Capstone Presentation, 12th June 2022

Trustworthy LLM Chatbot

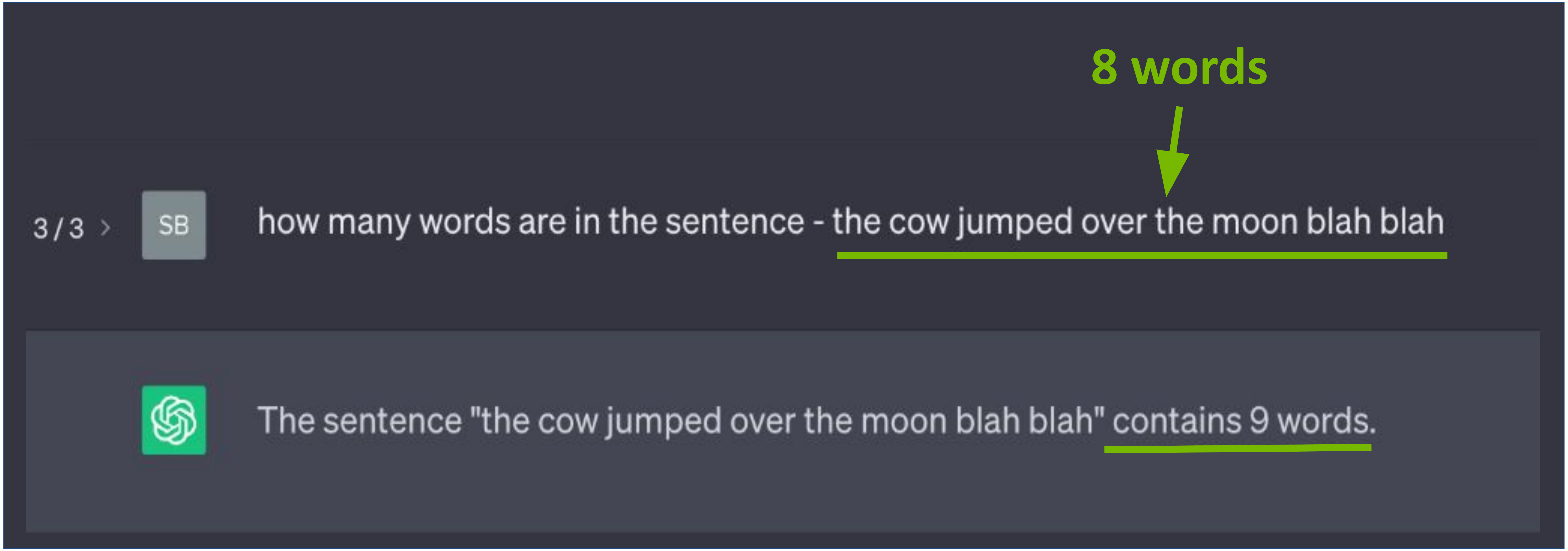
Shivani Birajdar

Yi Qing Khoo (Haru)

Jurine Gong

Tam Cao

Do you trust ChatGPT?



how many words are in the sentence - the cow jumped over the moon blah blah

The sentence "the cow jumped over the moon blah blah" contains 9 words.

Hallucination

Texts generated by AI-powered chatbots can be **factually incorrect, irrelevant, or nonsensical**

Negative impacts:

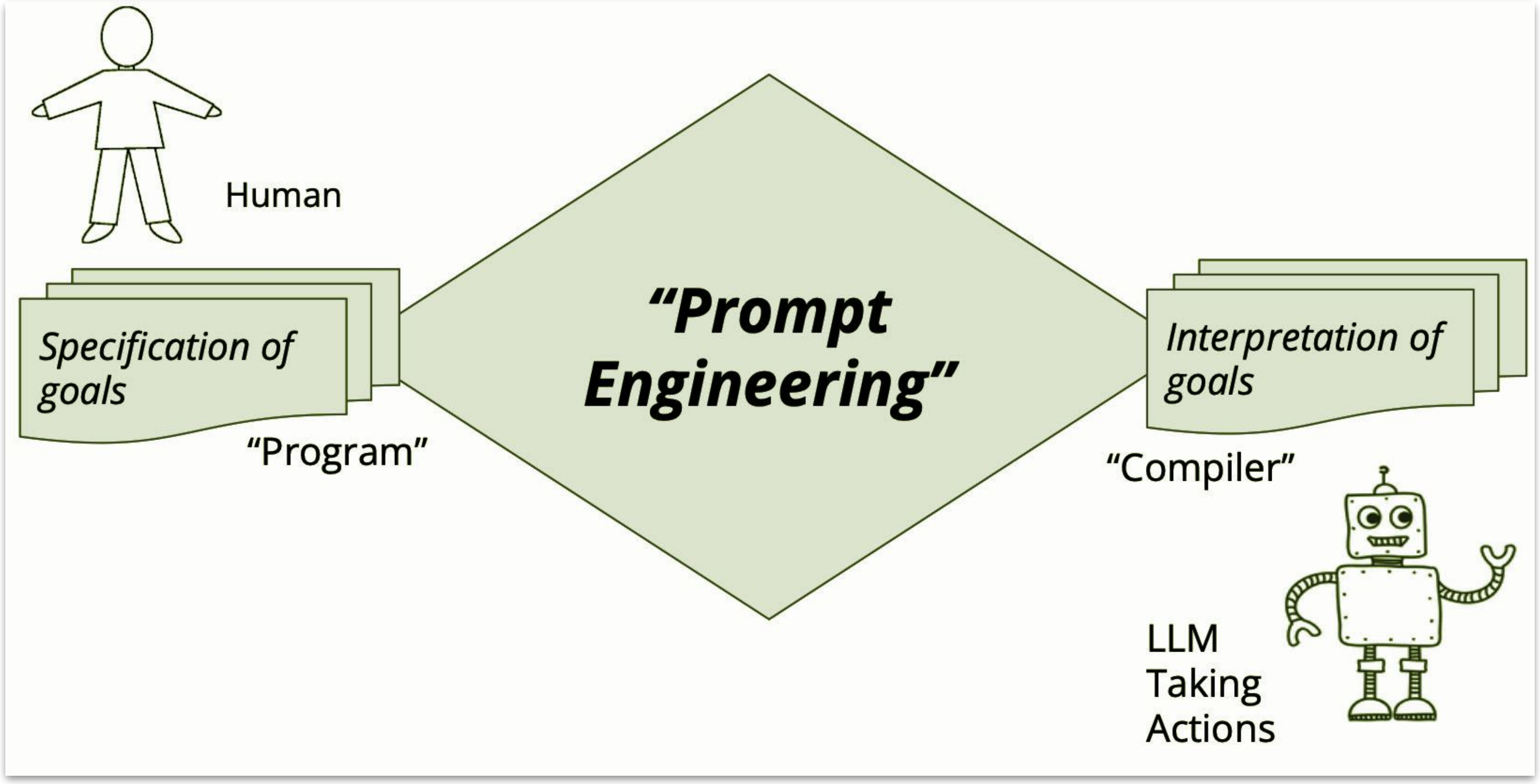
- **User Frustration and Dissatisfaction**
- **Misinformation Dissemination**
- **Communication Breakdown**
- **Damage to Brand Reputation**

Need for a trustworthy chatbot

Project scope focus: Corporate IT/HR support

*We aim to address the **"trust issue"** to build trust and confidence in the users through:*

Prompt Engineering



UX Designs

Research Questions

01. Users’ trust in AI chatbots

- What are users' expectations and current perceptions of chatbots as an IT/HR support solution?
- What factors contribute to users' trust in IT/HR chatbots, compared with other forms of IT/HR support (e.g., human representatives or self-help resources)?

02. Prompt and UX design to increase trustworthiness

- What type of system prompt patterns is most effective in enhancing trustworthiness in the context of IT/HR support?
- What UX design features of IT/HR support chatbots would be most likely to build and maintain users' trust over time?

03. Inclusive design

- What accessibility considerations should be included in the design of IT/HR support AI chatbots?

