Assignment: Zocket

"Prototype: RAG Agent for Marketing Query Research"

Documentation

By

Aditya Kumar Pandey

MTech (AI & Data Science)

Indian Institute of Information Technology Bhagalpur

Objective

The objective of this prototype is to build a lightweight AI agent that helps marketing teams quickly get actionable insights from existing marketing blog data by answering user queries using Retrieval-Augmented Generation (RAG).

Problem Statement

Marketing teams often need to extract insights for campaign planning from multiple blogs, resources, and notes, which is time-consuming. A lightweight AI agent can reduce manual search time and generate targeted, relevant marketing advice quickly.

Architecture and Tools Used

To design a lightweight AI agent for marketing-relevant research, we implemented a Retrieval-Augmented Generation (RAG) system using:

- FastAPI: To expose the agent as a REST API (POST /run-agent) for seamless Zocket integration.
- ChromaDB (PersistentClient): Used as a vector database to store and retrieve marketing blog chunks efficiently with metadata.
- SentenceTransformers (all-MiniLM-L6-v2): For embedding user queries and document chunks to enable semantic search.
- TinyLLaMA-1.1B-Chat-v1.0: A fast, open-source, lightweight LLM for coherent, marketing-focused response generation.
- Ngrok: To securely expose the FastAPI endpoint from Colab for demonstration and testing.

The pipeline:

- User sends a marketing-related query to the API.
- The query is embedded and used for semantic retrieval of top relevant blog chunks from ChromaDB.
- Retrieved context is compiled into a structured prompt.
- The LLM generates a clear, actionable marketing response.
- The response is returned to the user via the FastAPI endpoint.

Challenges Faced and Solutions

1. Environment Management: Ensuring GPU utilization for TinyLLaMA on Colab while managing VRAM constraints for stable generation.
2. ChromaDB Deprecation Issues: Migrated to the latest PersistentClient interface to resolve legacy configuration errors.
3. Ngrok Authentication: Integrated verified authtoken for stable, secure endpoint exposure.
4. Cutoff and Consistency: Adjusted max_new_tokens and eos_token_id to prevent incomplete responses and ensure prompt consistency.
5. LLM Selection: Extensively evaluated Mistral, OpenAI GPT, Phi-3, and other open models for balance between speed, quality, and cost before finalizing TinyLLaMA-1.1B-Chat-v1.0 for its fast, low-resource, high-quality generation suited for marketing RAG tasks.

These steps ensured the FastAPI endpoint remains stable, fast, and ready for production workflows.

Potential Improvements and Next Steps

1. Agentic RAG Expansion: Extend to multi-step reasoning by chaining sub-agents for query rephrasing or advanced reranking of retrieved documents.
2. Knowledge Graph Integration: Introduce a lightweight graph for ad platforms, user intents, and creative types to improve retrieval filtering and precision.
3. Evaluation Strategy:

- Relevance Scores: Manual checks on marketing queries.
- Hallucination Rate: Evaluate factual consistency in outputs.
- Latency: Ensure sub-3s response time for usability in real-time marketing workflows.
- Pattern Recognition and Improvement Loop: Integrate a feedback loop or memory nodes (via LangGraph) to refine retrieval and generation quality over time.

1. **Use of Graph RAG / Agentic RAG**

- Currently, the agent uses standard RAG, retrieving relevant blog chunks and generating answers using an LLM. It does not yet use Graph-based RAG or Agentic RAG.
- However, in the future, adding Agentic RAG could allow the agent to handle multi-step reasoning, such as first rephrasing unclear queries before retrieval or refining results through re-ranking steps.
- This would improve recall and precision when answering complex marketing queries by structuring the retrieval and generation process in smaller, clear steps.

2. **Knowledge Graph Integration**

- Currently, the system retrieves data using vector embeddings without using a Knowledge Graph.
- In the future, integrating a Knowledge Graph could help structure domain knowledge like ad platforms, audience intents, or creative types.
- For example, it could map relationships such as “Facebook Ads → Video Ads → Engagement Metrics” to filter and retrieve more relevant content for user queries. This would improve response relevance and precision by providing the LLM with structured context during generation.

