

# ChatGeo-Magi

Milan Nair<sup>2,3</sup>, Arnaud Chulliat<sup>2,3</sup>, Manoj Nair<sup>2,3</sup>, Rob Redmon<sup>3</sup>  
1) Fairview High School, Boulder CO 2) CIRES, University of Colorado Boulder 3) NOAA's NCEI, Boulder, CO

## A Retrieval-Augmented and Tool-Using Language Model for Geomagnetic Data Access and User Support

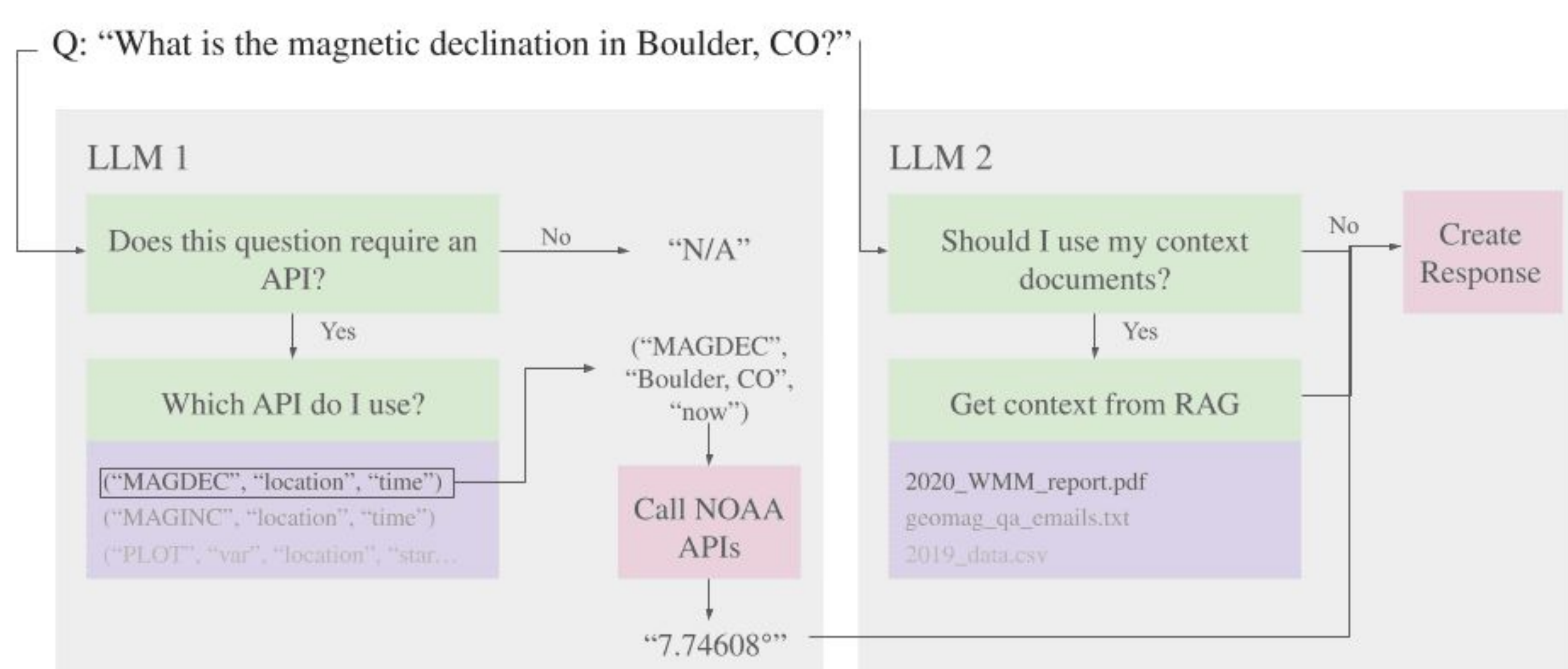


### 1. Introduction

The NOAA/CIRES geomagnetism group develops and distributes magnetic field models essential for navigation, research, and exploration, including the widely used World Magnetic Model. However, using these resources often requires domain-specific expertise. To address this, we developed ChatGeo-Magi, a conversational agent that provides a simpler way to query geomagnetic data, perform calculations, and generate visualizations. We compare two system architectures, a two-LLM approach and an agentic framework.

### 2. Two Approaches

The first uses a two-LLM approach that separates decision-making from response generation to improve reliability in handling user queries. The second uses an agentic framework that allows the system to reason through tasks in multiple steps and adapt its actions based on intermediate results. This comparison helps evaluate tradeoffs between reliability and flexibility in interacting with geomagnetic resources.