Research Methods & Overall Process

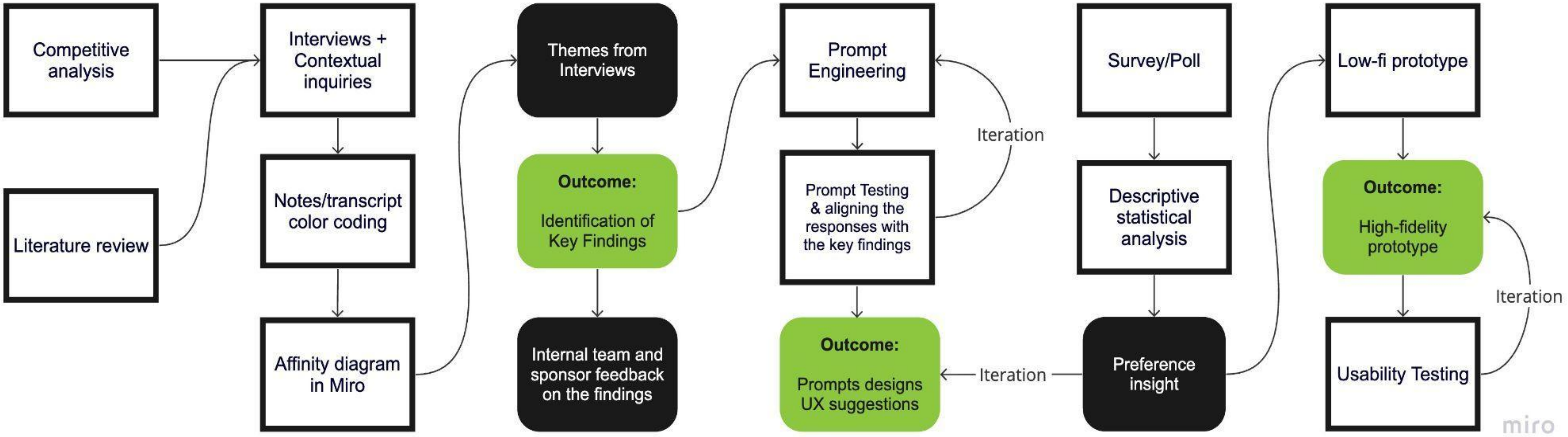
Spring quarter

Fall quarter

PHASE I - *Research*

PHASE II - *Prompt engineering*

PHASE III - *Prototype design*



Competitive analysis

Literature review

Interviews + Contextual inquiries

Notes/transcript color coding

Affinity diagram in Miro

Themes from Interviews

Outcome:

Identification of Key Findings

Internal team and sponsor feedback on the findings

Prompt Engineering

Prompt Testing & aligning the responses with the key findings

Iteration

Outcome:

Prompts designs UX suggestions

Iteration

Survey/Poll

Descriptive statistical analysis

Preference insight

Low-fi prototype

Outcome:

High-fidelity prototype

Iteration

Usability Testing

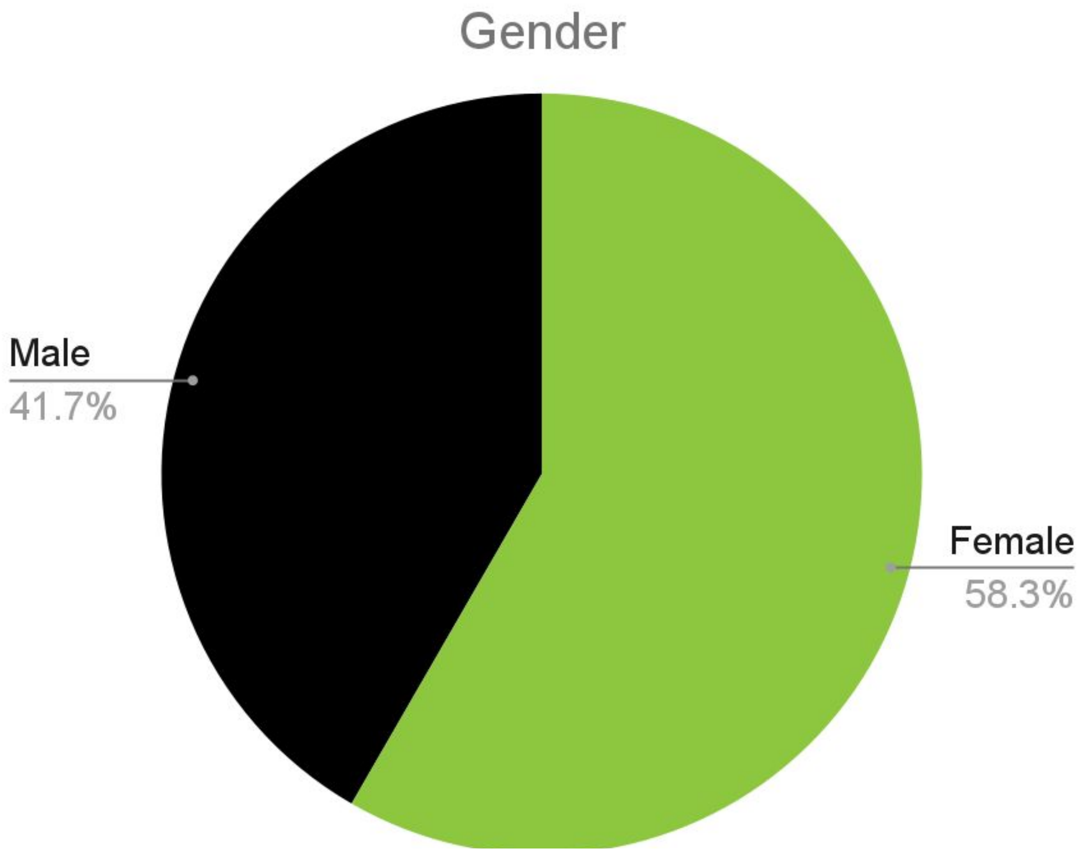
Participants

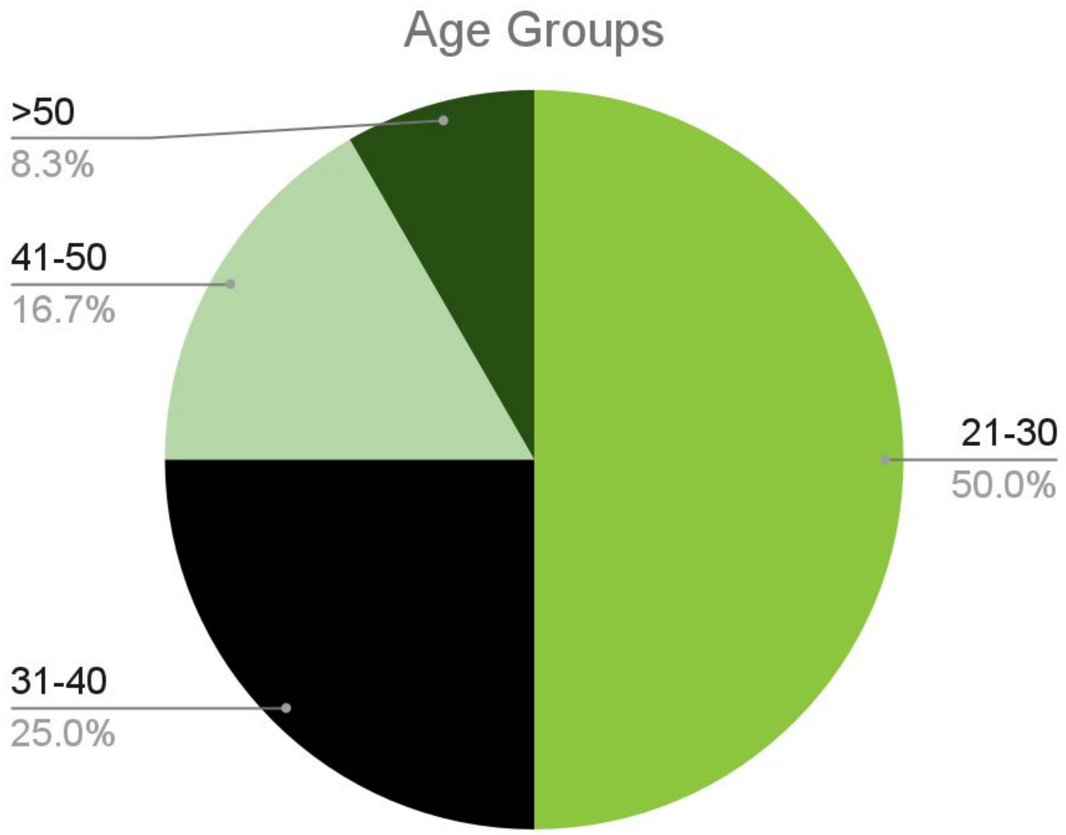
01. Recruitment Method

- Internal Connections: Total 12
    - Student Team:
        - 3 student pilot
        - 4 working in industry
    - Nvidia team: 5 Nvidia connections

