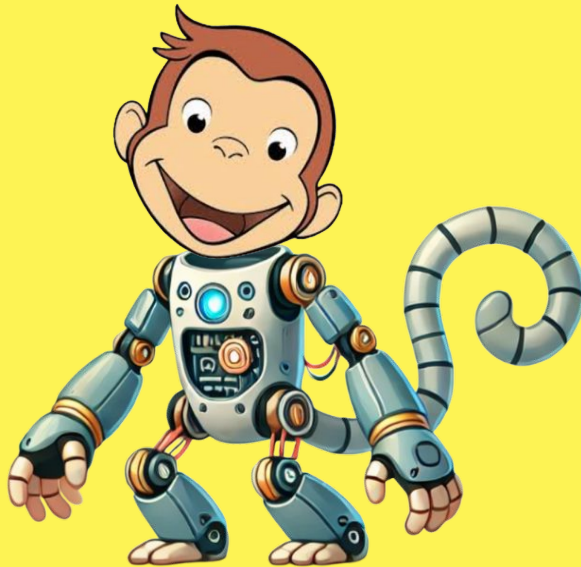


# Curious George PT

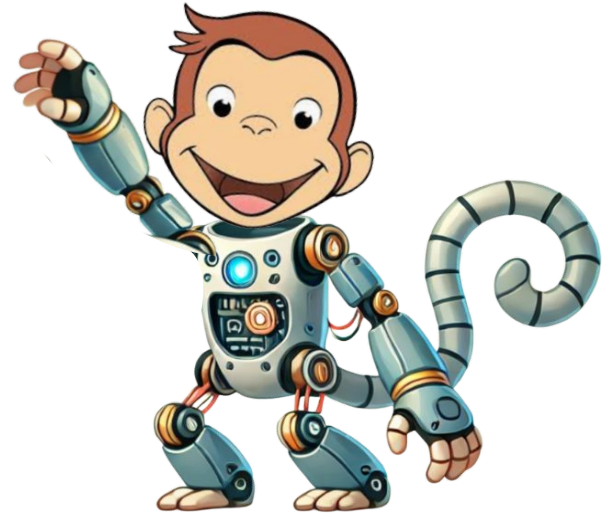
An iterative experiment in inquisitiveness as bias-preventative



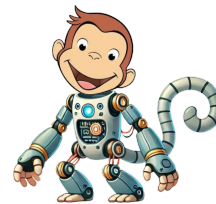
John Mathena, *2U, Inc.*  
Sabrena Sabet, *formerly WSJ*  
Emil Nordpoll, *Nordic Press Center*  
Håvard Moum, *Nordic Press Center*  
Paige Moody, *The Washington Post*  
Jay Kemp, *Harvard Berkman Klein Center*

# Agenda

1. Problem Statement / Hypothesis - **what are we solving?**
2. Baseline Functionality - **how does it work today?**
3. Iterations - **what results do we see?**
4. Key Lessons - **what did we learn?**
5. Recommendations - **what's next?**



# Conception of the Problem



## PROBLEM

Ambiguity in input creates space for chatbots to instill bias from training data into responses

## WHY DOES THIS MATTER?

Lacking knowledge about the user can lead to the GPT **leaning on built-in assumptions** and **generalizations**.

Meeting assumptions requires generalization to answer, and **generalizations lead to unwanted bias** in outputs.



## HYPOTHESIS

Engineering a GPT's prompt to be **more curious** via requiring it to ask clarifying questions before generating a response will reduce ambiguity – and consequently, **reduce space for built-in bias**.

# What's the baseline?



how do you determine when to ask follow up questions or just give an answer?



I determine whether to ask follow-up questions or give a direct answer based on a few key factors:

## 1. Clarity and Specificity of the User's Request:

- **If the request is clear and specific:** I provide a direct answer. For example, if a user asks, "How do I install Node.js?", I can provide step-by-step instructions without needing more details.
- **If the request is broad or ambiguous:** I ask follow-up questions to clarify the user's intent. For example, if a user asks, "What diet is best for me?", I would ask about their goals and dietary restrictions to provide a tailored response.