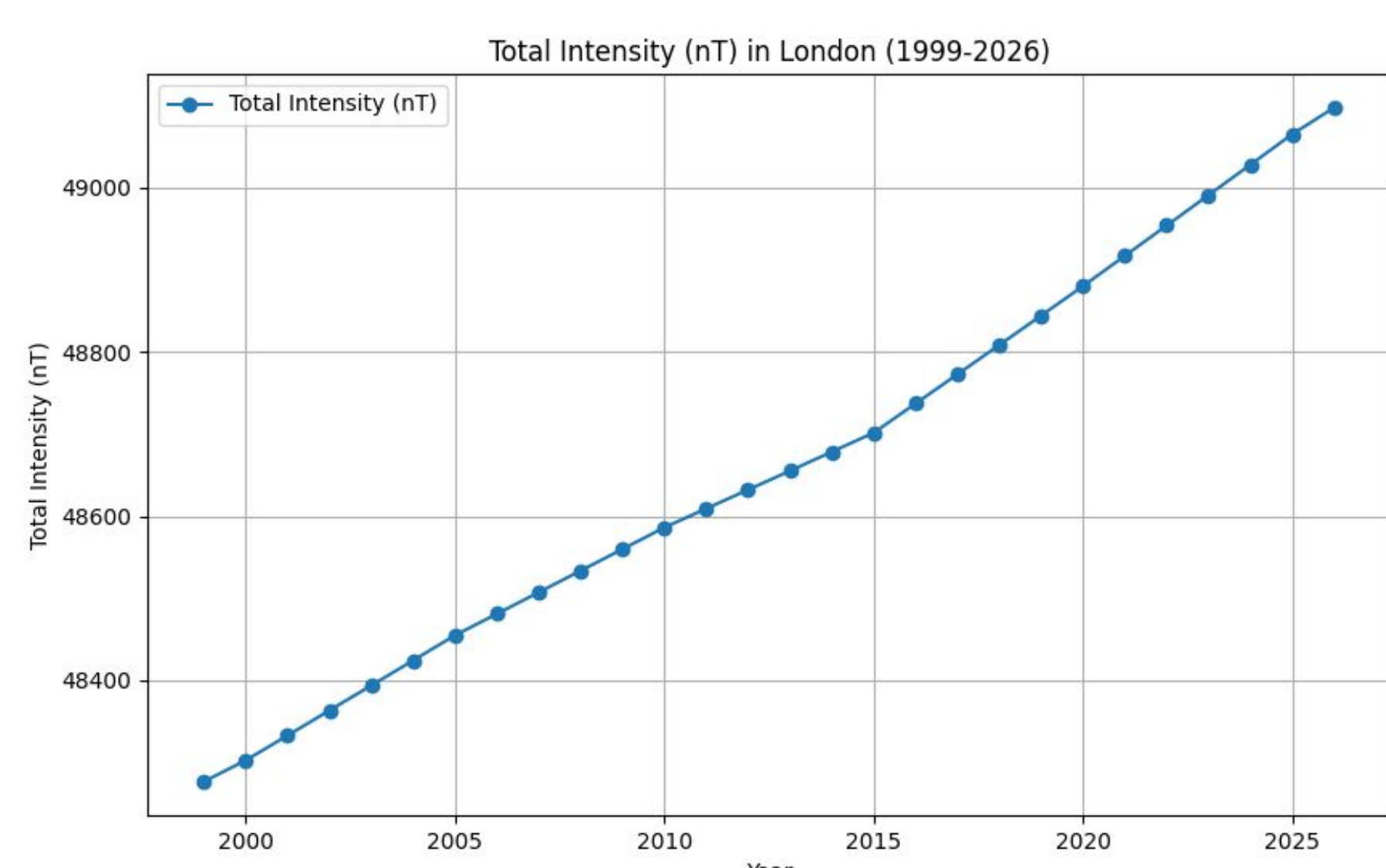


A: According to NOAA geomagnetic calculator, the magnetic declination in Boulder, CO is 7.74608° E

### 5. Qualitative Example

**Prompt:** Could you make me a plot of the magnetic field strength in London from 1999 until now?

**Response:** Here's the time-series plot of the total magnetic field intensity (total intensity) measured at London (51.5074 N, 0.1278 W) from 1999 up to the present. Feel free to let me know if you'd like additional statistics (e.g., trend slope, annual rates) or if you'd like to see other components such as declination or inclination.



