## Tegra Test - Lindon

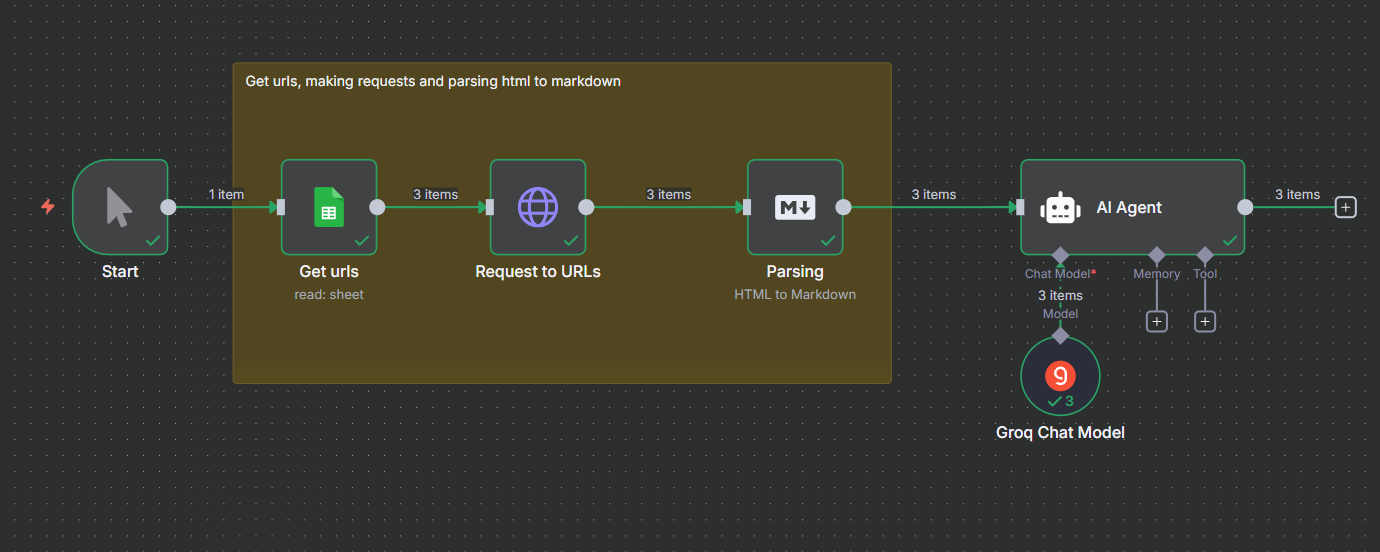
Beforehand, I’d like to thank you for this opportunity.

Sadly, I could not deliver what was properly expected as the test demanded many things that I did not have experience with, which in turn made me more aware of things that I need to study. Some things such as the self-hosted environment I managed to do.

Still, I’d like to thank you for the opportunity.

