M2U3 — Individual Assignment: RAG Mini Project (n8n)

github osama-ata/MAICEN2511

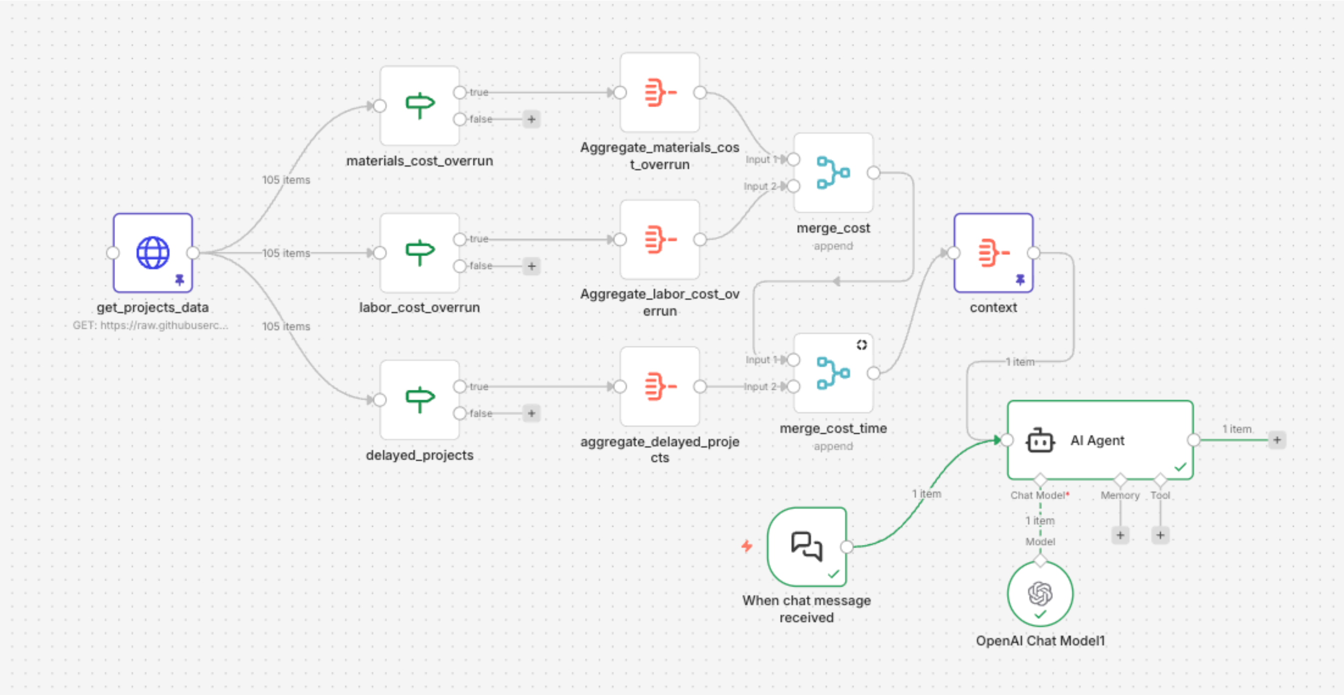
1) Goal

Build a simple Retrieval-Augmented Generation (RAG) workflow in **n8n** that answers questions **only using my own dataset**. The workflow loads structured project data, filters it into relevant subsets (retrieved context), and provides that context to an LLM to generate a constrained answer.