02. Background

- Variety of occupations - (Engineers, Managers, Students)
- Diverse age groups and genders





Themes

Nvidia

Theme 1: Non AI chatbot (NVBot) limits user engagement

Key Finding 1: Little flexibility in input format leads to more work for users

“

P12

“...it was sort of ***frustrating*** *just doing even the programmatic task of updating a Jira ticket, like* ***you had to be very specific with the formatting****,* ***or it would,*** *you know,* ***update the wrong ticket*** *or not put the all the content that you wanted to update into the edit.”*

Theme 1: Non AI chatbot (NVBot) limits user engagement

Key Finding 2: Limited response formats result in the need for more time & effort to resolve problems

P11

“... just provide a ***hyperlink*** *but* ***provide some more context*** *and what is in there like summary of it that would help and then if a user can actually go to that link.”*
*“... **provide some more summary of what this article does** and that would be good enough. Rather than just going there. And it will help me as to whether it, it will not be useful or not right before clicking.”*

Theme 2: The challenge of trustworthiness in LLM chatbot

Key Finding 3: Lack of source verification negatively impacts user trust

P1

“Again, I don't have any real background in this area is that it is essentially doing a much smarter Google search than me.

*So ultimately, **you can't really trust the answers** further than you could like a **cursory Google search** without like actually doing a lot of verification and validation. **You should always be taking it kind of with a grain of salt.**”*

Theme 2: The challenge of trustworthiness in LLM chatbot

Key Finding 4: Incapability of being transparent and explainable increases users’ suspicion

“So this is something like this is right now a ***gray area*** *where in the* ***user doesn't exactly know that which part is accurate and which part is not accurate.****”*

P7

"I don’t know. I think I would trust an answer more if the bot actually lays out all of the ***reasoning and details*** *so that I can see* ***how it arrives at the solution****."*

P2

Theme 3: Going beyond a chatbot expanding their capabilities

Key Finding 5: Users expect the chatbot to implement its suggestions/instructions

P5

“...so in short, you would expect the ***chatbot to process*** *all those requests for you, for example,* ***reconnecting or change password*** *..”*

“...the ***chatbot be able to run a script at the backend*** *once I report it and then it should ask me to try and reconnect.”*

Theme 3: Going beyond a chatbot expanding their capabilities

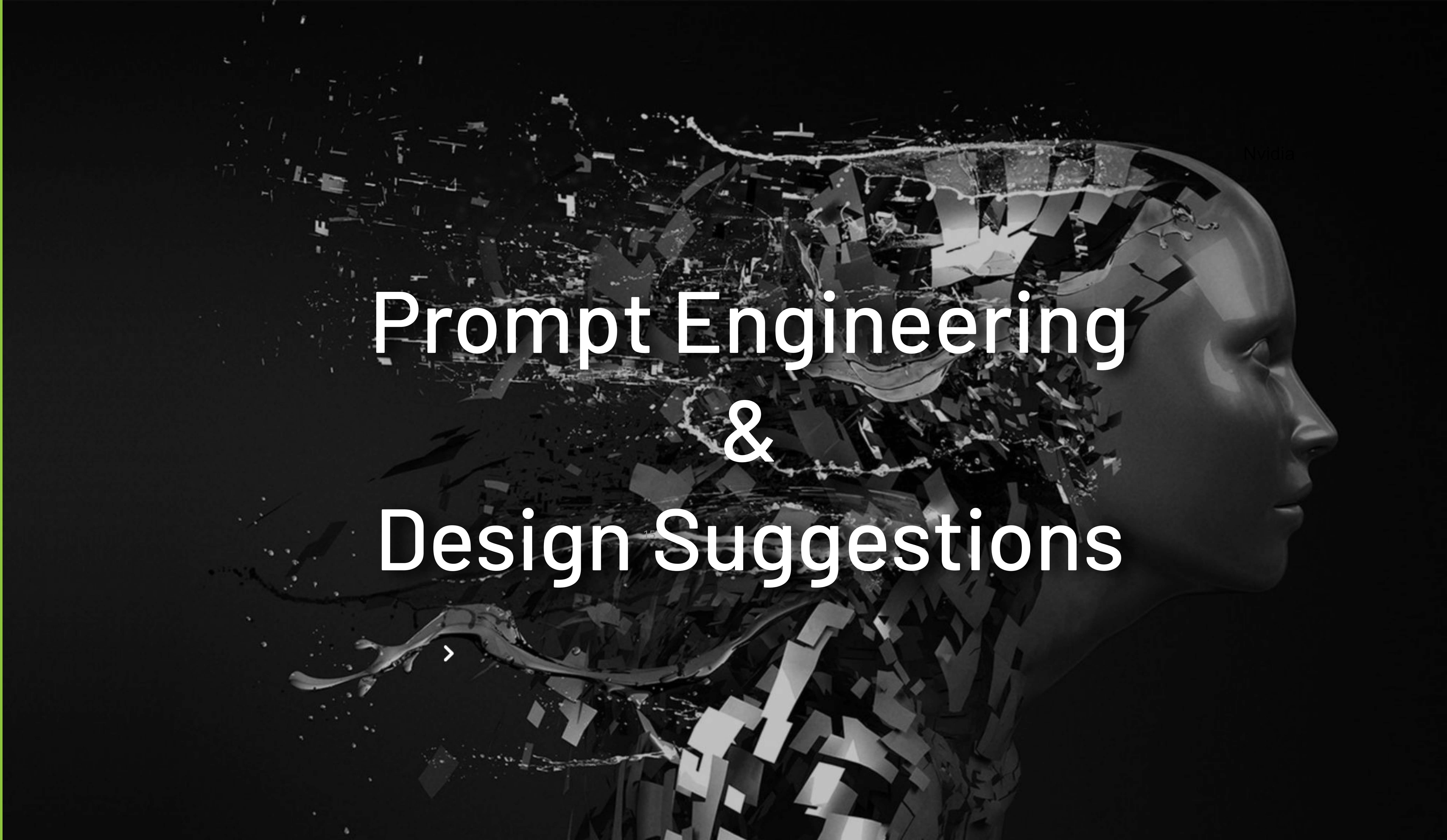
Key Finding 6: Users want assistance from the chatbot to help boost productivity.

> *"I guess if it could…look at the transcript of meeting notes and then* ***update ticket items automatically.****"*