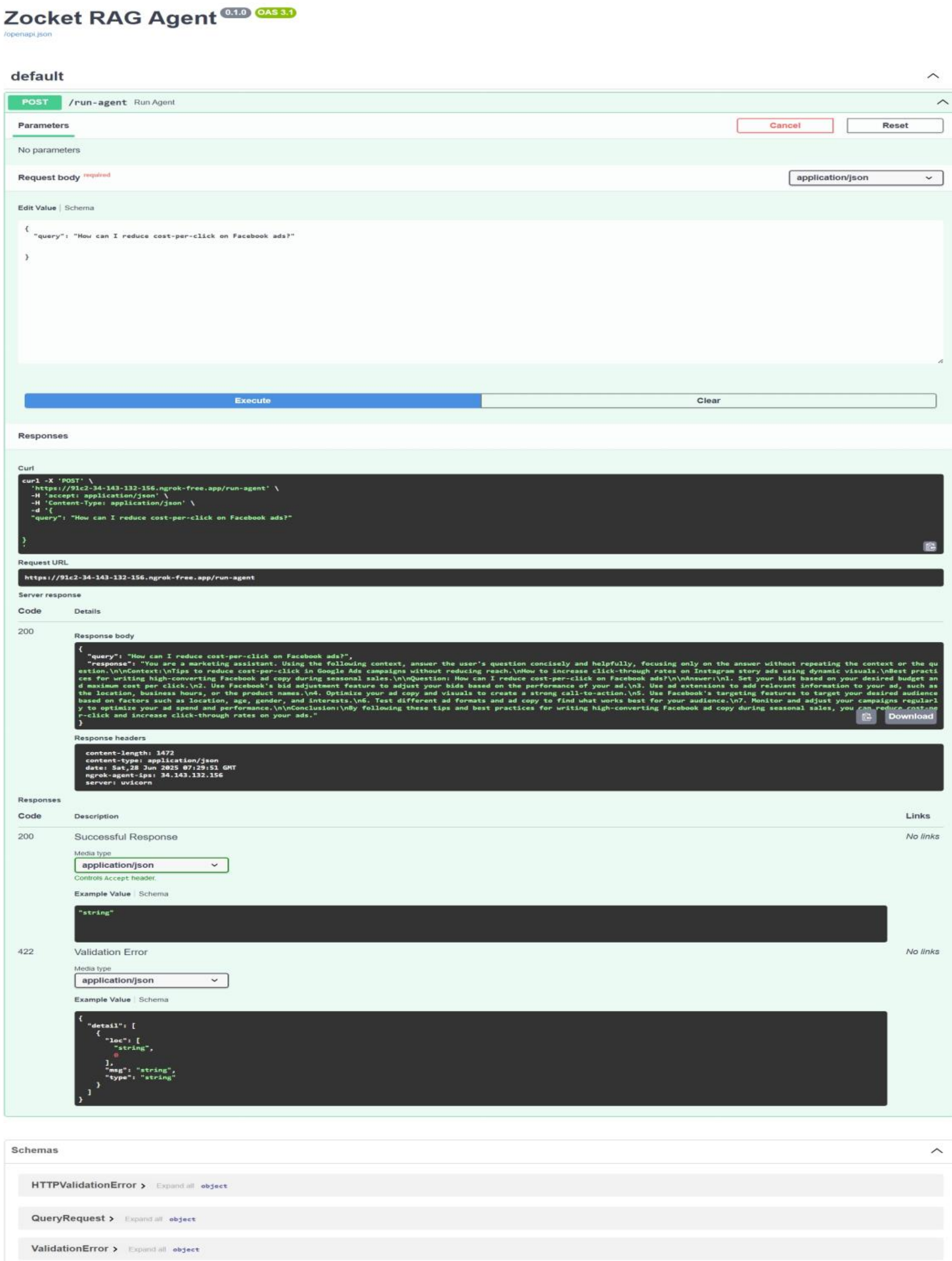
3. **Evaluation Strategy**

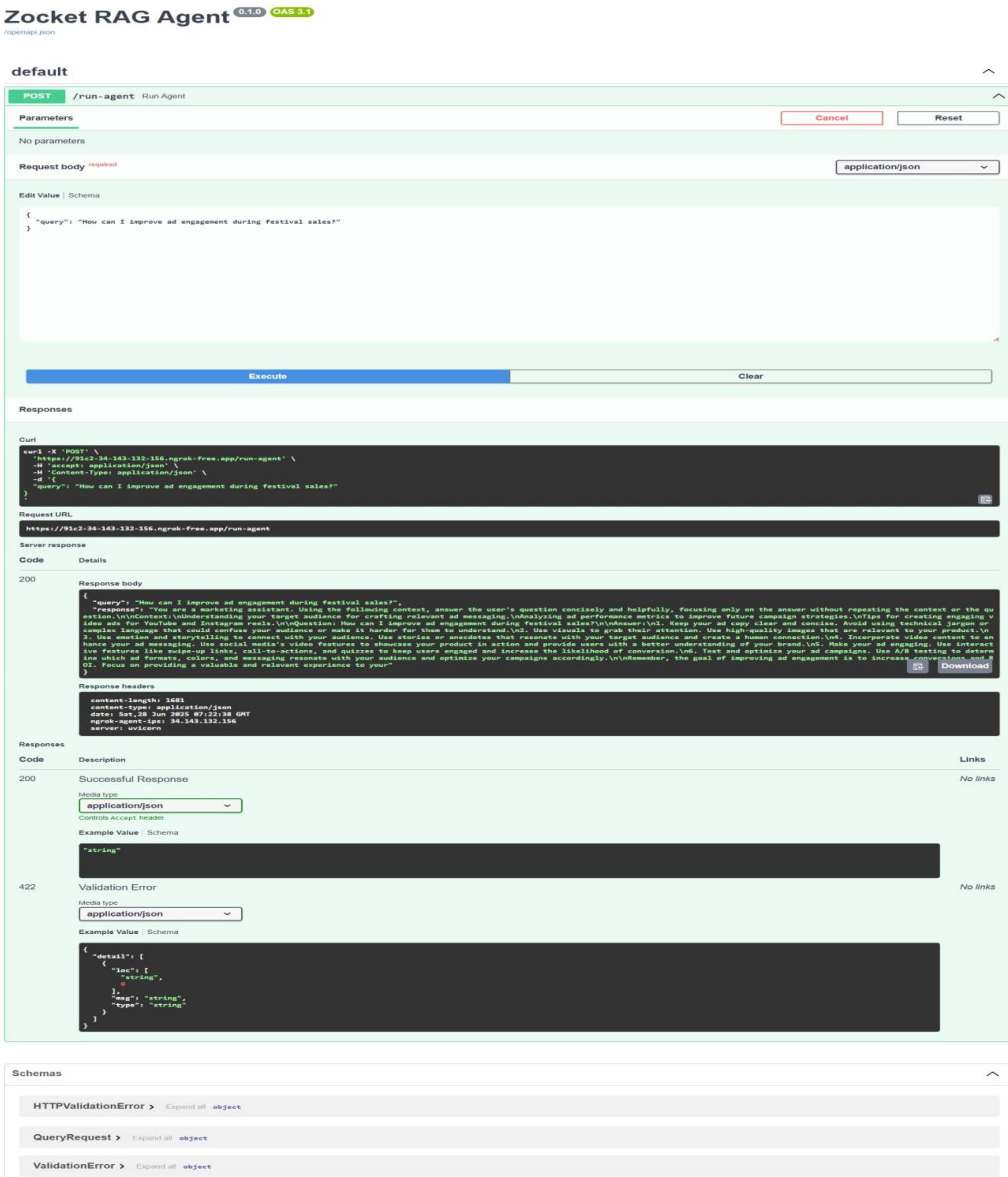
- The agent's performance can be evaluated using the following approach:
- Relevance: Manually check if the agent's answers accurately address the user's marketing questions.
- Hallucination Rate: Track how often the model generates incorrect or unrelated facts.
- Latency: Measure the time taken to generate responses, aiming for fast replies suitable for real-time workflows.
- In advanced stages, automated metrics like F1 score (for extraction accuracy) or ROUGE (for summarization quality) can be added. Initially, manual evaluation will ensure quality control before scaling automated tests.

4. **Pattern Recognition and Improvement Loop**

- The agent can improve over time by incorporating a feedback loop.
- User feedback (like thumbs up/down or relevance ratings) can be used to refine prompts and retrieval settings.
- Additionally, memory modules can be added to remember user preferences and common query patterns, helping the system adjust context and responses in future interactions. These steps will help the agent adapt, reduce repetitive errors, and improve the accuracy and quality of responses over time.