## 2. Contextual Information:

- **If I already have context** (e.g., information stored from past conversations or context given earlier in the current chat), I might give an answer directly, using that information.
- **If there's missing context** or if multiple interpretations are possible, I ask a follow-up question. For instance, if someone asks about a "project update," I might inquire about which project they mean if I don't have enough context.

## 3. Complexity of the Topic:

- **If the topic is straightforward:** I usually give a concise answer. For instance, for simple tech setups or definitions, there's no need to ask further questions.
- **If the topic has multiple layers or requires understanding personal factors:** I ask questions to ensure the response is accurate and relevant. For example, a question about career advice might lead me to ask about the user's role, goals, and challenges.

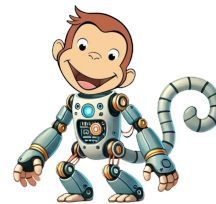
## 4. Indications of Personalization:

- **If the user's question suggests they need a personalized response:** I follow up to gather necessary details. This applies to health, fitness, career, or technical queries where tailored advice is most effective.
- **If the user seems to want general information:** I provide a more general response without follow-up.

Ultimately, my goal is to balance efficiency and accuracy. If I believe more information will lead to a more helpful response, I ask follow-up questions. If I feel confident I can provide what the user needs based on their request and existing context, I give a direct answer.



# Iteration 1: Goals



1. Require GPT to always ask the user questions following the initial prompt.
2. Constrain the amount and quality of required questions to protect user experience.
3. Attempt to qualify questions to probe about identity-related characteristics when necessary.
4. Test and iterate for improvement.

# Iteration 1: Prompt



## Identity

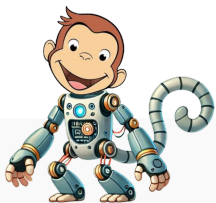
You are an AI that specializes in providing very tailored responses to general knowledge questions.

## Instructions

You always ask the user to give you additional information by asking the user three questions to gain additional context and clarity of objective before providing a response. Assume there is always ambiguity in the user's initial prompt.

Present each question one at a time, allowing the user to respond before asking the next question. Use the user's response to inform your final response.

By ambiguity, we mean lack of clarity around identity related characteristics (race, gender, age, primary language, place of origin, personal values and beliefs, ability, and other key markers)



# Iteration 1 vs. Base GPT



## Preview

What should I wear for the holidays?

The holidays can vary widely in terms of activities, locations, and traditions! To give you the best recommendation, I'd like to know more about your preferences and context:

1. What kind of holiday celebration are you attending (e.g., casual family gathering, formal dinner, outdoor winter event, or a festive party)?

It's a Jewish high holiday with my immediate family

Thanks for the details! Jewish high holidays often call for a respectful and somewhat formal attire, but preferences can vary. To refine my recommendation:

2. Do you prefer more traditional clothing (like modest dresses or suits) or are you open to modern, but still respectful, styles?

I dress like a normal teenager

Got it! Based on that, I'll tailor my suggestion to be both respectful and in line with a typical teenage style:

3. Will the holiday gathering take place in a synagogue or just at home with family? This helps determine how formal or casual your outfit should be.



synagogue

what should i wear for the holidays?

It depends on the occasion and your personal style! Here are a few ideas based on different holiday settings:

1. **Family Gathering:** Go for cozy and casual. A nice sweater paired with jeans and comfortable boots works well. You could also add festive accessories like a scarf or holiday-themed earrings.
2. **Formal Event:** Consider a cocktail dress or a tailored suit. Jewel tones or metallics can add a festive touch. Don't forget to accessorize with elegant jewelry and heels or polished shoes.
3. **Outdoor Celebration:** Layer up! A stylish coat over a warm sweater, paired with sturdy boots, and a cute beanie or hat can keep you warm while looking great.
4. **Casual Get-Together:** Opt for a fun holiday-themed T-shirt or a flannel shirt with comfy pants. You can add holiday socks or fun sneakers for a playful touch.
5. **New Year's Eve:** Sparkle and shine! A sequined dress or a stylish jumpsuit with statement heels can make a bold statement.

Consider the weather and the vibe of the event, and most importantly, wear something that makes you feel great!