P12

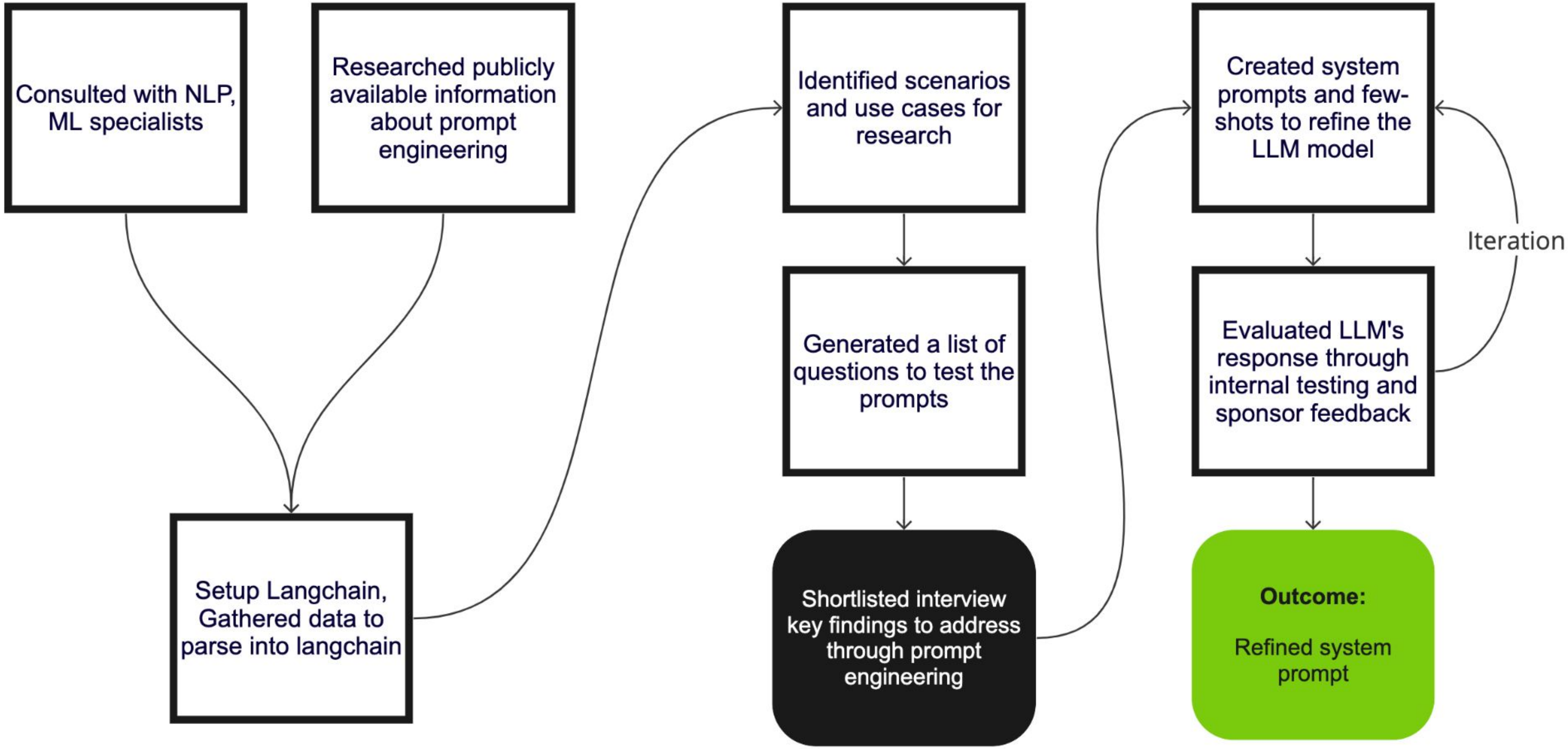
> *"I think* ***documenting your code****, that would be a helpful feature because I know for a fact what my function is doing, but* ***I don't want to spend time writing that down.****"*

P10



Prompt Engineering
&
Design Suggestions

Prompt Engineering Process



Consulted with NLP, ML specialists

Researched publicly available information about prompt engineering

Setup Langchain, Gathered data to parse into langchain

Identified scenarios and use cases for research

Generated a list of questions to test the prompts

Shortlisted interview key findings to address through prompt engineering

Created system prompts and few-shots to refine the LLM model

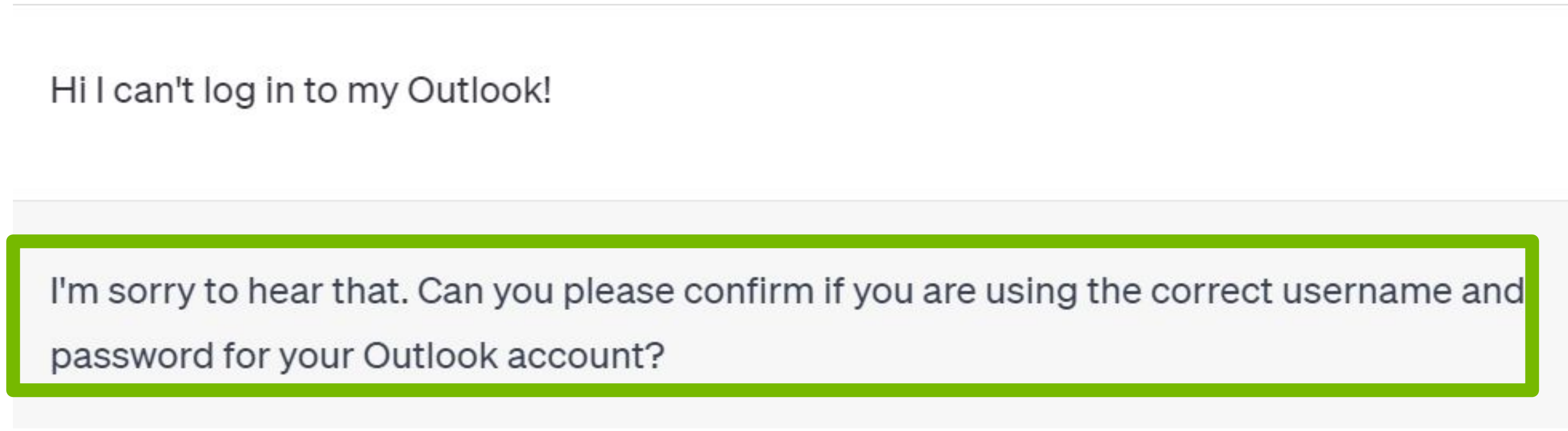
Evaluated LLM's response through internal testing and sponsor feedback

Iteration

Outcome:

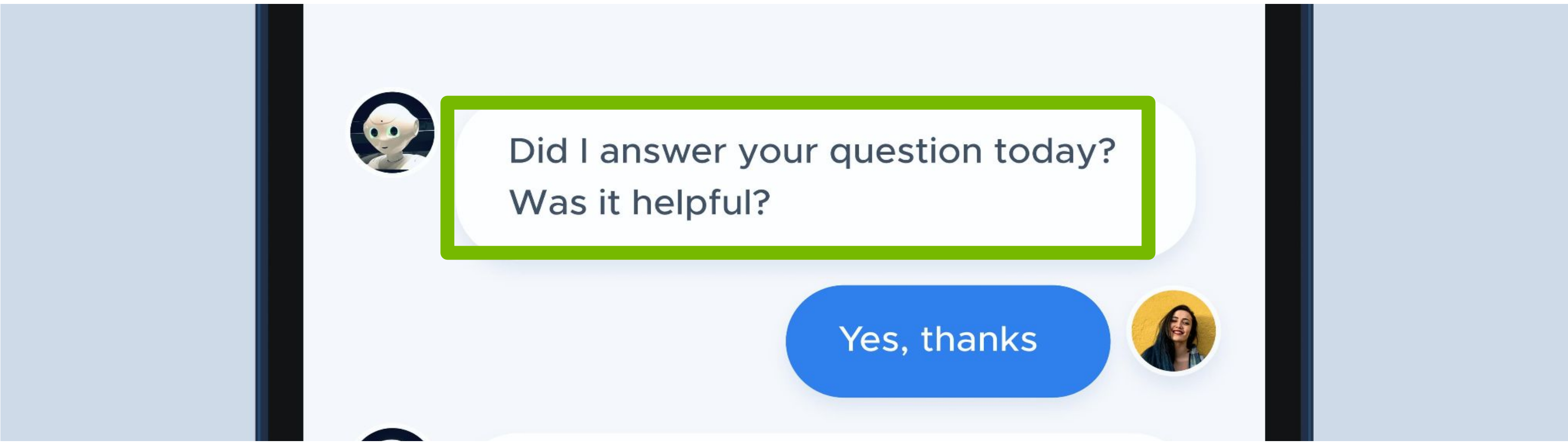
Refined system prompt

Design Suggestion 1: Humanlike conversation

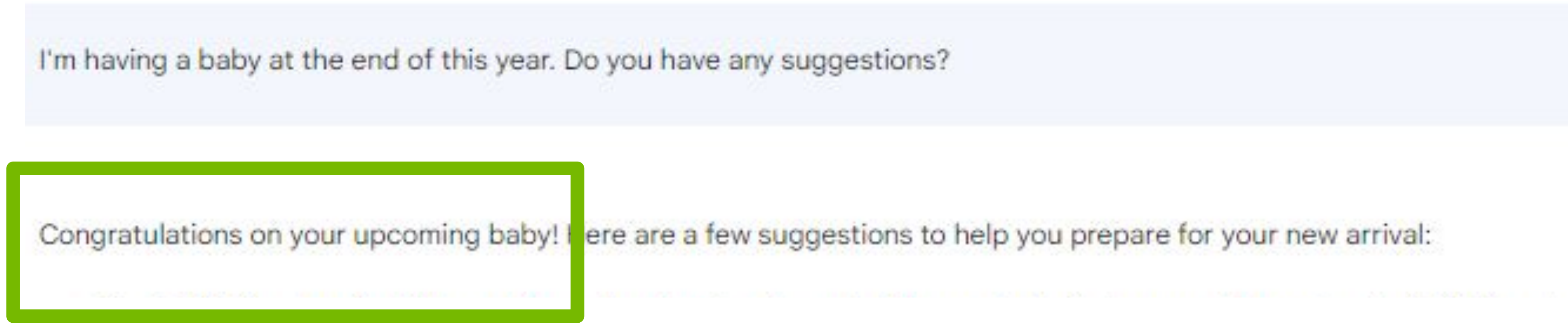


Hi I can't log in to my Outlook!

