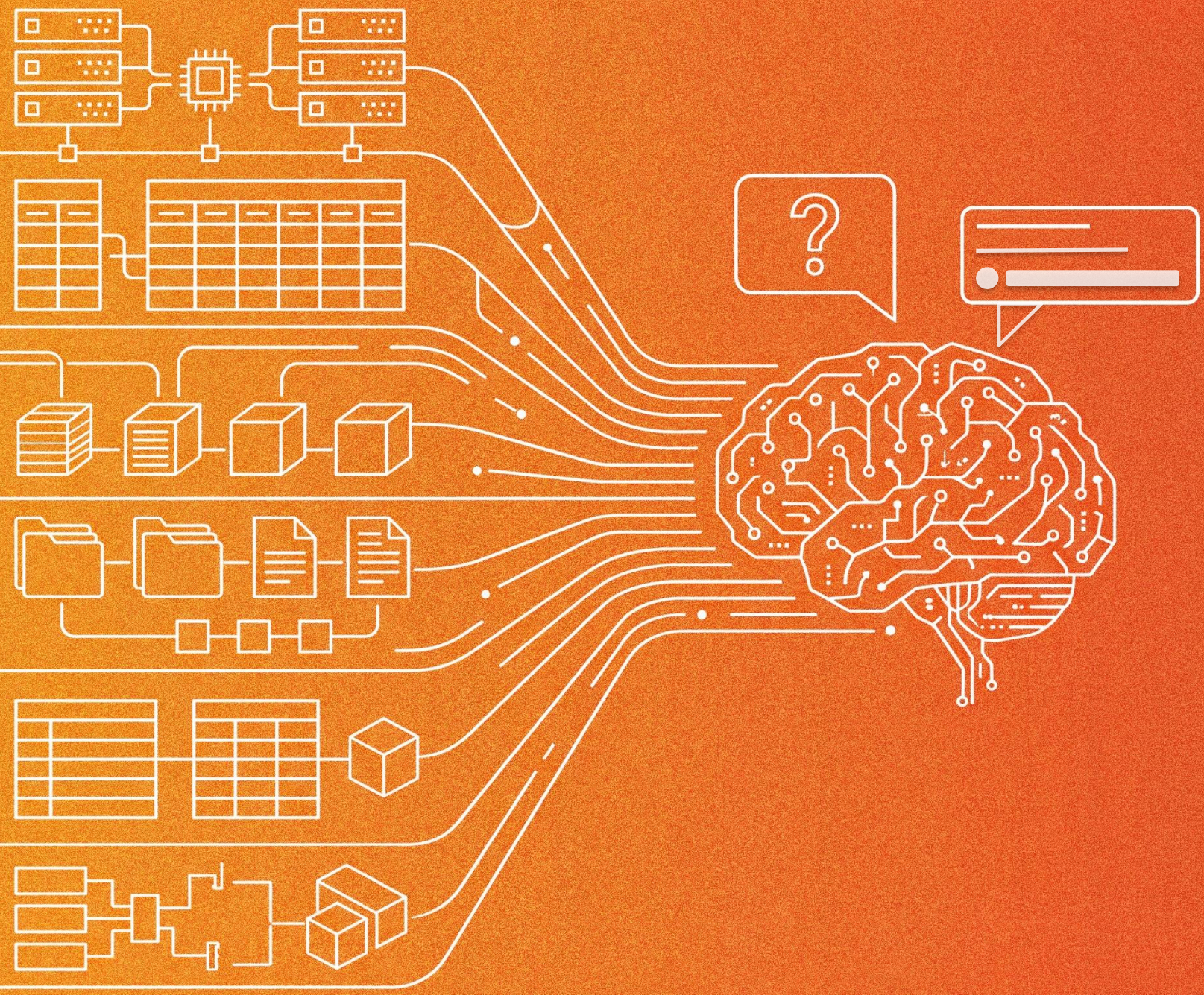
Optimizing Data for AI Chatbots



January 2026

Optimizing Data for AI Chatbots

By: Ralfs Rudzitis (LinkedIn) and Andrej Verity (LinkedIn)

Introduction

While AI chatbots can process vast amounts of information, they often lack the human-like ability to spot missing values, extreme outliers, or logical contradictions. Recent research indicates that even the most advanced AI models can see a **performance drop of up to 59% when moving from perfectly clean data to datasets with common real-world imperfections**. This report outlines practical guidelines for data providers to help increase the probability that their (downloadable) datasets can be better understood by AI, thereby reducing the risk of automated errors in high-stakes environments.

