**🧾 Problem Statement:**

**Objective:**  
You aim to integrate a high-throughput, cost-effective Large Language Model (LLM) API into an automated data cleaning pipeline that processes approximately **700 Reddit posts per day**, extracting structured "problem" and "solution" pairs for an automotive troubleshooting assistant used by **Car Clinic**.

**Current Constraints:**

1. **High Volume:** The pipeline must handle **700+ entries daily**, requiring fast and consistent inference capabilities.
2. **Cost Sensitivity:** You are seeking a **free or near-free solution**, with minimal token-based billing or rate limiting.
3. **Deployment Automation:** The pipeline must be compatible with **GitHub Actions**, requiring stateless, headless operation with no manual intervention.
4. **Model Requirements:** The LLM must:
   * Accurately process Reddit post content and comments.
   * Follow a custom prompt to extract structured JSON with high fidelity.
   * Support moderate-length prompts (~200–600 tokens per request).
5. **Platform Reliability:** The API must maintain **uptime and stable inference**, without unexpected downtimes or aggressive throttling that would halt the automation.

**Challenges:**

* Most free LLM endpoints (e.g., Together AI, LiteLLM proxies) impose **low rate limits** or limited availability for production-scale usage.
* Fully-featured models (like DeepSeek R1) often require **paid access** for consistent high-throughput performance.
* Uncertainty about **which model/platform** can sustainably support your current and future volume without breaking cost or speed constraints.

**🔍 Research Question**

*"What LLM provider and integration architecture can reliably and cost-effectively support automated processing of ~700 Reddit posts daily, extracting structured repair data, under free or near-free usage limits, and with compatibility for GitHub Actions-based deployment?"*