Minesweeper LLM Agent: Fine-Tuning Qwen2.5-14B for Competitive Play

Team 92 · February 2026

Problem

An LLM must play Minesweeper by outputting a single JSON action per turn on boards from 6×6 to 50×50. Scoring: +15 safe reveal, +15 correct flag, −25 mine hit, +50 win.

Approach

1. Frontier prompt format (novel). LLMs cannot reason spatially over ASCII grids (7–15% valid moves). We replace the grid with an explicit *frontier format* listing each numbered cell's value, flag count, and hidden neighbor coordinates. This converts spatial reasoning into constraint lookup → **100% valid moves**.

2. Three-tier constraint solver for data generation. We built a custom solver (Tier 1: single-cell propagation, Tier 2: set-based coupled constraints, Tier 3: backtracking Tank solver with Union-Find partitioning) achieving 94% deducible actions. This generates high-quality training labels across 13 board sizes including rectangular boards.

3. Supervised fine-tuning with LoRA. Qwen2.5-14B-Instruct fine-tuned with LoRA (r=64, α=128, 275M/15B params) on 37K curated examples for 1 epoch. Continued SFT on 5K all-frontier examples produced the final model. Loss: 0.91 → 0.09.

4. Three custom GRPO reward functions. Format compliance (R_{format}), gameplay outcome simulation (R_{game}), and strategic quality (R_{strat}). GRPO ultimately degraded performance—SFT already near-optimal with >95% correct actions, leaving insufficient reward variance for 4 generations/prompt.

5. Critical finding: prompt alignment. The system prompt at inference must *exactly* match training. Mismatch causes up to 7.4× degradation (+37.1 → +4.7).

Results

Board Size	Games	Avg Score	Valid JSON	Valid Moves
6 × 6	15	−1.9	100%	100%
8 × 8	15	+29.0	100%	100%
10 × 10	15	+43.3	100%	100%
16 × 16	8	+55.0	100%	100%
20 × 20	8	+49.4	100%	100%
50 × 50	2	+55.0	100%	100%
Rectangular (6 sizes)	28	+25.7	100%	100%
Core (5 sizes)	**61**	**+37.1**	**100%**	**100%**
All 12 sizes	**96**	**+29.1**	**100%**	**100%**

Key metrics: 100% valid JSON, 100% valid moves, ∼20 tokens/response (well under 128 limit), greedy decoding, no post-LLM processing. Final model: SFT-only, 28GB merged weights.