### **Workflow Architecture**

Workflow

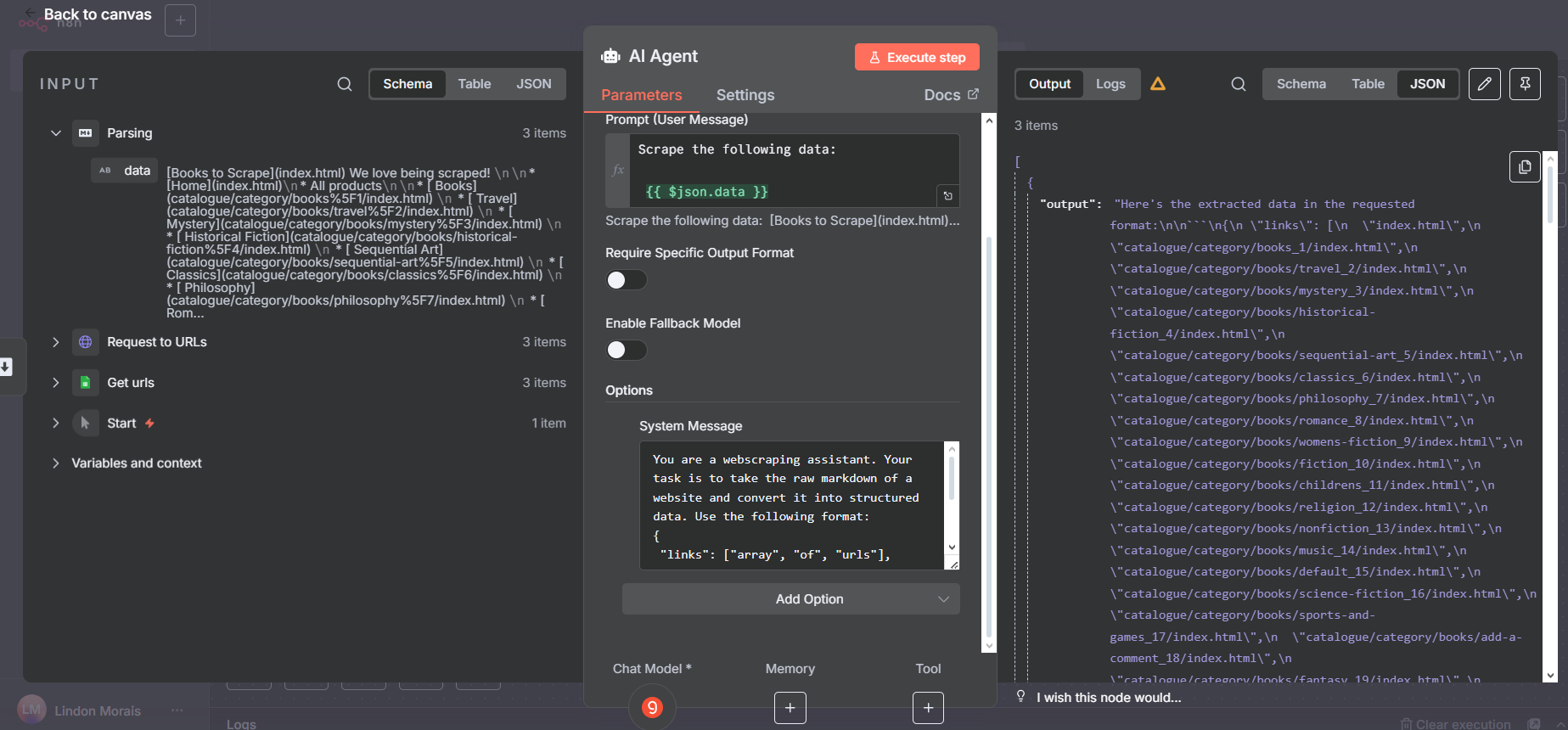


Here I made a simple workflow where we:

* Retrieved the spreadsheet from google drive
* Made separate requests for each one of the URLs
* Parsed the HTML to Markdown

After this part I got a bit overwhelmed with the tasks. I spent most of the time studying about Puppeteer, but I wasn’t really sure how I would make it work on my VPS.

I did use Groq and asked it to retrieve the first 10 links from each requests, however, due to the fact that the pages were made with JS, it retrieved partial links as you can see below:



### Vector Database

Originally I could not make the create table command work on supabase, however, upon checking the command with AI, I learned that the data type “vector” was not being originally used, and I had to fix that in the code as follows:

-- Enable the pgvector extension (if not already installed)

CREATE EXTENSION IF NOT EXISTS vector;

-- Create table

CREATE TABLE scraped\_content (

id SERIAL PRIMARY KEY,

source\_url TEXT NOT NULL,

chunk\_text TEXT NOT NULL,

chunk\_index INTEGER,

total\_chunks INTEGER,

embedding vector(1536), -- requires pgvector extension

metadata JSONB,

created\_at TIMESTAMP DEFAULT NOW()

);