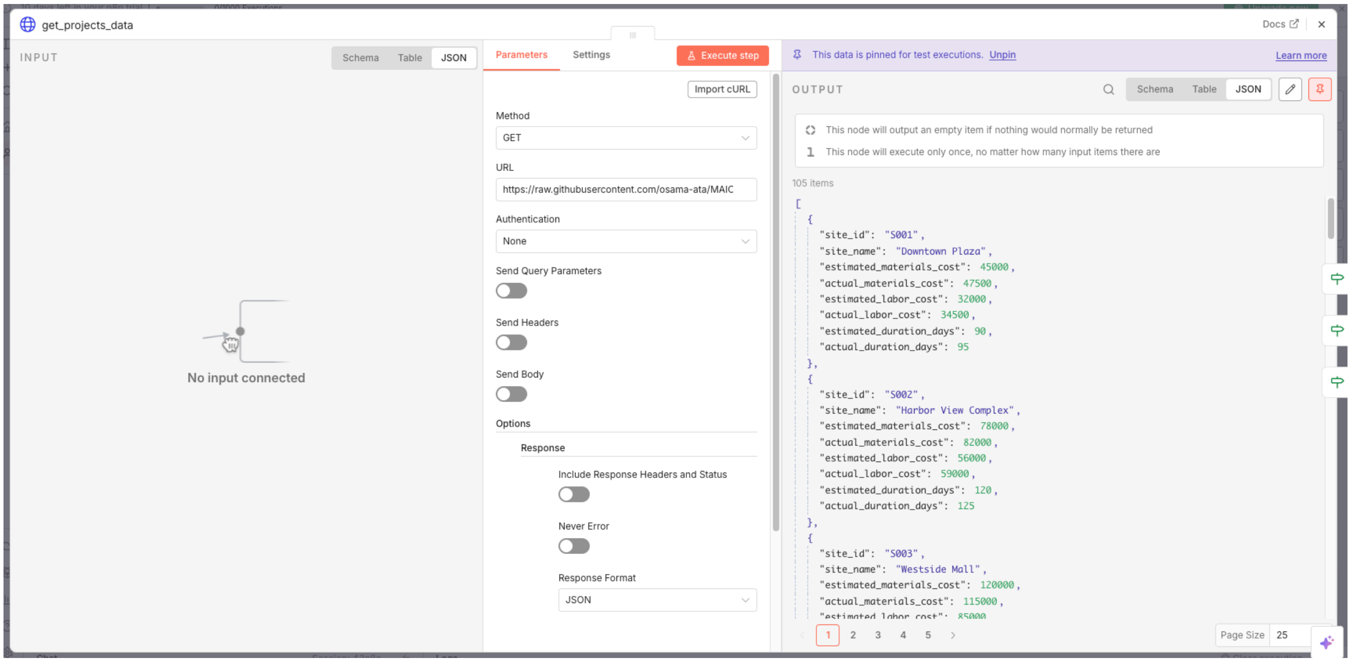
---

2) Submission Artifacts (Deliverables)

- **Workflow export (JSON):** `M2U3/RAG_Project_Workflow.json`
- **Screenshots (3–6 total):** `M2U3/screenshots/`
  Screenshot per major section of the workflow:
  - `01_workflow.png`
  - `02_getting_data.png`
  - `03_filtering_data.png`
  - `04_merging_datasets.png`
  - `05_merging_prompt_context.png`
  - `06_llm_call.png`