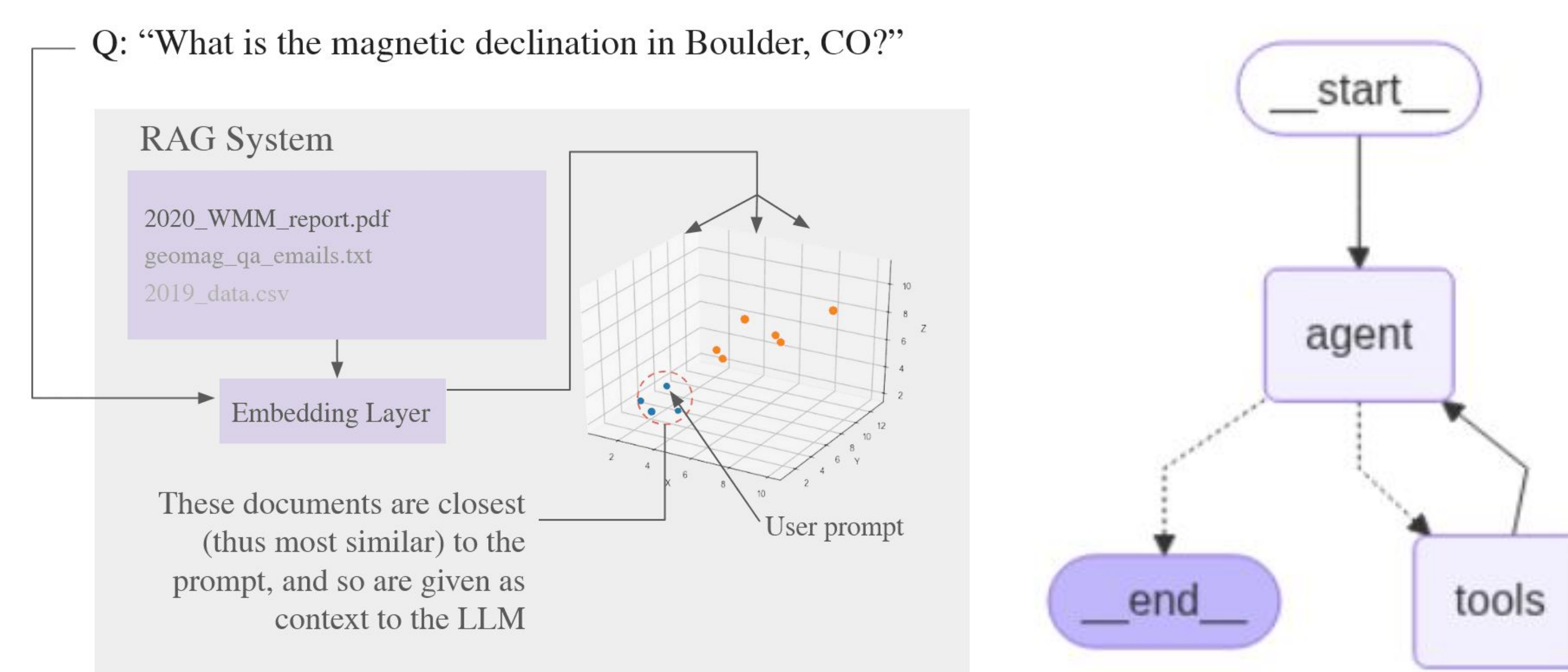
- Agent breaks the query into steps using the tools
- Capable of multi-step workflows, but final interpretation can introduce numerical errors

#### Tool Usage:

```
1.noaa_mag_api(data_name: "totalintensity", location: "51.5074,-0.1278", start_time: "1999-01-01", end_time: "2026-03-08")
2.plot(title: "Total Intensity (nT) in London (1999-2026)", x: years, y: intensity_values)
```

