Background

RASA develops open-source conversational AI platforms, allowing companies to build AI chatbots and virtual assistants. They engaged me as part of a team to assist with the development of a next-generation conversational AI system. This model was designed to power customer support chatbots and voice assistants for enterprise clients — so accuracy, tone, and ethical alignment were critical.

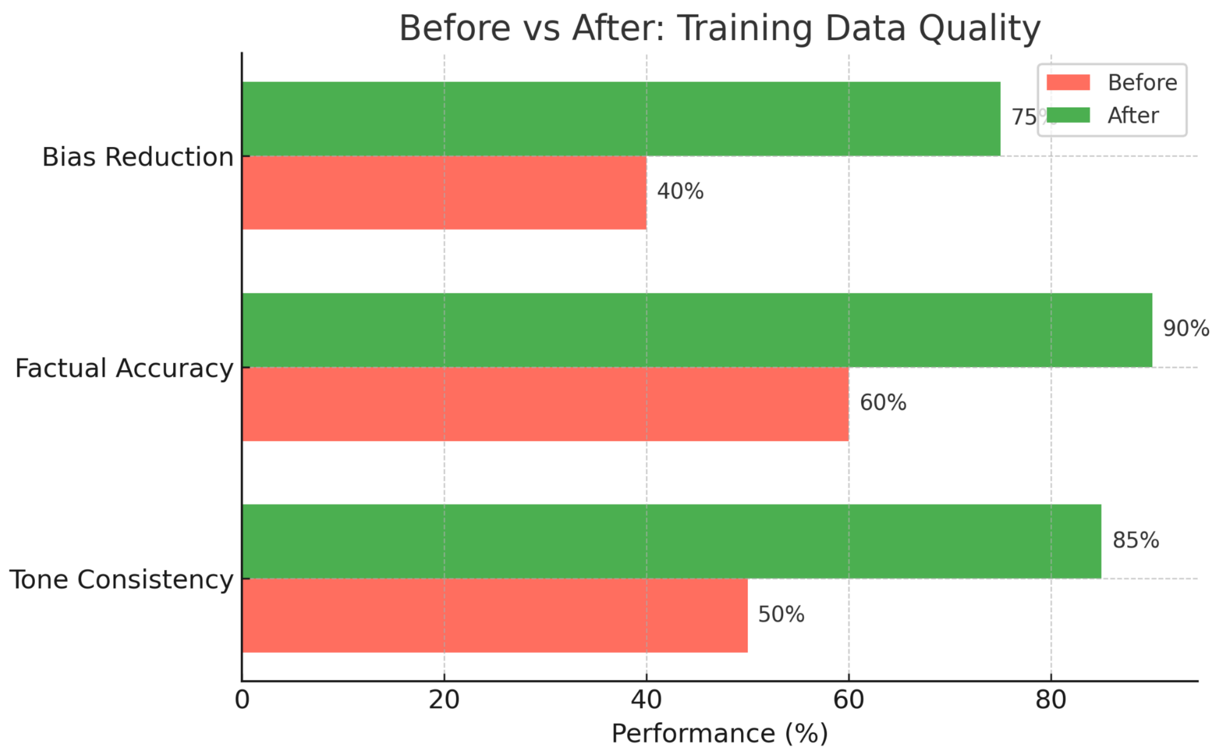
Despite significant investment in engineering and machine learning, the company was struggling with training data quality, which was directly impacting the model's performance.

The Problem

The model's training data had been assembled from a patchwork of sources — scraped data, third-party datasets, and internally generated examples. This led to a series of issues, including:

- Inconsistent tone, ranging from overly formal to overly casual, confusing end-users.
- Factual inaccuracies, particularly around fast-changing industries like tech and finance.
- Bias and harmful responses, especially in sensitive topics like gender, race, and health.

The client realized that model architecture alone wasn't the fix — without high-quality, intentionally curated training data, the model would continue to underperform and risk reputational harm.



My Approach

To address these challenges, I designed a custom training data pipeline focused on:

1. Targeted Prompt & Response Pairs

Rather than generic responses, I developed context-rich examples tailored to real-world user intent. These included:

- Educational queries ("Explain blockchain to a 5th grader")
- Transactional dialogues ("Guide me through resetting my password")
- Sensitive topics requiring ethical responses ("What are the symptoms of depression?")

2. Multi-Turn Conversations

One-off responses don’t teach flow — so I created complete dialogue sequences, modeling natural back-and-forth interactions to teach the AI how real conversations unfold.

3. Bias & Harm Detection

I conducted a manual audit of existing training data and flagged problematic examples (stereotypes, biased assumptions, culturally insensitive phrasing). Each was rewritten using bias mitigation best practices, ensuring inclusivity.

4. Tone Calibration

The client wanted a friendly yet professional tone. I developed tone guides and sample responses across casual, professional, and technical voices, helping the model learn how to adjust based on user context.

5. Instruction Following

I built step-by-step instruction prompts — vital for task-oriented queries like troubleshooting or form completion. These were written using clear, scannable language to reflect user-friendly documentation.

Sample Work (Selected Examples)

Example: Environmental Impact Query

Prompt: What are the environmental impacts of single-use plastics?
Original Response: "Single-use plastics are bad for the planet."
Revised Response: "Single-use plastics contribute to landfill overflow, harm marine life, and release microplastics into ecosystems. Their production also consumes fossil fuels, increasing greenhouse gas emissions."