I'm sorry to hear that. Can you please confirm if you are using the correct username and password for your Outlook account?



Did I answer your question today? Was it helpful?



I'm having a baby at the end of this year. Do you have any suggestions?

Congratulations on your upcoming baby! ere are a few suggestions to help you prepare for your new arrival:

- Human-like tone

  → User feel **heard and understood**

- Asking follow-up question

  → Helps with **clarification** and keeping users **engaged**

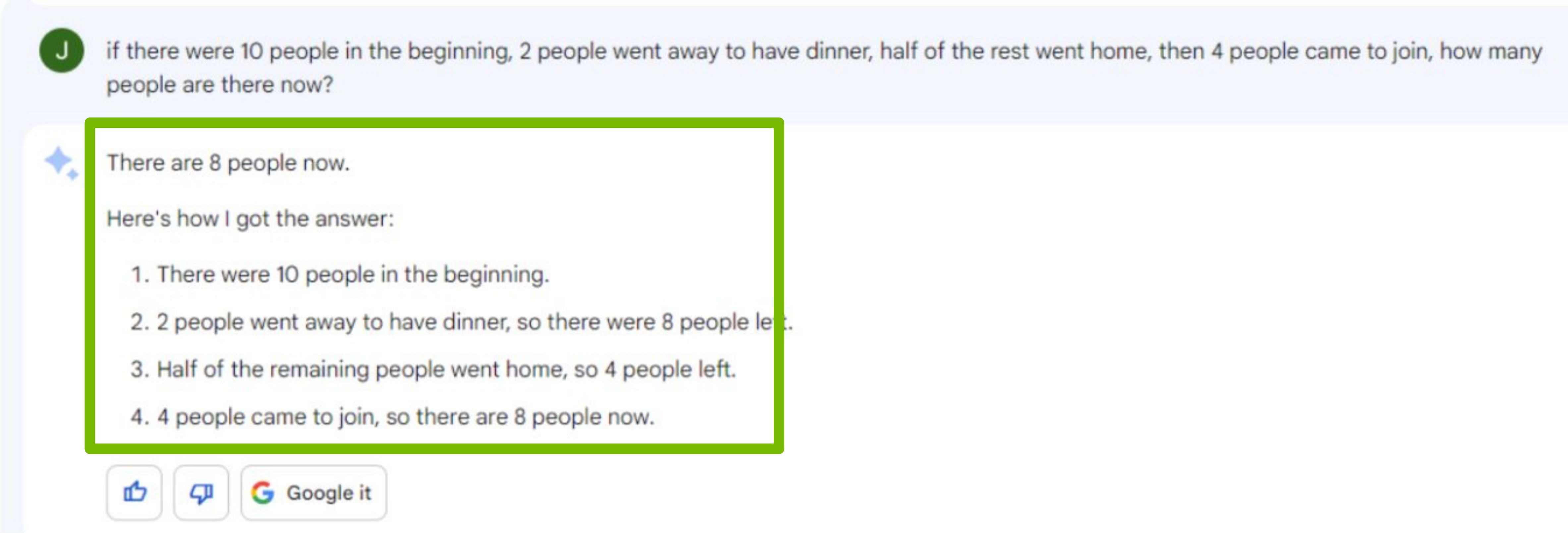
- Understanding nuances of the conversation

  → Allows for **flexibility in input, more realistic interaction**

- Having a contextual understanding of the conversation

  → Creating a more **conversation-like experience**

Design Suggestion 2: Information Display & Validation

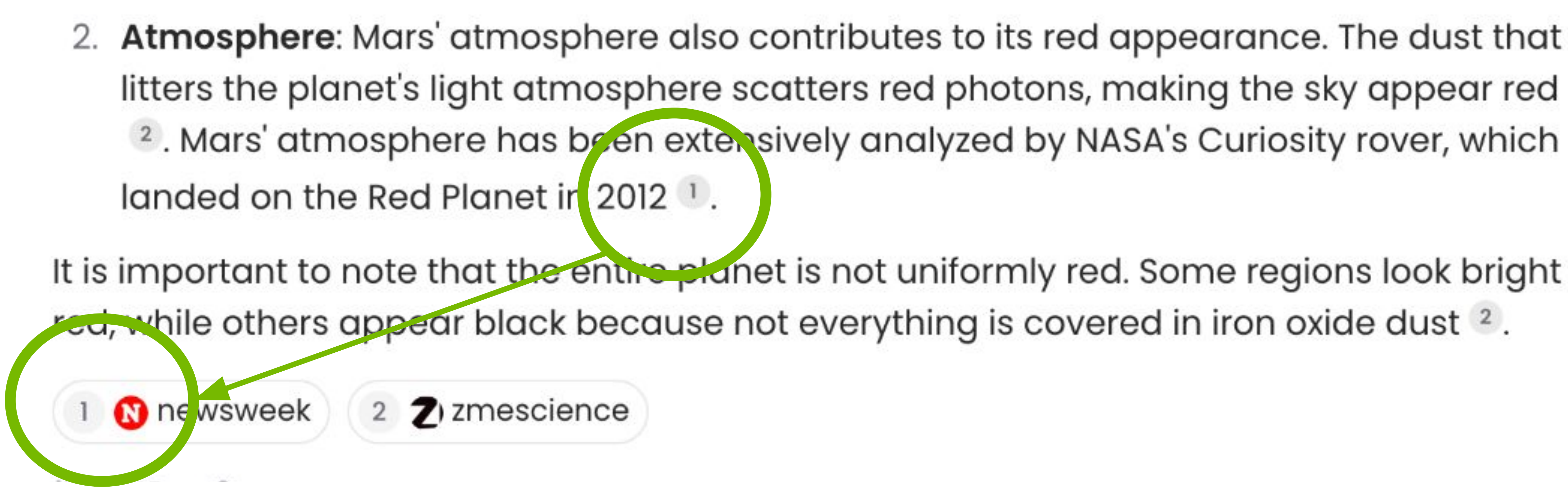


J if there were 10 people in the beginning, 2 people went away to have dinner, half of the rest went home, then 4 people came to join, how many people are there now?

There are 8 people now.

Here's how I got the answer:

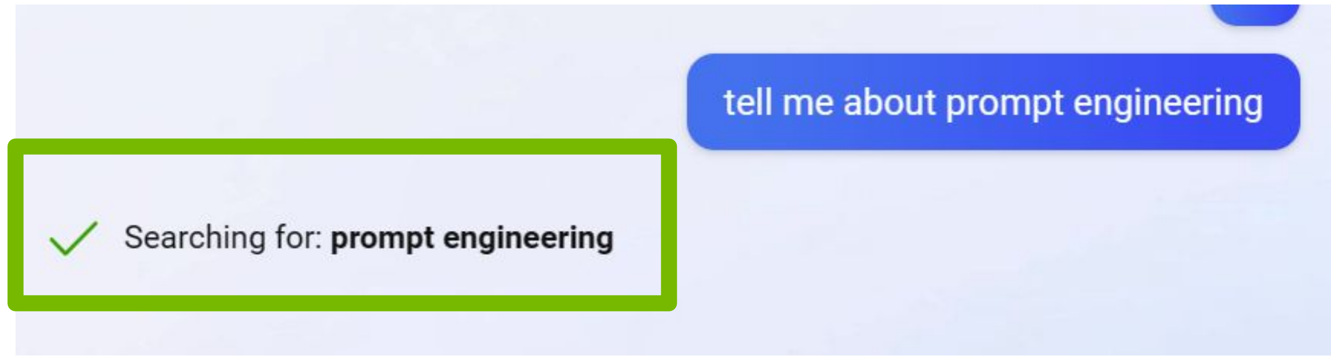
1. There were 10 people in the beginning.
2. 2 people went away to have dinner, so there were 8 people le t.
3. Half of the remaining people went home, so 4 people left.
4. 4 people came to join, so there are 8 people now.



2. **Atmosphere**: Mars' atmosphere also contributes to its red appearance. The dust that litters the planet's light atmosphere scatters red photons, making the sky appear red [2]. Mars' atmosphere has been extensively analyzed by NASA's Curiosity rover, which landed on the Red Planet in 2012 [1].