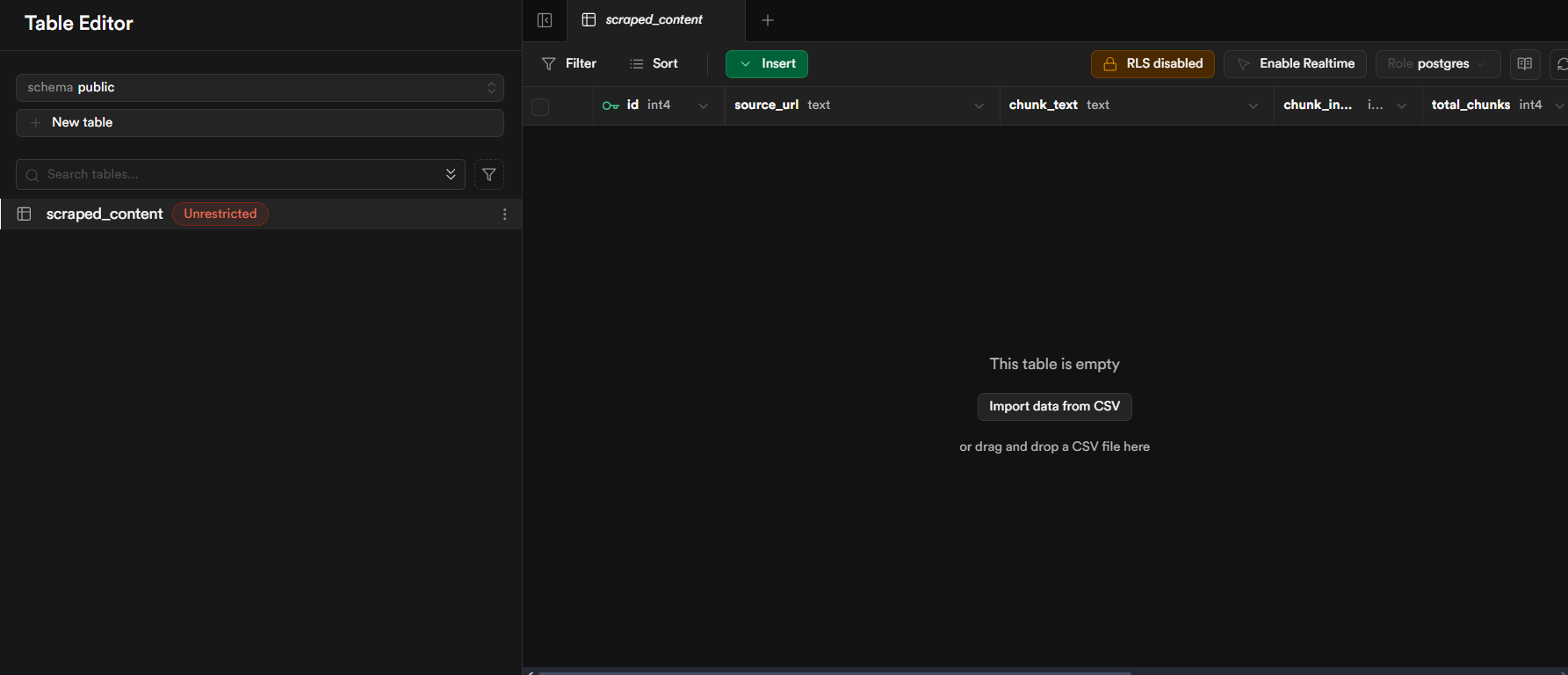
-- Create index for vector similarity search

CREATE INDEX ON scraped\_content

USING ivfflat (embedding vector\_cosine\_ops)

WITH (lists = 100);

Still, since the scraping part did not deliver as expected. I did not insert the data in the database, but the database was created and integrated.



### Points of improvement

As I made clear, I am obviously not happy about my performance, so I decided to list things that I can improve about myself and about the process:

* Learn how to use tools such as Puppeteer, CloudScraper, 2captcha and Anti-captcha
* Instead of using Groq we can use a more robust AI tool, which improves the output quality
* Create a subworkflow for the webscraping so we can use it for scraping both the main links and the sub-links.