Overall Workflow



Method: GET

URL: https://raw.githubusercontent.com/osama-ata/MAIC

Authentication: None

Send Query Parameters

Send Headers

Send Body

Options

Response

Include Response Headers and Status

Never Error

Response Format: JSON

This data is pinned for test executions. Unpin Learn more

This node will output an empty item if nothing would normally be returned

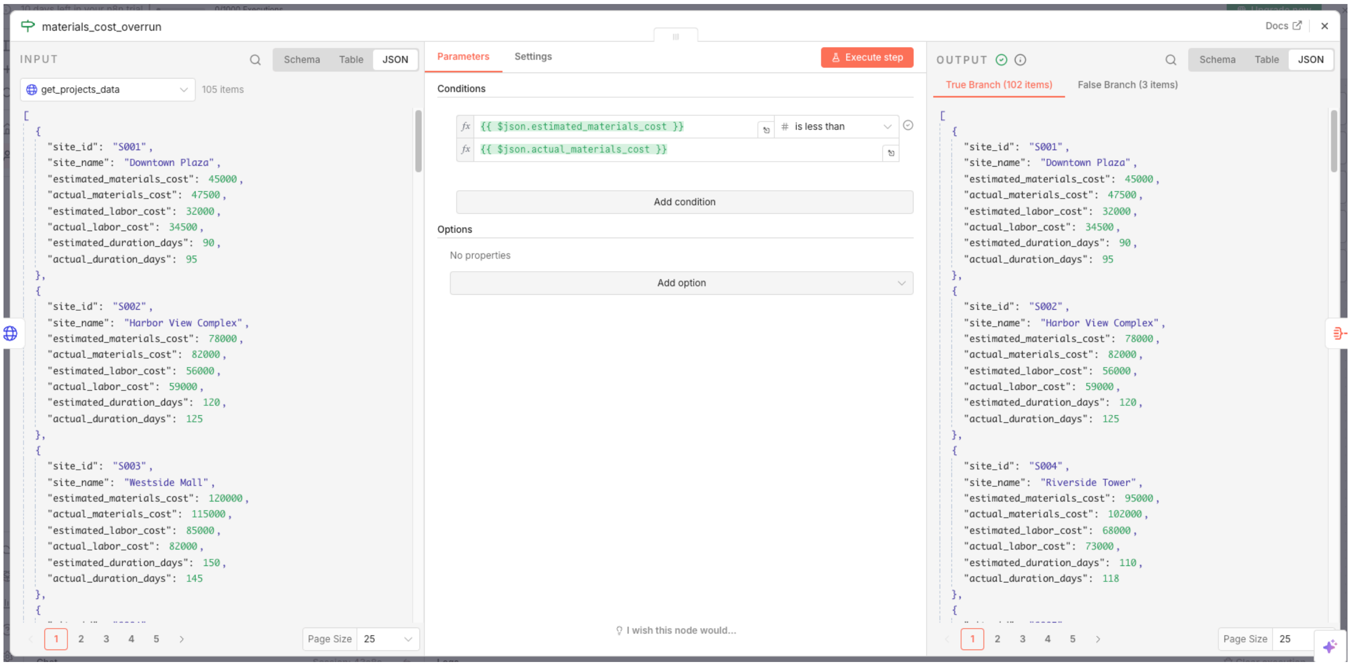
1 This node will execute only once, no matter how many input items there are

105 items

```
[
  {
    "site_id": "S001",
    "site_name": "Downtown Plaza",
    "estimated_materials_cost": 45000,
    "actual_materials_cost": 47500,
    "estimated_labor_cost": 32000,
    "actual_labor_cost": 34500,
    "estimated_duration_days": 90,
    "actual_duration_days": 95
  },
  {
    "site_id": "S002",
    "site_name": "Harbor View Complex",
    "estimated_materials_cost": 78000,
    "actual_materials_cost": 82000,
    "estimated_labor_cost": 56000,
    "actual_labor_cost": 59000,
    "estimated_duration_days": 120,
    "actual_duration_days": 125
  },
  {
    "site_id": "S003",
    "site_name": "Westside Mall",
    "estimated_materials_cost": 120000,
    "actual_materials_cost": 115000,
    "estimated_labor_cost": 85000
```

1 2 3 4 5 Page Size 25

Getting Data



```
[
  {
    "site_id": "S001",
    "site_name": "Downtown Plaza",
    "estimated_materials_cost": 45000,
    "actual_materials_cost": 47500,
    "estimated_labor_cost": 32000,
    "actual_labor_cost": 34500,
    "estimated_duration_days": 90,
    "actual_duration_days": 95
  },
  {
    "site_id": "S002",
    "site_name": "Harbor View Complex",
    "estimated_materials_cost": 78000,
    "actual_materials_cost": 82000,
    "estimated_labor_cost": 56000,
    "actual_labor_cost": 59000,
    "estimated_duration_days": 120,
    "actual_duration_days": 125
  },
  {
    "site_id": "S003",
    "site_name": "Westside Mall",
    "estimated_materials_cost": 120000,
    "actual_materials_cost": 115000,
    "estimated_labor_cost": 85000,
    "actual_labor_cost": 82000,
    "estimated_duration_days": 150,
    "actual_duration_days": 145
  },
  {
```