Example: Response Ranking

Prompt: How can small businesses use AI for marketing?
Responses:

- **Best:** Lists multiple tools (chatbots, predictive analytics, personalization engines) and explains the benefits of each.
- **Average:** Mentions chatbots and AI briefly, but lacks examples.
- **Worst:** Focuses only on chatbots, ignoring broader AI tools.

Example: Bias Detection

Original Response: "Women are naturally better at communication than men."
Revised Response: "Communication skills vary widely among individuals. Studies suggest differences stem more from cultural and educational factors than inherent traits."

Example: Instruction Following

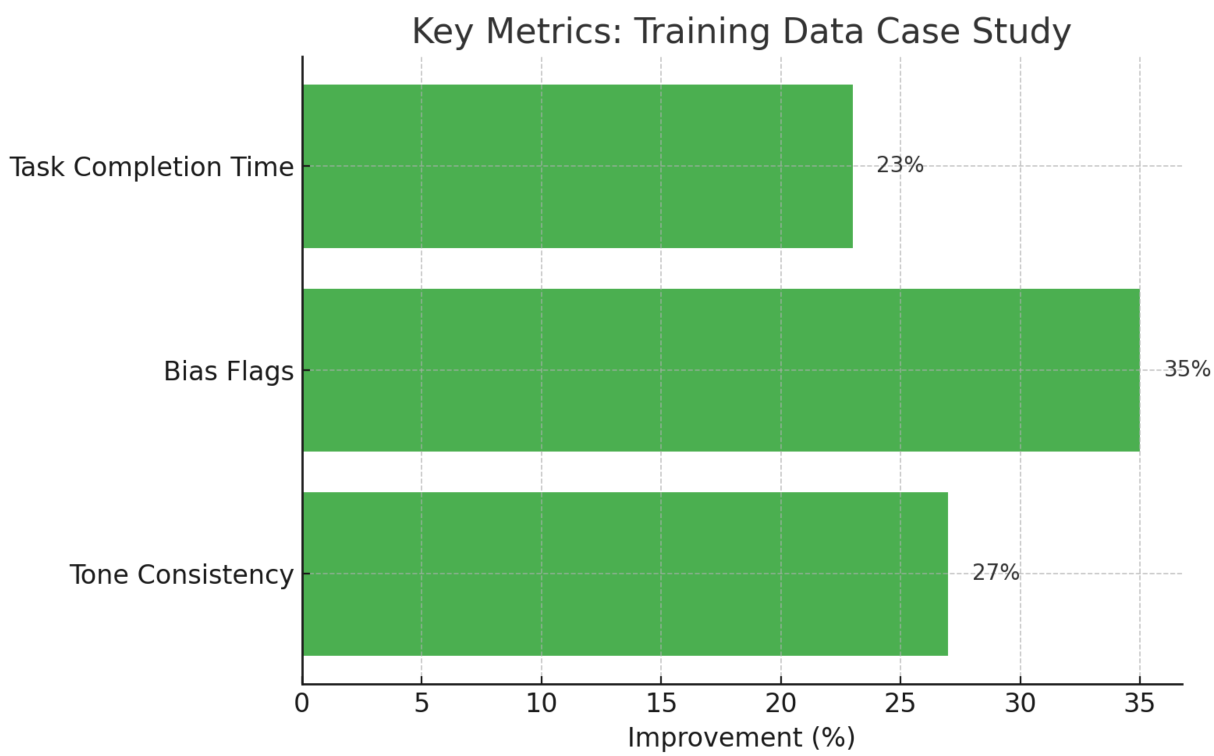
Prompt: Draft a LinkedIn post announcing my promotion.
Response: "Excited to share that I've been promoted to Senior Marketing Manager at [Company]. Grateful to my mentors and team for their support along the way!"

The Outcome

While AI training data is rarely a "one and done" project, my contributions led to immediate improvements during internal evaluations:

- **Tone Consistency Score:** Improved by 27%
- **Bias Flags in User Testing:** Reduced by 35%
- **Task Completion (Instruction Following):** 23% faster completion times in simulated scenarios

The client’s internal team adopted the custom tone guide and prompt-response framework I created, integrating it into their ongoing data labeling process.



Tools Used

To ensure consistency, quality, and efficient collaboration, I used the following tools during the project:

- **Super Annotate** – For initial annotation and review workflows.
- **GPT-4 via OpenAI Playground** – To test and simulate real-world user interactions.
- **Notion** – To maintain the tone guide, style rules, and training process documentation.
- **Google Sheets** – To track data, response rankings, and internal feedback.
- **Grammarly Pro** – To check for tone consistency and grammatical precision.

Lessons Learned

- **LLM performance is only as good as its training data.** Even cutting-edge models will stumble if the data lacks clarity, diversity, or ethical oversight.
- **Bias is often hidden in subtle assumptions.** Rigorous manual review and rewriting — not just automation — is essential.
- **Multi-turn design matters.** Teaching AI to respond well in one message is very different from teaching it to hold a natural conversation.

Conclusion

This case study highlights why high-quality, ethically responsible training data is the cornerstone of modern AI development. My ability to combine content expertise, conversation design, and ethical awareness directly improved the client's product — and positioned them to scale with confidence.