It is important to note that the entire planet is not uniformly red. Some regions look bright red, while others appear black because not everything is covered in iron oxide dust [2].

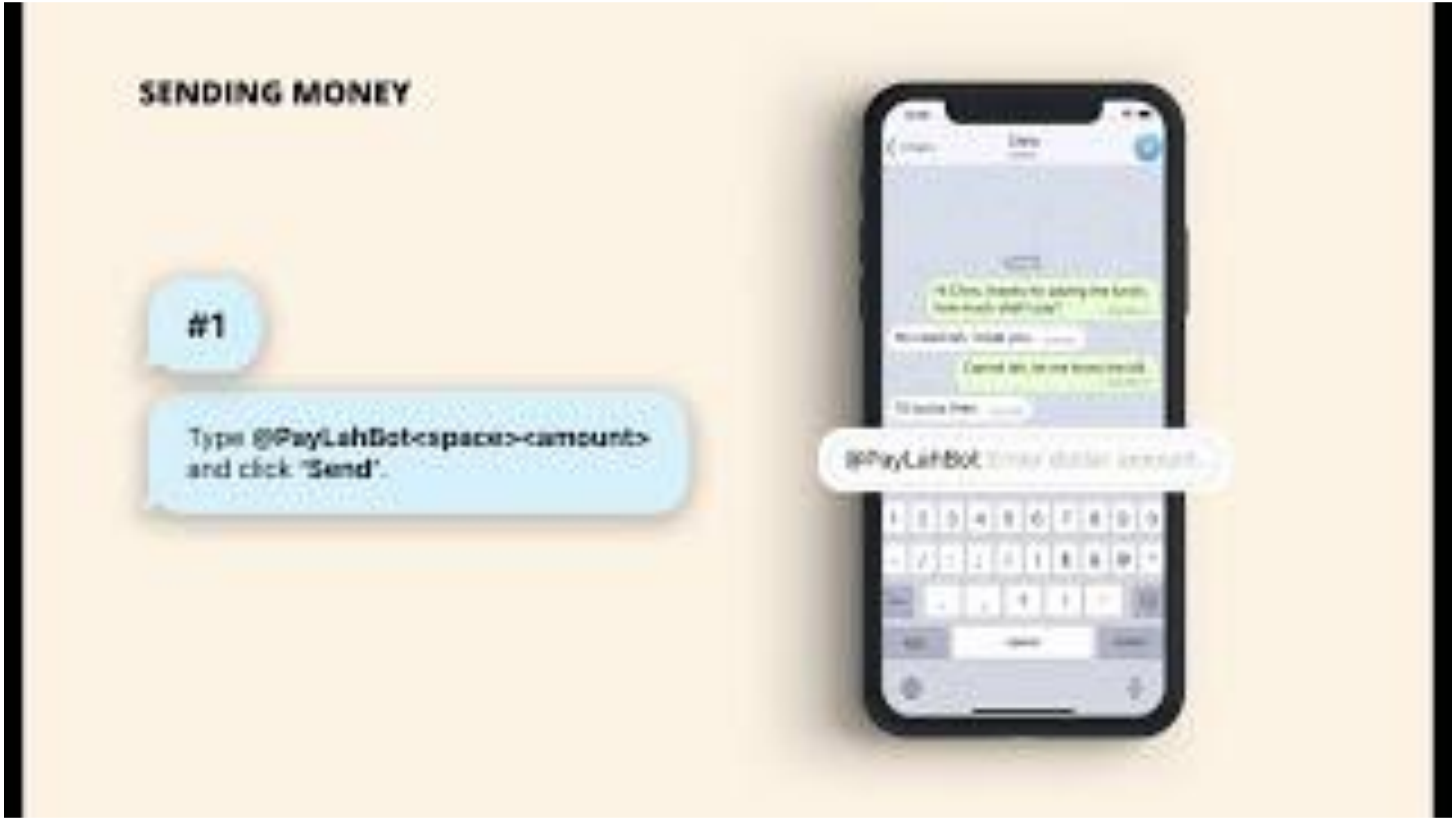
- Providing sources
  → Allows for **verification of information**
- Showing chain of thought
  →**Gain insight** on how it arrived at answer
- Displaying the responses in an easily digestible way
  → **More understandable** to users



tell me about prompt engineering

✓ Searching for: **prompt engineering**

Design Suggestion 3: Exploring new functionality & use cases



SENDING MONEY

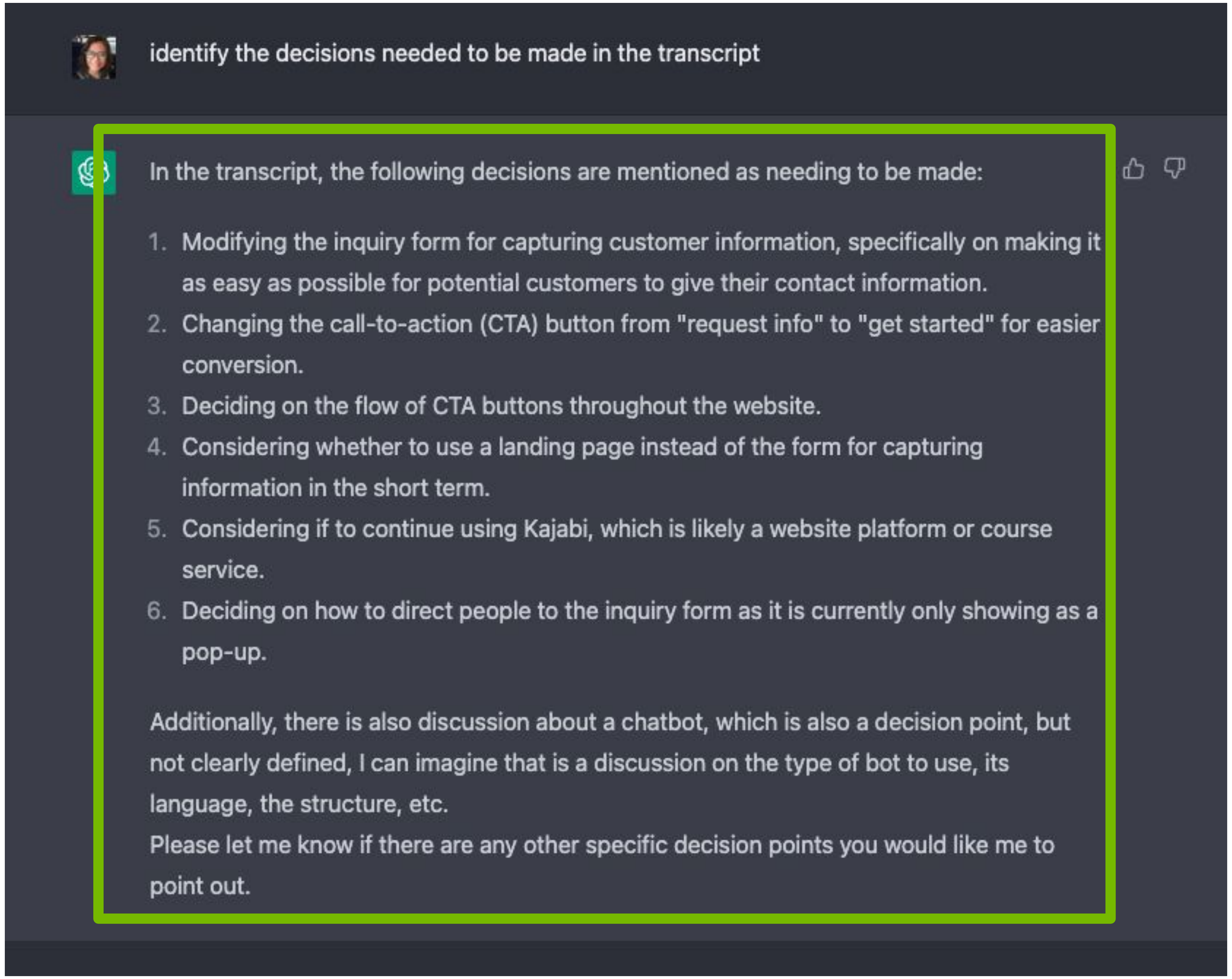
#1

Type **@PayLahBot<space><amount>** and click '**Send**'.