Conditions

{{ $json.estimated_materials_cost }} is less than
{{ $json.actual_materials_cost }}

Add condition

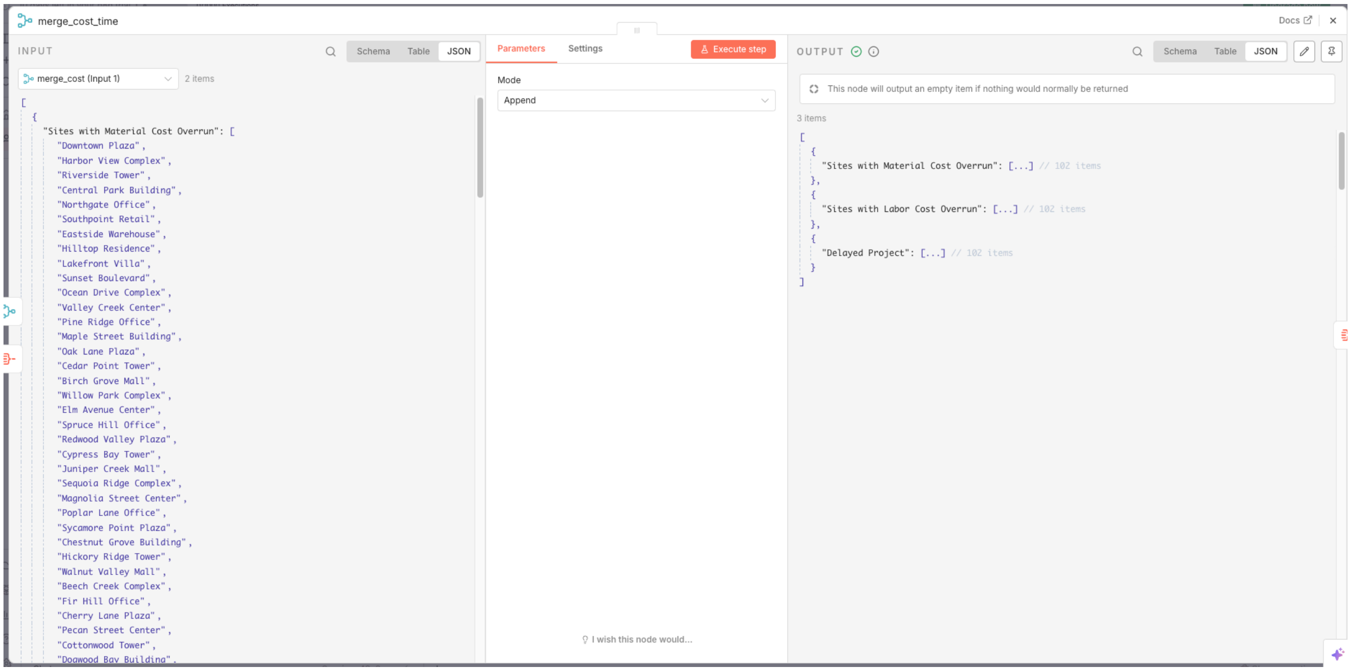
Options

No properties

Add option

```
[
  {
    "site_id": "S001",
    "site_name": "Downtown Plaza",
    "estimated_materials_cost": 45000,
    "actual_materials_cost": 47500,
    "estimated_labor_cost": 32000,
    "actual_labor_cost": 34500,
    "estimated_duration_days": 90,
    "actual_duration_days": 95
  },
  {
    "site_id": "S002",
    "site_name": "Harbor View Complex",
    "estimated_materials_cost": 78000,
    "actual_materials_cost": 82000,
    "estimated_labor_cost": 56000,
    "actual_labor_cost": 59000,
    "estimated_duration_days": 120,
    "actual_duration_days": 125
  },
  {
    "site_id": "S004",
    "site_name": "Riverside Tower",
    "estimated_materials_cost": 95000,
    "actual_materials_cost": 102000,
    "estimated_labor_cost": 68000,
    "actual_labor_cost": 73000,
    "estimated_duration_days": 110,
    "actual_duration_days": 118
  },
  {
```