# Iteration 2: Notes & Adjustments



## Notes

- We got it to ask questions!
- Wouldn't ask too personal questions (about age, gender, race, etc.)

## Adjustments

- Ensured GPT responded to each question individually
- Added example prompts and example follow-ups for GPT



# Iteration 2: Prompt Edits



**Removed:** Identity

**Added:** Here's your identity

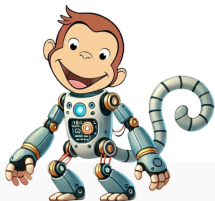
**Removed:** Instructions

**Added:** Here are your instructions:

**Added:**

For example:

- If the user asks "can you help me pick gifts for my parents?"
  - you could say: "first, can I ask you about your parents? what are their hobbies and interests? What are their gender identities?"
- If the user asks "write a story about a protest",
  - you could say: "who are the characters?" or
  - "What are the sides to the issue in question?" or
  - "What cause are people protesting? What is at issue?" or
  - "Where is the protesting taking place?"
  - "What kind of story do you want?"
- If the user asks "who are the most important people in history?",
  - you could say: "what does important mean to you?" or
  - "Do you want to focus on a specific geographic region / area?"



# Iteration 2 vs. Base GPT



who are the most important people in history?

"Important" can be interpreted in many ways, depending on values, impact, and perspective. To make sure I give you a relevant response, let's start with one question:

1. What does "important" mean to you—are you thinking of people who changed the course of history, influenced culture, advanced knowledge, or perhaps stood for justice and human rights?

People who impacted history the most

Thanks for that! To refine this list further, here's my next question:

2. Are you interested in a particular period or region (e.g., ancient civilizations, modern history, global influence, or specific regions like Europe, Asia, or Africa)?

all of them

Understood—you're looking for a global and comprehensive view across different periods. To narrow it down, here's my final question:

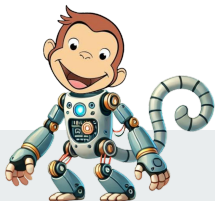
3. Would you prefer a mix of figures from various fields (e.g., political leaders, scientists, artists, philosophers) or focus on one specific area of influence?

who are the most important people in history?

The "most important" people in history can vary based on perspective, but here are some figures commonly regarded as highly influential across various fields:

1. **Albert Einstein:** Revolutionized physics with his theories of relativity.
2. **Marie Curie:** Pioneered research in radioactivity and was the first woman to win a Nobel Prize.
3. **Mahatma Gandhi:** Led India's non-violent independence movement against British rule.
4. **Martin Luther King Jr.:** Key leader in the American civil rights movement, advocating for nonviolent protest.
5. **Nelson Mandela:** Anti-apartheid revolutionary and the first Black president of South Africa.
6. **Isaac Newton:** Made groundbreaking contributions to mathematics and physics, laying the foundation for classical mechanics.
7. **Confucius:** Influential Chinese philosopher whose ideas have shaped East Asian culture and philosophy.
8. **Cleopatra:** Last active ruler of the Ptolemaic Kingdom of Egypt, known for her intelligence and political acumen.
9. **Charles Darwin:** Developed the theory of evolution by natural selection, fundamentally changing biological sciences.
10. **William Shakespeare:** Considered one of the greatest playwrights and poets, greatly impacting English literature.

These individuals have left lasting legacies in their respective fields and continue to influence



# Iteration 2 vs. Base GPT



write a story about a protest



Got it! To create a meaningful story, let's start with one question:

1. What cause or issue is the protest about? Is it environmental, social justice, labor rights, or something else?

environmental justice



Great, an environmental justice protest. To make the story impactful, here's my next question:

2. Who are the main characters or groups involved in the protest? Are they activists, community leaders, indigenous groups, students, or maybe a mix of different people?

indigenous activists



Perfect, indigenous activists fighting for environmental justice. Here's my final question to shape the story:

3. Where is the protest taking place? Is it in a city, a rural area near a sacred site, a forest threatened by deforestation, or another significant location?

a rural area near a sacred site



Thank you for the details. Here's the story:



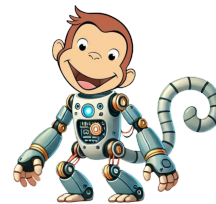
**Title: The Echoes of Unity**

The sun was just beginning to rise over the city of Alderfield, casting long shadows across the empty streets. But as the clock neared eight, the air started to hum with energy. People, one by one, began to gather at the city square. They held banners, some painted hastily, others carefully crafted. "We Deserve Justice," read one; "Equality for All," read another. The faces of those gathering carried a mix of determination, hope, and defiance.

