Paper Validation Report for llm-agents-for-smart...

Correspondence: Yes
Percentages: 88.0%

Conclusion:

The repository demonstrates substantial compliance with the experiments and methods described in the paper. The code implements multi-agent LLM pipelines for smart city queries, integrates both function/tool-calling and RAG with ChromaDB, supports multi-model evaluation (including Llama, GPT-4o, Mistral), and provides implementations for agent routing, context retrieval, and API integration as described. There is evidence of using the same evaluation metrics (G-Eval, Answer Relevancy via DeepEval), pipelines, and domain-specific APIs as in the reported experiments. Scripts for timing, evaluation, and API-based/human baselines are present, supporting replication of the reported metrics (accuracy, work time, correctness). However, automatic data and experiment result reproducibility may require additional scripting, and precise dataset artifacts for 150 queries and context may not be fully included or easily reproducible from the current codebase. Therefore, a strong (but not perfect) alignment is given.