### 3. Techniques Used

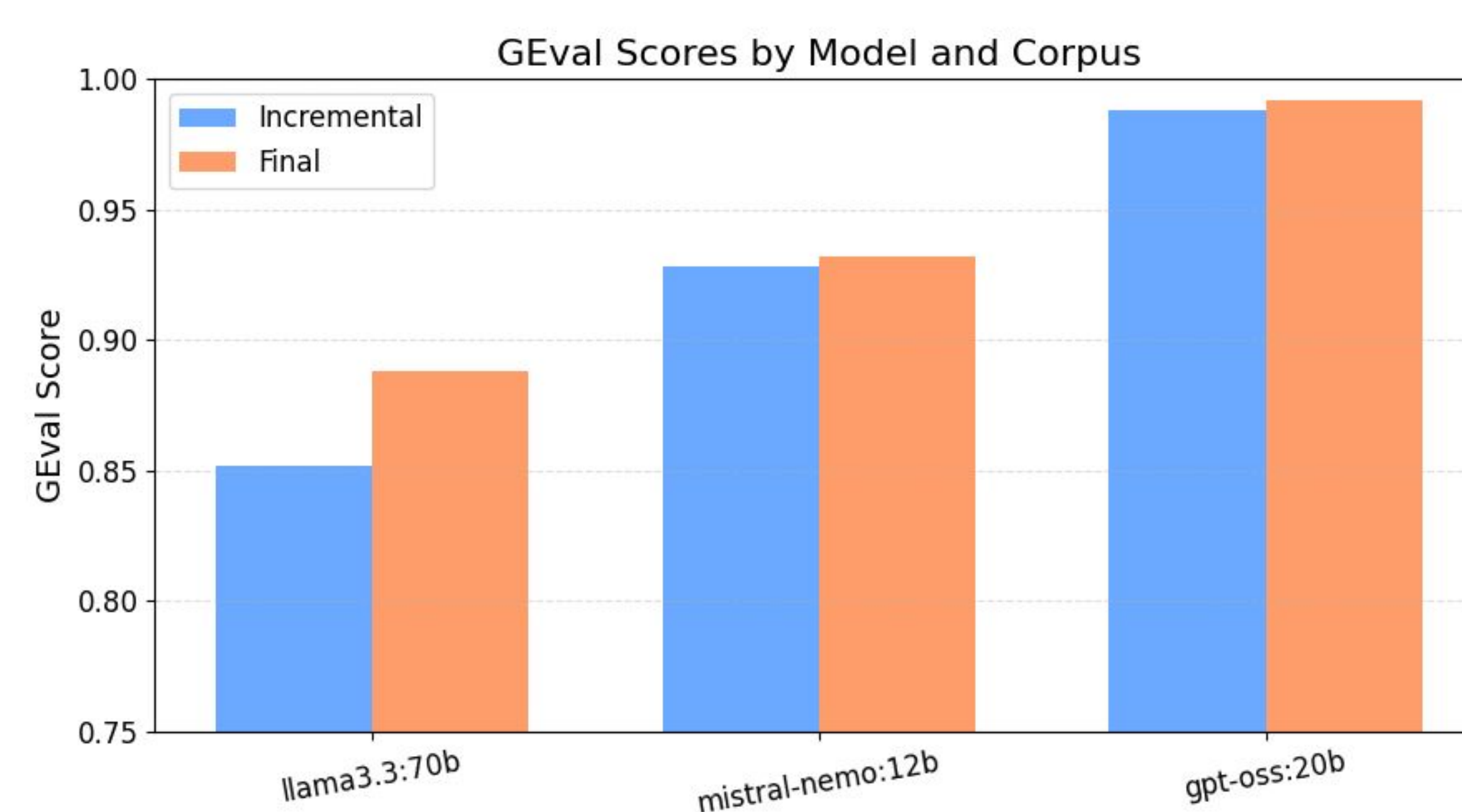
- **Retrieval Augmented Generation (RAG)** retrieves relevant documents and injects them into the model prompt
- Uses embeddings and vector search to match user queries with the **most relevant data chunks**



- **Tool Calling** converts user queries into structured calls for APIs, calculations, and plotting
- Two-LLM approach uses a single, fixed tool call per query, while the agentic approach supports multi-step tool use and iterative reasoning

### 4. Evaluation

An evaluation dataset was created using gpt-4o-mini to generate goldens, which are reference answers produced from documents in the RAG data. Model responses were then compared to these goldens using Opik, which scores correctness and factual alignment using gpt-4o-mini. We also tested the effect of adding proprietary support emails to the dataset to evaluate how different data sources impact system performance



Model performance depended more on the underlying model and retrieval quality than on simply increasing the amount of data. However, the agentic system proved better than the two-LLM, as it allowed for more complex tool and reasoning CoTs.

### 6. Limitations & Future Work

The evaluation used a small dataset of GPT-4o-mini-generated goldens and relied on automated scoring, which may not fully reflect real-world performance. All models were run locally, limiting results to smaller open-weight models. Additionally, the system was evaluated mainly on question answering rather than full end-to-end task completion. Future work includes human evaluation with domain experts, improving retrieval quality, testing larger models, and reducing computational costs.