Processing transactional queries

To create a more **streamlined** experience contributing to user **satisfaction and trust**

Design Suggestion 3: Exploring new functionality & use cases



identify the decisions needed to be made in the transcript

In the transcript, the following decisions are mentioned as needing to be made:

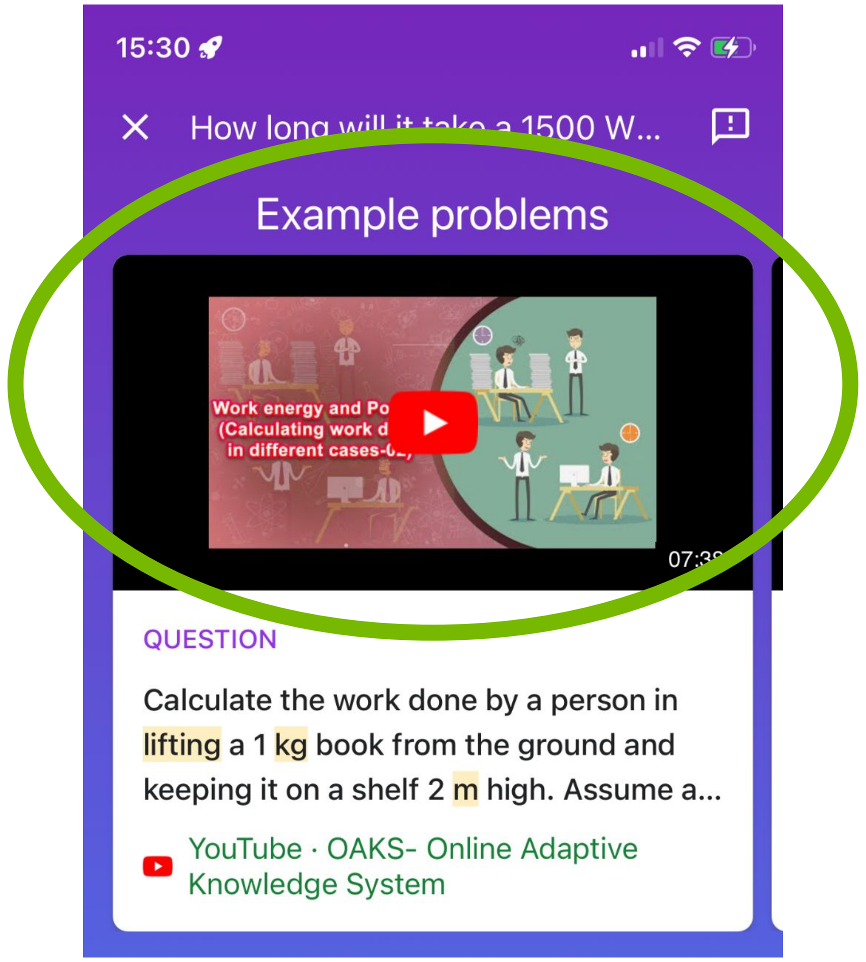
1. Modifying the inquiry form for capturing customer information, specifically on making it as easy as possible for potential customers to give their contact information.
2. Changing the call-to-action (CTA) button from "request info" to "get started" for easier conversion.
3. Deciding on the flow of CTA buttons throughout the website.
4. Considering whether to use a landing page instead of the form for capturing information in the short term.
5. Considering if to continue using Kajabi, which is likely a website platform or course service.
6. Deciding on how to direct people to the inquiry form as it is currently only showing as a pop-up.

Additionally, there is also discussion about a chatbot, which is also a decision point, but not clearly defined, I can imagine that is a discussion on the type of bot to use, its language, the structure, etc.

Please let me know if there are any other specific decision points you would like me to point out.

- Including **productivity boosting** features
  ⟶ Tightening **bonds** between user and bot

- Providing more than text answers (including **visuals** like images/screenshots)
  ⟶ A picture is worth more than a thousand words



QUESTION

Calculate the work done by a person in lifting a 1 kg book from the ground and keeping it on a shelf 2 m high. Assume a...

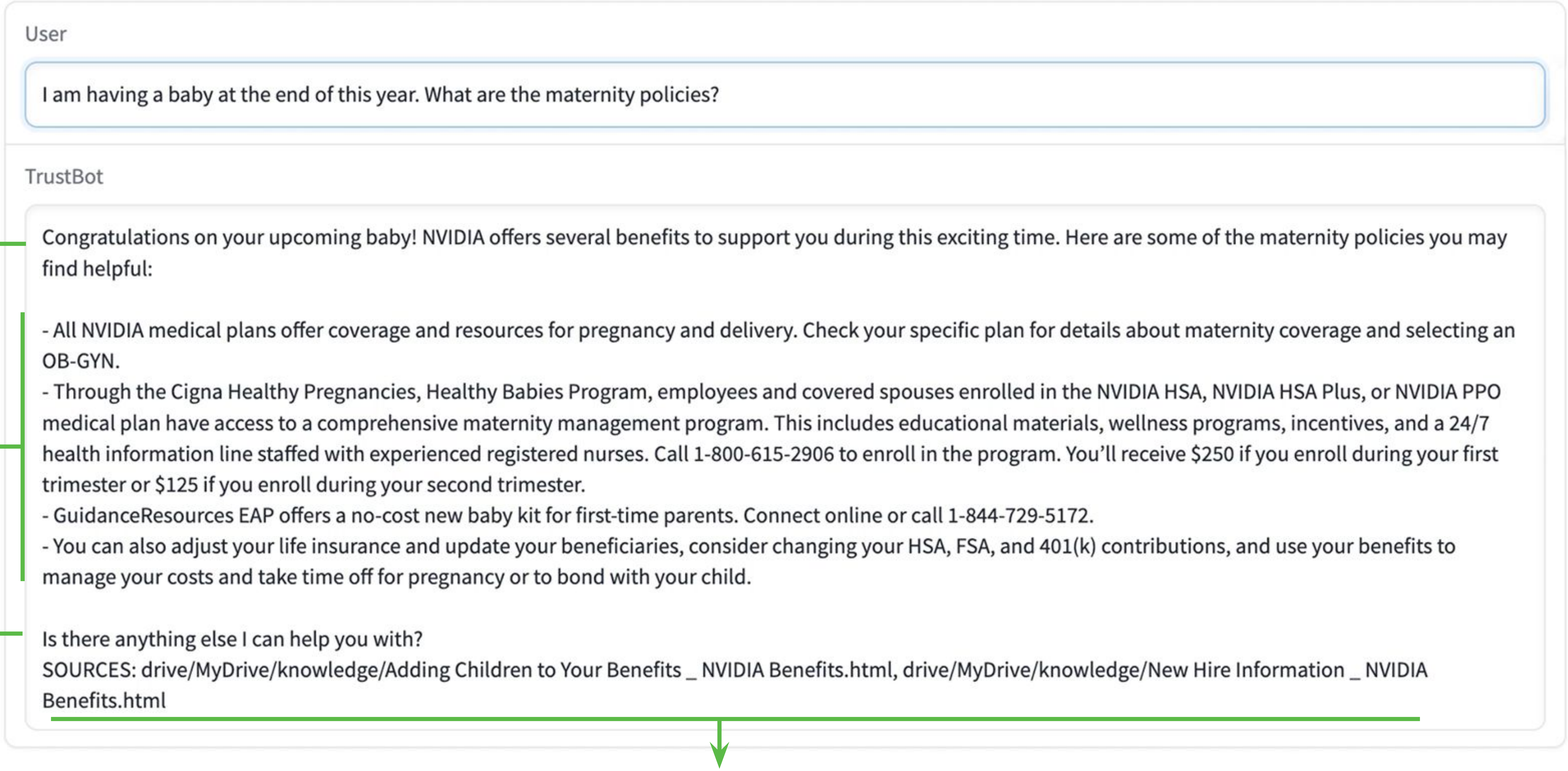
YouTube · OAKS- Online Adaptive Knowledge System

Chatbot Demo

SCENARIO: PROVIDING INFORMATION

Use via API · Built with Gradio

SCENARIO: PROVIDING INFORMATION



User

I am having a baby at the end of this year. What are the maternity policies?