API Test Results


Zocket RAG Agent 0.1.0 OAS 3.1
/openapi.json
default
POST /run-agent Run Agent
Parameters
Cancel
Reset
No parameters
Request body required
application/json
Edit Value | Schema
{
"query": "How can I reduce cost-per-click on Facebook ads?"
}
Execute
Clear
Responses
Curl
curl -X 'POST' \
'https://91c2-34-143-132-156.ngrok-free.app/run-agent' \
-H 'accept: application/json' \
-H 'Content-Type: application/json' \
-d '{
"query": "How can I reduce cost-per-click on Facebook ads?"
}'
Request URL
https://91c2-34-143-132-156.ngrok-free.app/run-agent
Server response
Code
Details
200
Response body
{
"query": "How can I reduce cost-per-click on Facebook ads?",
"response": "You are a marketing assistant. Using the following context, answer the user's question concisely and helpfully, focusing only on the answer without repeating the context or the question.\n\nContext:\nTips to reduce cost-per-click in Google Ads campaigns without reducing reach.\nHow to increase click-through rates on Instagram story ads using dynamic visuals.\nBest practices for writing high-converting Facebook ad copy during seasonal sales.\n\nQuestion: How can I reduce cost-per-click on Facebook ads?\n\nAnswer:\n1. Set your bids based on your desired budget and maximum cost per click.\n2. Use Facebook's bid adjustment feature to adjust your bids based on the performance of your ad.\n3. Use ad extensions to add relevant information to your ad, such as the location, business hours, or the product names.\n4. Optimize your ad copy and visuals to create a strong call-to-action.\n5. Use Facebook's targeting features to target your desired audience based on factors such as location, age, gender, and interests.\n6. Test different ad formats and ad copy to find what works best for your audience.\n7. Monitor and adjust your campaigns regularly to optimize your ad spend and performance.\n\nConclusion:\nBy following these tips and best practices for writing high-converting Facebook ad copy during seasonal sales, you can reduce cost-per-click and increase click-through rates on your ads."
}
Download
Response headers
content-length: 1472
content-type: application/json
date: Sat,28 Jun 2025 07:29:51 GMT
ngrok-agent-ips: 34.143.132.156
server: uvicorn
Responses
Code
Description
Links
200
Successful Response
No links
Media type
application/json
Controls Accept header.
Example Value | Schema
"string"
422
Validation Error
No links
Media type
application/json
Example Value | Schema
{
"detail": [
{
"loc": [
"string",
0
],
"msg": "string",
"type": "string"
}
]
}
Schemas
HTTPValidationError > Expand all object
QueryRequest > Expand all object
ValidationError > Expand all object


Zocket RAG Agent
0.1.0
OAS 3.1
/openapi.json
default
POST /run-agent Run Agent
Parameters
Cancel
Reset
No parameters
Request body required
application/json
Edit Value | Schema
{
"query": "How can I improve ad engagement during festival sales?"
}
Execute
Clear
Responses
Curl
curl -X 'POST' \
'https://91c2-34-143-132-156.ngrok-free.app/run-agent' \
-H 'accept: application/json' \
-H 'Content-Type: application/json' \
-d '{
"query": "How can I improve ad engagement during festival sales?"
}'
Request URL
https://91c2-34-143-132-156.ngrok-free.app/run-agent
Server response
Code
Details
200
Response body
{
"query": "How can I improve ad engagement during festival sales?",
"response": "You are a marketing assistant. Using the following context, answer the user's question concisely and helpfully, focusing only on the answer without repeating the context or the question.\n\nContext:\nUnderstanding your target audience for crafting relevant ad messaging.\nAnalyzing ad performance metrics to improve future campaign strategies.\nTips for creating engaging video ads for YouTube and Instagram reels.\n\nQuestion: How can I improve ad engagement during festival sales?\n\nAnswer:\n1. Keep your ad copy clear and concise. Avoid using technical jargon or complex language that could confuse your audience or make it harder for them to understand.\n2. Use visuals to grab their attention. Use high-quality images that are relevant to your product.\n3. Use emotion and storytelling to connect with your audience. Use stories or anecdotes that resonate with your target audience and create a human connection.\n4. Incorporate video content to enhance your ad messaging. Use social media's video features to showcase your product in action and provide users with a better understanding of your brand.\n5. Make your ad engaging. Use interactive features like swipe-up links, call-to-actions, and quizzes to keep users engaged and increase the likelihood of conversion.\n6. Test and optimize your ad campaigns. Use A/B testing to determine which ad formats, colors, and messaging resonate with your audience and optimize your campaigns accordingly.\n\nRemember, the goal of improving ad engagement is to increase conversions and ROI. Focus on providing a valuable and relevant experience to your"
}
Download
Response headers
content-length: 1681
content-type: application/json
date: Sat,28 Jun 2025 07:22:38 GMT
ngrok-agent-ips: 34.143.132.156
server: uvicorn
Responses
Code
Description
Links
200
Successful Response
No links
Media type
application/json
Controls Accept header
Example Value | Schema
"string"
422
Validation Error
No links
Media type
application/json
Example Value | Schema
{
"detail": [
{
"loc": [
"string",
0
],
"msg": "string",
"type": "string"
}
]
}
Schemas
HTTPValidationError > Expand all object
QueryRequest > Expand all object
ValidationError > Expand all object