TrialMatch — Stakeholder Update

Date: Feb 22, 2026 | Deadline: Feb 24, 2026 (Kaggle MedGemma Impact Challenge)

What We Built

An AI-powered clinical trial matching system that takes a cancer patient's medical record and automatically finds + evaluates matching clinical trials from ClinicalTrials.gov.

- **3-stage pipeline**: Extract patient facts → Search trials → Evaluate eligibility
- **3 AI models**: MedGemma 4B (medical terms), MedGemma 27B (clinical reasoning), Gemini 3 Pro (orchestration)
- **Live ClinicalTrials.gov integration** via agentic search (not static data)
- **37 real NSCLC patient cases** for demonstration
- **183 automated tests**, zero lint errors

Benchmark Results

20-pair criterion-level evaluation against expert annotations

Model	Accuracy	Notes
GPT-4 (gold standard)	75.0%	Built into reference dataset
MedGemma 27B	70.0%	Our best model, deployed on Google Cloud
Gemini 3 Pro	75.0%	General-purpose, not medical-specific
MedGemma 4B	35.0%	Limited by infra bug (max 512 output tokens)

Key takeaway: MedGemma 27B reaches 70% accuracy on clinical criterion matching — within 5 points of GPT-4, and dramatically better than the smaller 4B model. This validates the multi-model approach.

What's Done (17 of 20 deliverables)

✓ Full backend pipeline: patient ingestion, trial search, criterion evaluation

✓ MedGemma 27B deployed + benchmarked on Google Cloud Vertex AI

✓ Streamlit demo scaffold (patient selector, pipeline viewer, benchmark dashboard)

✓ ClinicalTrials.gov API integration with rate limiting + error handling

✓ Benchmark infrastructure: metrics, run artifacts, cost tracking

✓ 8 Architecture Decision Records documenting key technical choices

What's Left (3 deliverables, ~2 days)

→ **Wire VALIDATE into Streamlit** — connect criterion evaluation to the UI

→ **Playwright QA + demo video** — automated testing + 3-min demo recording

→ **Kaggle submission** — 3-page writeup + code + video upload

Competitive Advantages

★ **Live CT.gov search** — most competitors use static data

★ **Real agentic architecture** — multi-turn tool use, not prompt chaining

★ **Honest benchmarking** — transparent comparison (judges value this)

★ **Multi-model orchestration** — qualifies for Agent Workflows special award ($75K pool)

★ **Production-quality code** — 183 tests, modular architecture, full reproducibility

Risks

Risk	Mitigation
Endpoint cold-start during demo	Cached replay mode as fallback
CT.gov API downtime	Pre-cached trial data for demo patients
MedGemma 27B not beating GPT-4	Narrative = complementary roles, not superiority

Budget Spent

$ **Cloud compute**: ~$70 total (Vertex AI GPU hours for deployment + benchmark)

$ **API costs**: <$1 (20-pair benchmarks across all models)

$ **Infrastructure**: All endpoints torn down after use, no ongoing costs