Filtering Data



```
[
  {
    "Sites with Material Cost Overrun": [
      "Downtown Plaza",
      "Harbor View Complex",
      "Riverside Tower",
      "Central Park Building",
      "Northgate Office",
      "Southpoint Retail",
      "Eastside Warehouse",
      "Hilltop Residence",
      "Lakefront Villa",
      "Sunset Boulevard",
      "Ocean Drive Complex",
      "Valley Creek Center",
      "Pine Ridge Office",
      "Maple Street Building",
      "Oak Lane Plaza",
      "Cedar Point Tower",
      "Birch Grove Mall",
      "Willow Park Complex",
      "Elm Avenue Center",
      "Spruce Hill Office",
      "Redwood Valley Plaza",
      "Cypress Bay Tower",
      "Juniper Creek Mall",
      "Sequoia Ridge Complex",
      "Magnolia Street Center",
      "Poplar Lane Office",
      "Sycamore Point Plaza",
      "Chestnut Grove Building",
      "Hickory Ridge Tower",
      "Walnut Valley Mall",
      "Beech Creek Complex",
      "Fir Hill Office",
      "Cherry Lane Plaza",
      "Pecan Street Center",
      "Cottonwood Tower",
      "Dogwood Bay Building",
```

Mode

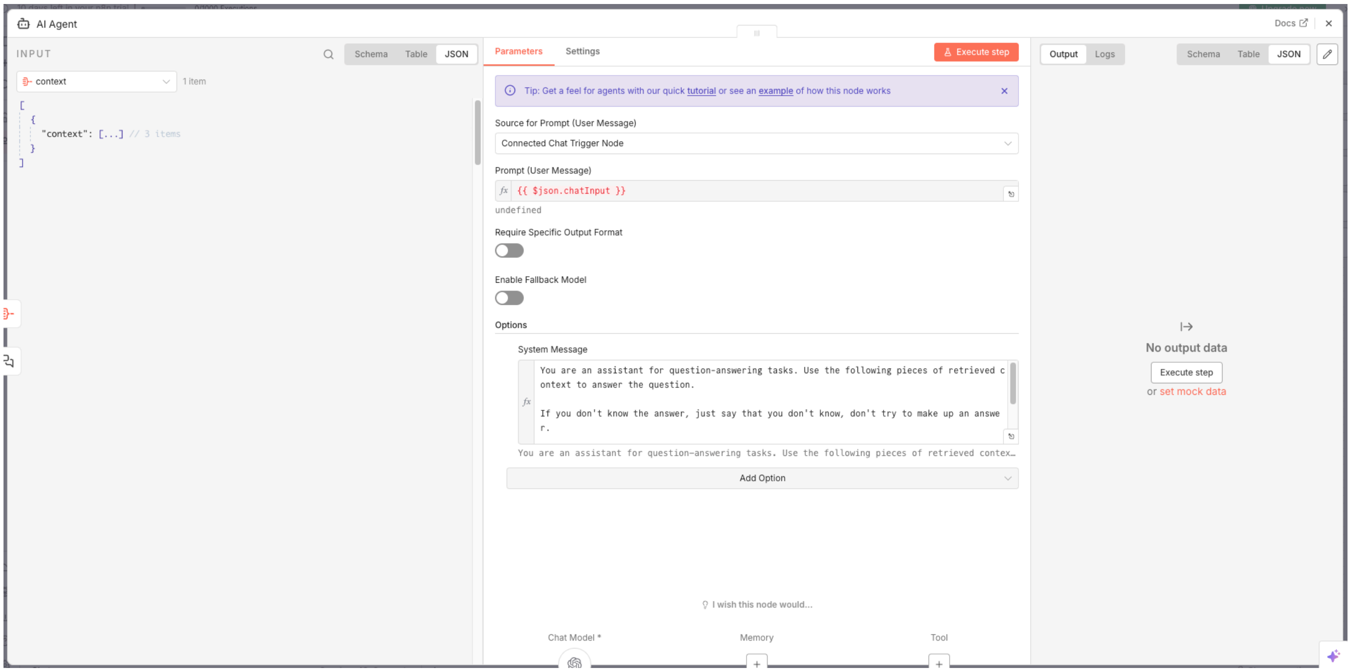
Append

This node will output an empty item if nothing would normally be returned

3 items

```
[
  {
    "Sites with Material Cost Overrun": [...] // 102 items
  },
  {
    "Sites with Labor Cost Overrun": [...] // 102 items
  },
  {
    "Delayed Project": [...] // 102 items
  }
]
```

Merging Datasets



AI Agent

INPUT

context — 1 item

```
[
  {
    "context": [...] // 3 items
  }
]
```

Parameters | Settings

Tip: Get a feel for agents with our quick tutorial or see an example of how this node works

Source for Prompt (User Message): Connected Chat Trigger Node

Prompt (User Message):

```
{{ $json.chatInput }}
```

undefined

Require Specific Output Format: ○

Enable Fallback Model: ○

Options

System Message

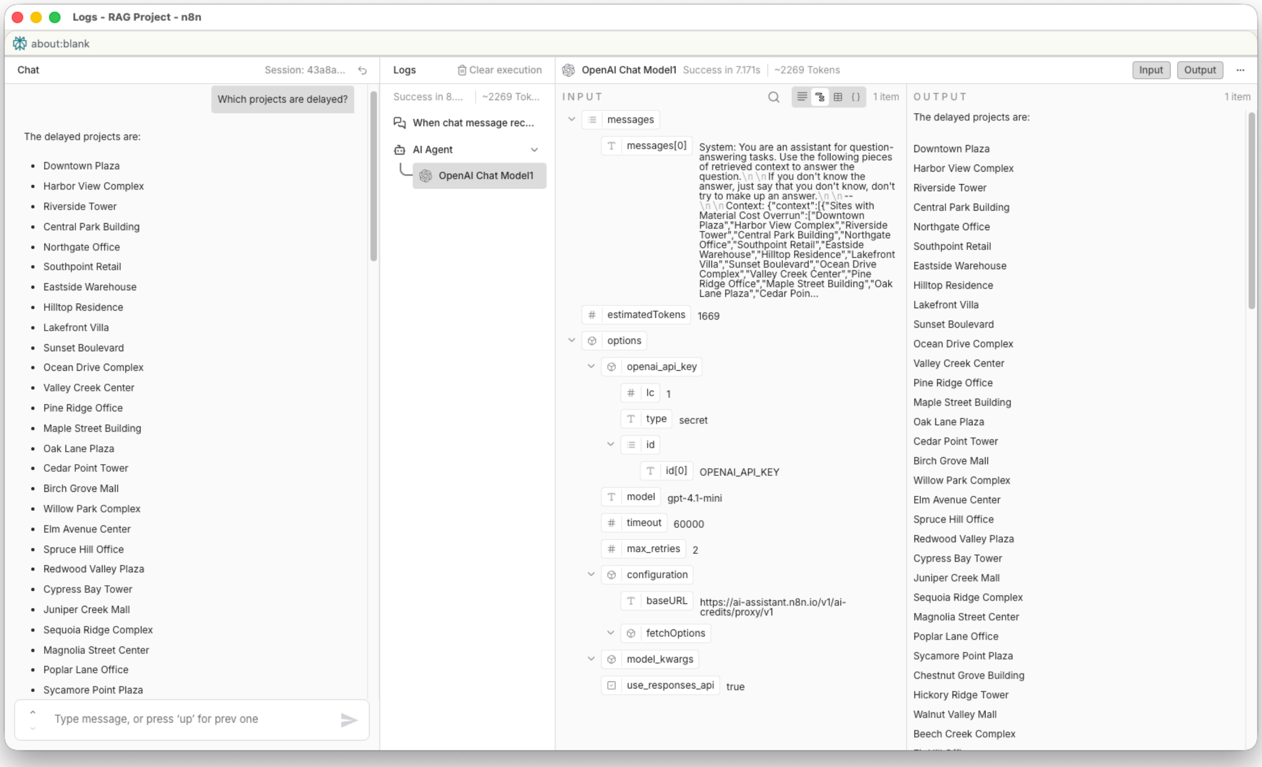
```
You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question.

If you don't know the answer, just say that you don't know, don't try to make up an answer.
```