Preparing Your Data for AI[1]

1	**Limit the Size**	Keep dataset size under 25MB to prevent the AI from making mistakes or crashing.
2	**Clear the Grid**	Each sheet should contain only one table with no blank rows or merged cells.
3	**Name Things Clearly**	Use full words (e.g., Donor_Organization instead of DnrOrg). Use underscores or blanks instead of dots.
4	**Wide Formatting**	AI chatbots are more accurate on files with more columns and fewer rows.
5	**Provide Instructions**	If the data is provided in a .xlsx file, create a 2nd sheet that includes a short explanation of the column names & give some context about the data.

[1] Based on internal testing and RADAR research paper: https://arxiv.org/abs/2506.08249

Empowering the End-User

Beyond cleaning the data, providing guidance to the people analysing your data can significantly improve outcomes. Encourage users to follow these strategies:

1. **Setting Roles and Goals**
    - Users should not simply upload a file and ask, "What's in this?" They need to tell the AI what the data represents and exactly what they want to clarify/analyze.
    - Example: *"Act as a senior financial analyst. I want you to identify the top three trends in our donor growth over the last quarter."*

2. **Requesting Data Overviews**
    - Before asking for complex calculations, users should ask the AI to summarize column names and row counts to ensure it hasn't "hallucinated" non-existent data.

3. **Utilizing "Chain-of-Thought" Prompting**
    - For complex math, asking the AI to "show its work step-by-step" forces it to process data methodically rather than jumping to incorrect conclusions.

4. **Provide instructions**
    - In case the dataset does not come with a tab including explanations for column names and abbreviations, the user should add some instructions at the end of their input prompt (e.g., Here is a rundown of the important columns in the dataset: "TFnd" = Total Funding Given in a year).

5. **Respect the Context Window**
    - While the table above mentions limiting file size, remind users that chat history also takes up space. If they've been chatting for a long time, the AI might "forget" earlier parts of the data. Starting a fresh chat for a new complex analysis is often necessary.

Provide sample prompts to your users

Example prompt for AI Chatbot Data Analysis:

> *I have uploaded a dataset regarding [Topic, e.g., Donor Contributions 2025].*
>
> *Your Role: Act as a [Role, e.g., Data Scientist/Marketing Consultant].*
> *The Data: Note that [Abbreviation] stands for [Full Word].*
> *Please ignore any rows where [Column Name] is blank.*
> *Your Task:*
> *1. First, provide a brief 2-sentence summary of what this data contains to confirm you've read it.*
> *2. Then, calculate the [Specific Metric] and tell me which [Category] performed the best.*
> *3. Please show your reasoning step-by-step.*

Benefits vs. Limitations of Using AI Chatbots for Data Analysis[2]

Why end-users might rely on AI chatbots for data analysis and why that concerns you.

Benefits	Limitations
Efficiency: Can perform complex data analysis in seconds.	**Inconsistency:** The same AI might give two different answers to the same question if the data is messy.
Multi-Modality: It can filter, perform calculations, reformat, and generate clean summary tables or charts ready for a report.	**Scaling Issues:** The bigger the tables, the more problems AI chatbots have with processing and analysis.
Accessibility: Allows non-technical users to understand complex datasets.	**Artifact Sensitivity:** Extreme outliers or "bad values" (like using "-1" for missing data) can skew all AI generated results.

[2] How AI can make data analysis more efficient: https://www.youtube.com/watch?v=KXqR6VuPBlk and RADAR research paper https://arxiv.org/abs/2506.08249

Why do AI Chatbots Misinterpret Data?

