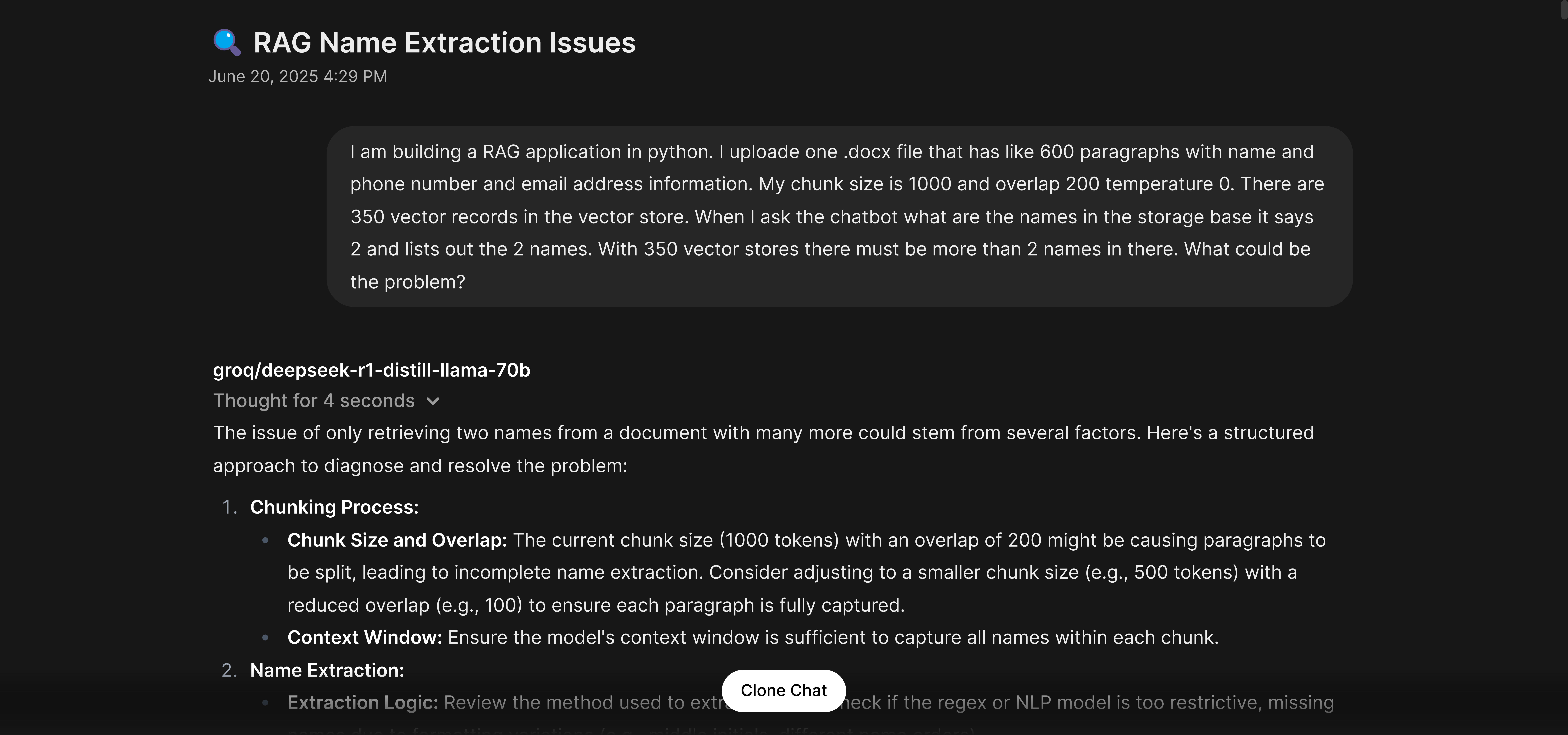

🔍 RAG Name Extraction Issues

June 20, 2025 4:29 PM

I am building a RAG application in python. I uploade one .docx file that has like 600 paragraphs with name and phone number and email address information. My chunk size is 1000 and overlap 200 temperature 0. There are 350 vector records in the vector store. When I ask the chatbot what are the names in the storage base it says 2 and lists out the 2 names. With 350 vector stores there must be more than 2 names in there. What could be the problem?

groq/deepseek-r1-distill-llama-70b

Thought for 4 seconds

The issue of only retrieving two names from a document with many more could stem from several factors. Here's a structured approach to diagnose and resolve the problem:

1. **Chunking Process:**
   - **Chunk Size and Overlap:** The current chunk size (1000 tokens) with an overlap of 200 might be causing paragraphs to be split, leading to incomplete name extraction. Consider adjusting to a smaller chunk size (e.g., 500 tokens) with a reduced overlap (e.g., 100) to ensure each paragraph is fully captured.
   - **Context Window:** Ensure the model's context window is sufficient to capture all names within each chunk.
2. **Name Extraction:**
   - **Extraction Logic:** Review the method used to extr heck if the regex or NLP model is too restrictive, missing