You are an assistant for question-answering tasks. Use the following pieces of retrieved contex…

Output — No output data

Merging Prompt Context



Logs - RAG Project - n8n

about:blank

Chat — Session: 43a8a...

Which projects are delayed?

The delayed projects are:

- Downtown Plaza
- Harbor View Complex
- Riverside Tower
- Central Park Building
- Northgate Office
- Southpoint Retail
- Eastside Warehouse
- Hilltop Residence
- Lakefront Villa
- Sunset Boulevard
- Ocean Drive Complex
- Valley Creek Center
- Pine Ridge Office
- Maple Street Building
- Oak Lane Plaza
- Cedar Point Tower
- Birch Grove Mall
- Willow Park Complex
- Elm Avenue Center
- Spruce Hill Office
- Redwood Valley Plaza
- Cypress Bay Tower
- Juniper Creek Mall
- Sequoia Ridge Complex
- Magnolia Street Center
- Poplar Lane Office
- Sycamore Point Plaza

Logs — Success in 8.... | ~2269 Tok...

- When chat message rec...
- AI Agent
  - OpenAI Chat Model1

OpenAI Chat Model1 — Success in 7.171s | ~2269 Tokens

INPUT — 1 item

```
messages
  messages[0]   System: You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question.\n\nIf you don't know the answer, just say that you don't know, don't try to make up an answer.\n\n--\n\nContext: {"context":[{"Sites with Material Cost Overrun":["Downtown Plaza","Harbor View Complex","Riverside Tower","Central Park Building","Northgate Office","Southpoint Retail","Eastside Warehouse","Hilltop Residence","Lakefront Villa","Sunset Boulevard","Ocean Drive Complex","Valley Creek Center","Pine Ridge Office","Maple Street Building","Oak Lane Plaza","Cedar Poin...
estimatedTokens   1669
options
  openai_api_key
    lc   1
    type   secret
    id
      id[0]   OPENAI_API_KEY
  model   gpt-4.1-mini
  timeout   60000
  max_retries   2
  configuration
    baseURL   https://ai-assistant.n8n.io/v1/ai-credits/proxy/v1
    fetchOptions
  model_kwargs
    use_responses_api   true
```

OUTPUT — 1 item

The delayed projects are:

Downtown Plaza
Harbor View Complex
Riverside Tower
Central Park Building
Northgate Office
Southpoint Retail
Eastside Warehouse
Hilltop Residence
Lakefront Villa
Sunset Boulevard
Ocean Drive Complex
Valley Creek Center
Pine Ridge Office
Maple Street Building
Oak Lane Plaza
Cedar Point Tower
Birch Grove Mall
Willow Park Complex
Elm Avenue Center
Spruce Hill Office
Redwood Valley Plaza
Cypress Bay Tower
Juniper Creek Mall
Sequoia Ridge Complex
Magnolia Street Center
Poplar Lane Office
Sycamore Point Plaza
Chestnut Grove Building
Hickory Ridge Tower
Walnut Valley Mall
Beech Creek Complex

LLM Call

3) Data Used

- **Dataset:** construction project/site cost + duration data (JSON derived from CSV)
- **Primary columns:**
  - `site_id`, `site_name`
  - `estimated_materials_cost`, `actual_materials_cost`
  - `estimated_labor_cost`, `actual_labor_cost`
  - `estimated_duration_days`, `actual_duration_days`
- **Source file referenced by workflow:** `M2U3/project_data.json` (fetched via HTTP in the workflow)

---

4) Workflow Summary (What happens)

1. Load the dataset (JSON).
2. "Retrieve" relevant subsets by filtering:
   - sites where **actual materials cost > estimated**
   - sites where **actual labor cost > estimated**
   - sites where **actual duration > estimated**
3. Aggregate filtered results into a single **context** object.
4. Provide that context to the LLM so answers are grounded in retrieved data.