AI chatbots process data using two primary methods, each with its own shortcomings:

- **The Calculator Mode[3]:** AI writes a computer script to analyze your file. It is more mathematically precise but also more blind when it comes to understanding the context of the data, which often leads to ignoring impossible values (e.g., an age of 200).
- **The Reader Mode[4]:** AI reads the data like a book. It has a better contextual understanding but suffers memory fatigue, causing it to ignore rows or invent numbers in large files.

Note: As of 2025, modern chatbots typically employ a hybrid approach combining both modes. Based on our testing, ChatGPT leans toward the Calculator mode for precision, while Gemini favors the Reader mode for context.

Current AI Chatbot Data Analysis Capabilities[5]

Frontier AI chatbot model performance with regards to data analysis and processing.

Tool (2025)	Performance	Observations
Microsoft Copilot chat (free)	Medium	Often provides incorrect totals on large files (e.g., miscalculating multilateral funding).
Google Gemini 3 Pro (free)	High	Currently the most robust at handling large OCHA datasets correctly on the first attempt.
ChatGPT 5 (free)	Low	Strictly fails on files over ~30MB, despite advertised higher limits.
Claude 4.5 (free)	Medium	Has problems with files over 25MB, however successfully gives correct answers once the dataset is trimmed.
DeepSeek V3.2 (free)	Low	Failed to process files at 35MB and 25MB due to persistent memory/RAM limits.

[3] How ChatGPT processes data: https://help.openai.com/en/articles/8437071-data-analysis-with-chatgpt More in-depth guide on ChatGPT processing data: https://medium.com/@cubode/comprehensive-guide-using-ai-agents-toanalyze-and-process-csv-data-a0259e2af761

[4] How Google Gemini processes data: https://blog.google/technology/ai/google-gemini-next-generation-model-february2024/#sundar-note

5 Based on our internal testing of free version models. For more rigorous testing (for paid version models), see page 8: https://arxiv.org/abs/2506.08249

Checklist: Ensure data is AI Chatbot ready

1. **Table Structure & Dimensions**
    - **[ ] Is it "Wide" vs. "Long"?** Models perform better on tables with more columns and fewer rows.
    - **[ ] One Table Per Sheet:** Does each sheet contain only one distinct table?
    - **[ ] Sheet Volume:** Does the dataset include many sheets?
    - **[ ] Header Placement:** Are column names in the very first row?
    - **[ ] Clean Headers:** Can a non-expert reader understand the column headers?
    - **[ ] No Merged Cells:** Have merged cells been removed?
    - **[ ] File Size Check:** Is the file under 25MB? (If not, consider splitting it by category or year).

2. **Data Cleaning & "Artifact" Removal**
    - **[ ] Consistent Formatting:** Are values represented identically (e.g., "5km" vs "5 kilometres")?
    - **[ ] Logical Sanity:** Do the values make sense? (e.g., "No "start dates" after "end dates").
    - **[ ] No Placeholder "Bad Values":** Are there placeholders like -1, 999, or #REF! that could skew calculations?
    - **[ ] No Blank Values:** Have blank values been removed?
    - **[ ] Extreme Outliers:** Have contextually impossible values (e.g., age of 200) been flagged or removed?

3. **Naming & Documentation**
    - **[ ] Human-Readable Headers:** Do headers use full words and underscores (e.g., Donor Organization) instead of dots or cryptic abbreviations?
    - **[ ] The "Read Me" Tab:** Is there a sheet defining every column header and abbreviation?
    - **[ ] Core Schema:** Have unnecessary columns been removed to reduce the file complexity and size?

4. **Model Compatibility (Testing)**
    - **[ ] Tool Selection:** If the data is context-heavy and public, use Gemini (favours Reader mode).
    - **[ ] Calculation Check:** If the data requires complex math and is public, use ChatGPT (favours Calculator mode). Otherwise, ensure the user knows to ask for "step-by-step" reasoning.