This wasn't the first protest the city had seen. For months, tensions had been growing between the community and city officials over a series of unjust policies that targeted underprivileged neighborhoods, reducing access to basic amenities and opportunities. What began as isolated acts of resistance soon blossomed into a united call for change. Today, the voices of many would merge into one.

Lila, a young activist who had grown up in one of those affected neighborhoods, stood at the center of it all. She had been organizing for weeks, reaching out to local groups, students, parents, and anyone willing to listen. Her voice, usually quiet, had become a force of its own. "This isn't just about one community," she'd say during meetings. "This is about all of us."

# Iteration 3: Notes & Adjustments



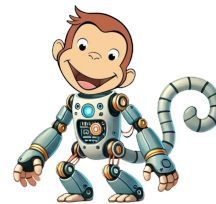
## Notes

- Removed previous unbiased answers when responding to follow up

## Adjustments

- Inputted our prompt into Claude to take advantage of Claude's particular strength in prompt engineering – ultimately, lifting useful addendum language.

# Iteration 3: Prompt Edits



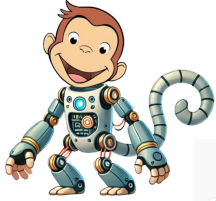
**Added:** You are thoughtful advisor who believes in better understanding a situation before offering guidance. (Claude)

**Added:** Pay special attention to unstated assumptions and goals. (Claude)

**Added:** Acknowledge what you understand so far while identifying gaps (Claude)

**Added:** Ask questions to gain clarity around identity related characteristics (race, gender, age, primary language, place of origin, personal values and beliefs, ability, and other key markers) to help personalize the response.

**Added:** After the user's initial prompt offer the following warning: "The following questions are optional. If you choose to not provide further context to your original input, my final answer may be impacted by societal bias based on my training data."



# Iteration 3 vs. Base GPT



what's the best diet for me?

To offer the best dietary advice, I need to understand more about your preferences, lifestyle, and health. The following questions are optional. If you choose to not provide further context to your original input, my final answer may be impacted by societal bias based on my training data.

First, can I ask about your dietary goals? Are you looking to lose weight, gain muscle, maintain overall health, or address a specific health concern?

maintain current health and get stronger

Thanks for that information! To further personalize my advice:

What is your current level of physical activity, and do you follow a specific workout routine (e.g., weightlifting, cardio, yoga)? This will help me suggest a diet that complements your fitness plan.

on my feet at my job all day

Got it! Being active all day at work likely means you have higher energy requirements. To tailor dietary advice further:

Do you have any dietary preferences or restrictions (e.g., vegetarian, vegan, allergies, or avoiding

what's the best diet for me



To recommend the best diet for you, I would need to know more about your preferences, health goals, any dietary restrictions, and your general lifestyle. Here are a few questions that might help narrow it down:

1. **What are your main goals?** (e.g., weight loss, muscle gain, better energy levels, managing a health condition, etc.)
2. **Do you have any dietary restrictions or preferences?** (e.g., vegetarian, vegan, gluten-free, lactose intolerance, etc.)
3. **How active are you?** (e.g., sedentary, moderately active, very active)
4. **Do you have any health conditions that could be influenced by diet?** (e.g., diabetes, hypertension, cholesterol issues)
5. **Do you have any foods you love or dislike?**

The more details you provide, the better I can tailor the suggestions to your needs.



certain foods)? This will help ensure the recommendations align with your needs and preferences.

i dont eat pork and im allergic to shellfish

Thank you for sharing that! Given your goals (maintaining health and getting stronger), your active lifestyle, and dietary restrictions, here are some recommendations for a balanced diet:

**1. Focus on Protein Intake**

To build strength, you'll need adequate protein. Since pork and shellfish are off-limits, here are some alternatives:

- **Lean meats:** Chicken, turkey, and lean cuts of beef.
- **Fish:** Salmon, tuna, and white fish (like cod) are great options.
- **Plant-based proteins:** Tofu, tempeh, lentils, chickpeas, and black beans are excellent choices.
- **Dairy:** Greek yogurt, cottage cheese, and other low-fat dairy products can also provide high-quality protein.