This design supports questions such as:

- "Which sites have material cost overruns?"
- "Which sites are delayed?"
- "List sites with labor overruns."

---

5) One Sentence Per Node (as required)

- **When chat message received**: Receives the user's question (chat input) to trigger the workflow.
- **get_projects_data**: Loads the project dataset (JSON) into the workflow.

- **materials_cost_overrun (IF)**: Filters rows where `actual_materials_cost` exceeds `estimated_materials_cost`.
- **labor_cost_overrun (IF)**: Filters rows where `actual_labor_cost` exceeds `estimated_labor_cost`.
- **delayed_projects (IF)**: Filters rows where `actual_duration_days` exceeds `estimated_duration_days`.
- **Aggregate_materials_cost_overrun**: Collects the `site_name` values for material-overrun sites into one list.
- **Aggregate_labor_cost_overrun**: Collects the `site_name` values for labor-overrun sites into one list.
- **aggregate_delayed_projects**: Collects the `site_name` values for delayed sites into one list.
- **merge_cost**: Merges the two cost-overrun aggregations into a single stream.
- **merge_cost_time**: Merges cost results with delay results to form a combined dataset.
- **context (Aggregate)**: Aggregates merged results into a single `context` payload for the LLM.
- **OpenAI Chat Model1**: Provides the chat model used by the agent for generation.
- **AI Agent**: Generates the final answer using only the provided `context`.

6) RAG Prompt Used

```
Answer only using the data below.
DATA:
{{context}}
QUESTION:
{{query}}
If the answer is not in the data, reply exactly: "No information available."
```

> Note: In this workflow, the “retrieved context” is the filtered + aggregated lists produced upstream (overruns/delays), which constrains what the model can truthfully answer.

7) How to Run (n8n)

1. Open n8n.
2. Import the workflow JSON: `M2U3/RAG_Project_Workflow.json`
3. Configure credentials for the OpenAI node (or replace with your preferred provider).
4. Execute the workflow (or use the chat trigger) and ask questions like:
   - "Which sites are delayed?"
   - "Which sites have labor cost overruns?"

---

8) Short Reflection (≤ 500 words)

What the data is:
The dataset represents construction sites with estimated vs. actual costs (materials and labor) and estimated vs. actual durations. Each row is a site, which makes it suitable for retrieval-by-filtering.

What questions people will ask:
Typical stakeholder questions are list-based and exception-focused: which sites are over budget (materials/labor), which are delayed, and which sites require attention due to negative variance.

What the AI should reply with:
Short, factual answers grounded in the retrieved results (e.g., a list of site names that match the condition). If a question cannot be answered from the retrieved context, the assistant must respond exactly with: **"No information available."**

Why clean data matters:
RAG depends on matching and filtering. If column names, types, or units are inconsistent (e.g., costs stored as strings, missing values, mixed currencies), retrieval will either miss relevant rows or include incorrect ones—both of which degrade the quality of the final answer.

Why filtering is "search before generate":
The workflow first narrows the dataset to only the rows that meet the query-relevant conditions (overrun or delay). This step plays the role of retrieval: it reduces the information

space before the model generates any natural language output.

How RAG reduces hallucination:
By supplying a bounded context (filtered lists) and instructing the model to only use that context, the model has less room to invent unsupported details.

What was easy vs. confusing:
Using n8n's required extra learning to understand how to chain nodes and pass data between them. However, the visual interface made it easier to conceptualize the workflow steps compared to coding from scratch.