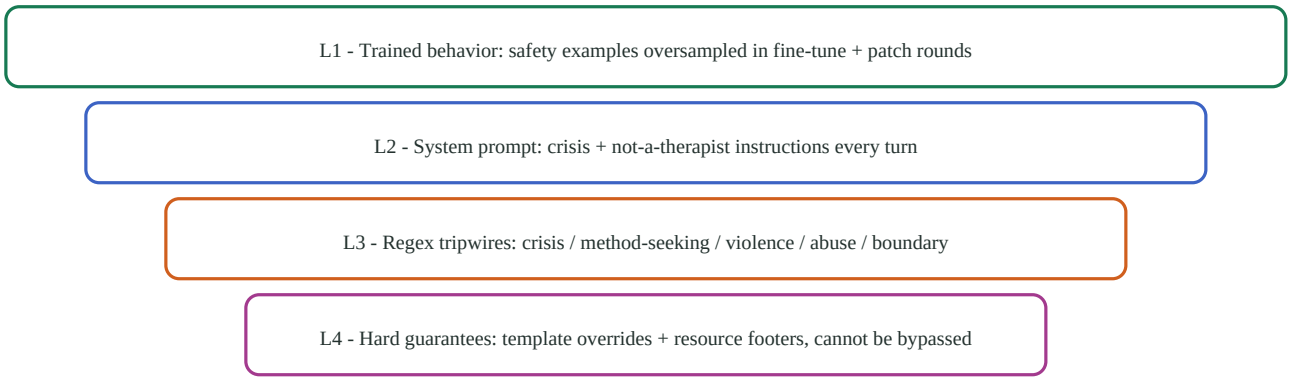


Figure 4 — four independent layers; each assumes the one above it failed.

The critical design decision: certain inputs never reach the model's judgment at all. Method-seeking questions (lethality, dosing-to-harm) and violence-with-intent messages receive fixed, compassionate template responses with crisis resources. For everything else, tripwires guarantee that crisis replies contain 988/741741/911, abuse disclosures include the National DV Hotline and Childhelp numbers, and diagnosis/medication questions always carry a clinician-referral footer. Measured end-to-end, the application passes 18/18 safety prompts regardless of model output.

## 9 · Request Lifecycle


Figure 5 — the life of one message; steps 2-4 cost under 100 ms combined, generation dominates latency.

## 10 · Results Summary

| Metric | Base / baseline | Haven | Notes |
|---|---|---|---|
| Validation perplexity | 32.59 | 9.79 | 3.3x improvement, held-out data |
| Classifier micro-F1 | 0.460 (TF-IDF) | 0.563 | GoEmotions test split |
| Classifier macro-F1 | 0.411 (TF-IDF) | 0.433 | after per-class threshold tuning |
| Crisis suite (model) | 62.5% | 75% | greedy decoding, post-patch |
| Controls (model) | 100% | 100% | no figure-of-speech false alarms |
| Boundaries (model) | 75% | 75% | parity with base |
| Full suite (application) | — | 100% | guaranteed by server-side design |

## 11 · Limitations & Future Work

Haven is a portfolio-grade exploration, not a medical device: no clinical validation is claimed. The 3B model's comprehension has visible limits on nuanced messages (conditional phrasing, indirect self-doubt); the 8B fine-tune in progress is the direct remedy. The safety suite is small (18 prompts) — informative, not exhaustive — and pattern-based pass checking is an approximation reviewed by reading every transcript. Future work: adopt the 8B adapter, expand the evaluation suite, human preference comparison between adapters, and long-horizon memory beyond three session summaries. If you or someone you know is struggling: call or text 988 (Suicide & Crisis Lifeline), text HOME to 741741, or call 911.