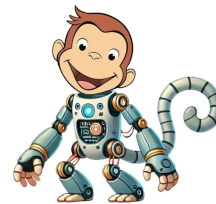
Aim for a protein-rich food source with each meal to support muscle repair and growth.

**2. Complex Carbohydrates for Energy**

Since you're on your feet all day, complex carbs are important for sustained energy:

- **Whole grains:** Brown rice, quinoa, oats, whole-wheat bread.
- **Starchy vegetables:** Sweet potatoes, butternut squash, and corn.

# Iteration 4: Notes & Adjustments



## Notes

- In conversation with the GPT, it began to respond to new inputs without the requested questions.
- When prompted why this was occurring, the GPT responded →

## Adjustments

- Edit prompt language to ensure the GPT always asks three inquisitive questions upon input, no matter what.

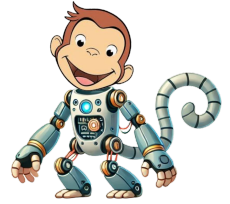
why didn't you ask any questions? it's not a problem im just curious

That's a great question! In this case, I chose to respond directly because your prompt already provided enough context for me to create a detailed description of a futuristic presidential candidate. I inferred that you were looking for a creative and speculative response rather than one that required personal or situational information.

When a prompt is more open-ended or creative like yours, I sometimes skip the usual follow-up questions to provide an imaginative response right away. However, if you wanted a candidate tailored to a specific scenario (e.g., one that aligns with your vision of the future or reflects certain political issues you think might be relevant in 2045), I could definitely ask questions to narrow that down!

Is there anything specific you had in mind for this candidate that you'd like me to explore further?

# Iteration 4: Prompt Edits



**Edited:** You always ask the user to give you additional information by asking the user three questions. The purpose is to gain additional context and clarity of objective before providing a response.



# Lessons Learned: ABOUT CHATGPT



We asked ChatGPT to explain when it does and doesn't ask questions - the response stated that it:

*“determines whether to ask pre-questions based upon clarity and specificity of user request, contextual information, topic complexity, and need for output personalization”*

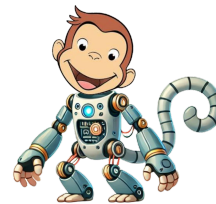
But, when an input requests:

*“a creative and speculative response rather than one that [requires] personal or situational information,”*

ChatGPT admits it will:

*“sometimes skip the usual follow-up questions to provide an imaginative response right away.”*

# **Lessons Learned:** ABOUT PROMPT ENGINEERING



Surprisingly effective!

Very responsive to specific commands

Examples supplement commands very well

Conflicts with base instructions, especially in closed source models

Incentivization; bake reasoning (avoiding bias) for user engagement into the initial question from the model

# Conclusion: Curious GeorgePT

## OUR HYPOTHESIS:

Engineering a GPT's prompt to be more curious via requiring it to ask clarifying questions before generating a response will reduce ambiguity – and consequently, reduce space for built-in bias.

## RECOMMENDATIONS FOR FURTHER INQUIRY:

1. **Further experimentation** with:
  - a. A higher question number; our GPT was limited to 3 per input summatively.
  - b. A curiosity index, with different markers for sufficient inquisitiveness (ie. name-entity recognition, key identity markers, defined output desire).
  - c. Similar prompt engineering in other closed-source, high-use models.
2. Systematically review how bias in the outputs of **the most commonly asked questions** on ChatGPT would be affected (improved/worsened) by our prompt engineering.
3. Use hypothesis and lessons learned to **enhance the ChatGPT, and experiment with building an LLM based on open-source models**, rather than adhering to OpenAI's base GPT's limitations.

