Assignment: Synthetic Data Generator with Quality Guardrails

Task

Build a synthetic data generator for **service/tool reviews** in a domain of your choice (SaaS, dev tools, design software, etc.).

Requirements

Generation:

- 300-500 samples (fewer acceptable if hardware/model speed is limiting)
- Configurable via YAML/JSON: personas, rating distribution, review characteristics
- Use at least 2 different models/providers (OpenAI, Anthropic, Ollama, local - your choice)

Quality Guardrails:

- Diversity metrics (vocabulary overlap, semantic similarity)
- Bias detection (sentiment skew, unrealistic patterns)
- Domain realism validation
- Automated rejection/regeneration for low-quality samples

Engineering:

- CLI or API interface
- Quality report (markdown/text with metrics)
- Compare against 30-50 real reviews you collect from the web
- Track quality/time per model

Deliverables

1. GitHub repo
2. Generated dataset with quality scores
3. Quality report: metrics, synthetic vs. real comparison, model performance
4. README: setup, design decisions, trade-offs made

Constraints

- Document hardware/model limitations if relevant
- Don’t use AI assistant for ideation and approach (you can vibe-code)