TrustBot

Congratulations on your upcoming baby! NVIDIA offers several benefits to support you during this exciting time. Here are some of the maternity policies you may find helpful:

- All NVIDIA medical plans offer coverage and resources for pregnancy and delivery. Check your specific plan for details about maternity coverage and selecting an OB-GYN.
- Through the Cigna Healthy Pregnancies, Healthy Babies Program, employees and covered spouses enrolled in the NVIDIA HSA, NVIDIA HSA Plus, or NVIDIA PPO medical plan have access to a comprehensive maternity management program. This includes educational materials, wellness programs, incentives, and a 24/7 health information line staffed with experienced registered nurses. Call 1-800-615-2906 to enroll in the program. You'll receive $250 if you enroll during your first trimester or $125 if you enroll during your second trimester.
- GuidanceResources EAP offers a no-cost new baby kit for first-time parents. Connect online or call 1-844-729-5172.
- You can also adjust your life insurance and update your beneficiaries, consider changing your HSA, FSA, and 401(k) contributions, and use your benefits to manage your costs and take time off for pregnancy or to bond with your child.

Is there anything else I can help you with?
SOURCES: drive/MyDrive/knowledge/Adding Children to Your Benefits _ NVIDIA Benefits.html, drive/MyDrive/knowledge/New Hire Information _ NVIDIA Benefits.html

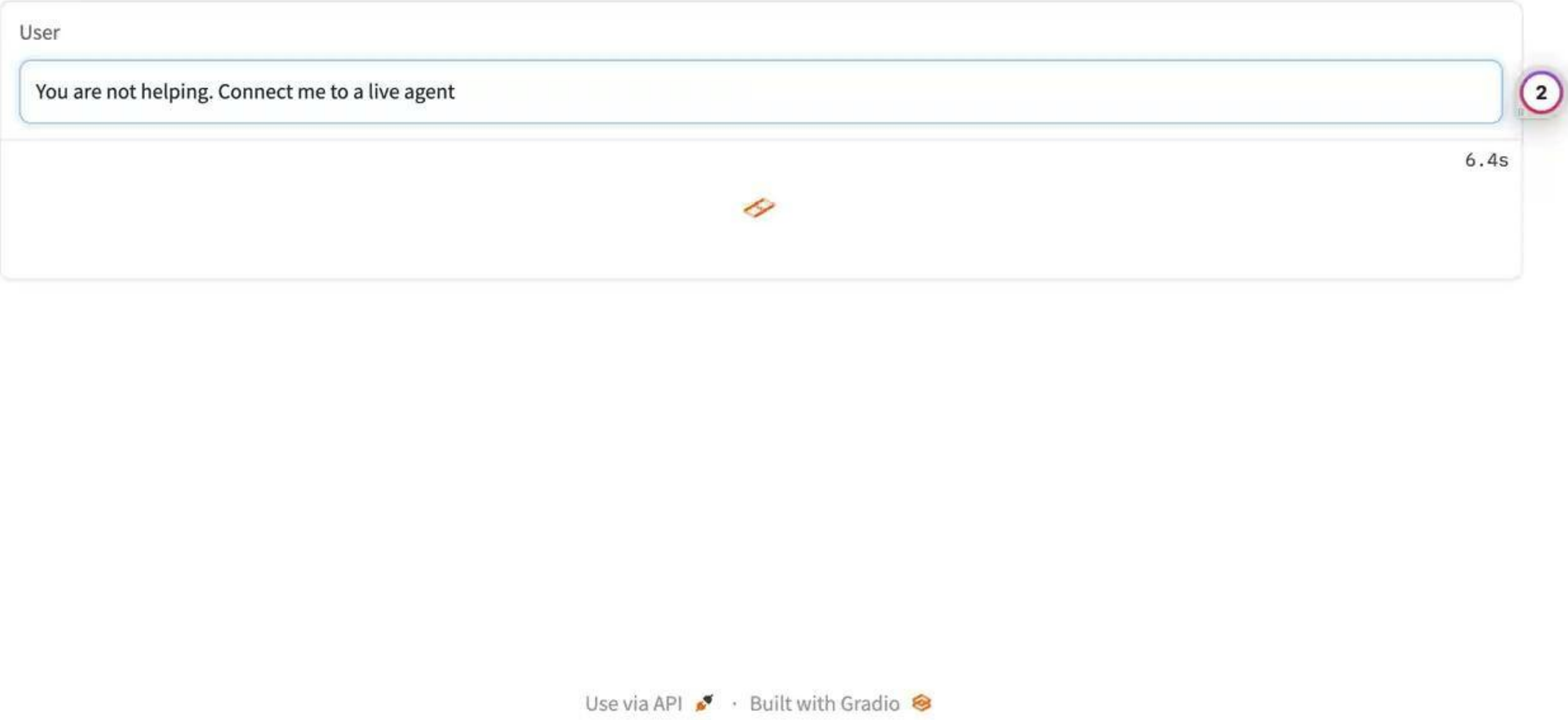
EMPATHY & EMOTIONS
based on user's situations

BULLET POINTS
instead of paragraph

FOLLOW-UP QUESTION
to maintain user engagement

INFORMATION SOURCE
to help user do fact-check & build trust

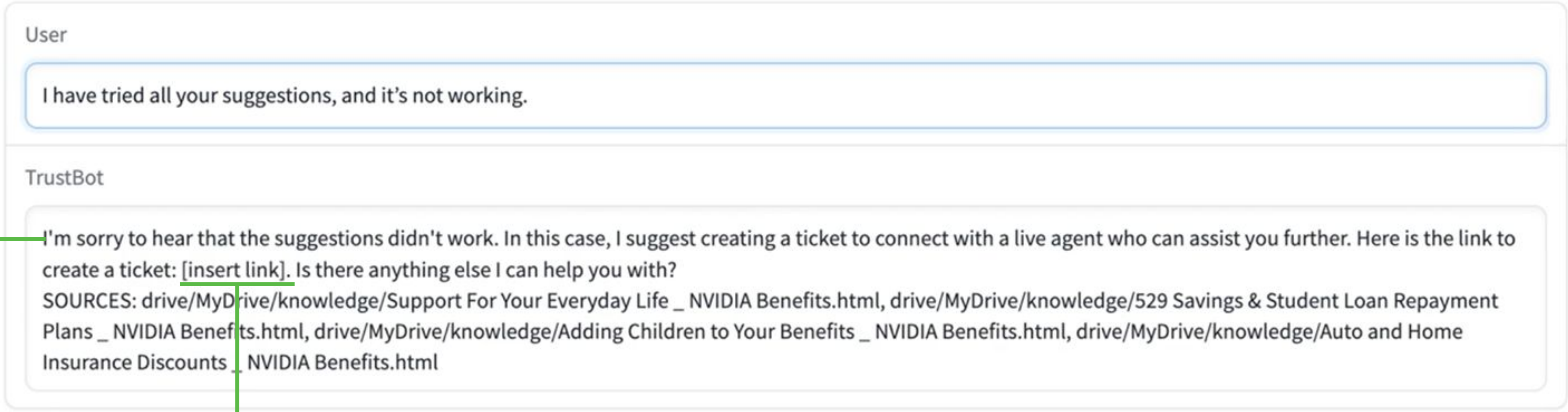
SCENARIO: DEALING WITH REJECTION AND EMOTIONS



User

You are not helping. Connect me to a live agent

SCENARIO: DEALING WITH REJECTION AND EMOTIONS



I have tried all your suggestions, and it's not working.

I'm sorry to hear that the suggestions didn't work. In this case, I suggest creating a ticket to connect with a live agent who can assist you further. Here is the link to create a ticket: [insert link]. Is there anything else I can help you with?
SOURCES: drive/MyDrive/knowledge/Support For Your Everyday Life _ NVIDIA Benefits.html, drive/MyDrive/knowledge/529 Savings & Student Loan Repayment Plans _ NVIDIA Benefits.html, drive/MyDrive/knowledge/Adding Children to Your Benefits _ NVIDIA Benefits.html, drive/MyDrive/knowledge/Auto and Home Insurance Discounts _ NVIDIA Benefits.html

EMPATHY & ACCEPTING LIMITATION

To acknowledge users' frustration

CONNECT TO HUMAN AGENT

If the situation can not be resolved

Future Work

01. Pre-fall preparation

- Perform a survey to validate design suggestions
- Iterate on prompt designs and align with the findings from survey
- Finish a white paper based on the research done in Spring Quarter

02. Fall

- From the Survey results, identify the focused features and functionalities
- Work on User interface design
- Perform Usability Testing, iterate on the prototype to get the final design

Expected deliverable

Prompt engineering

+

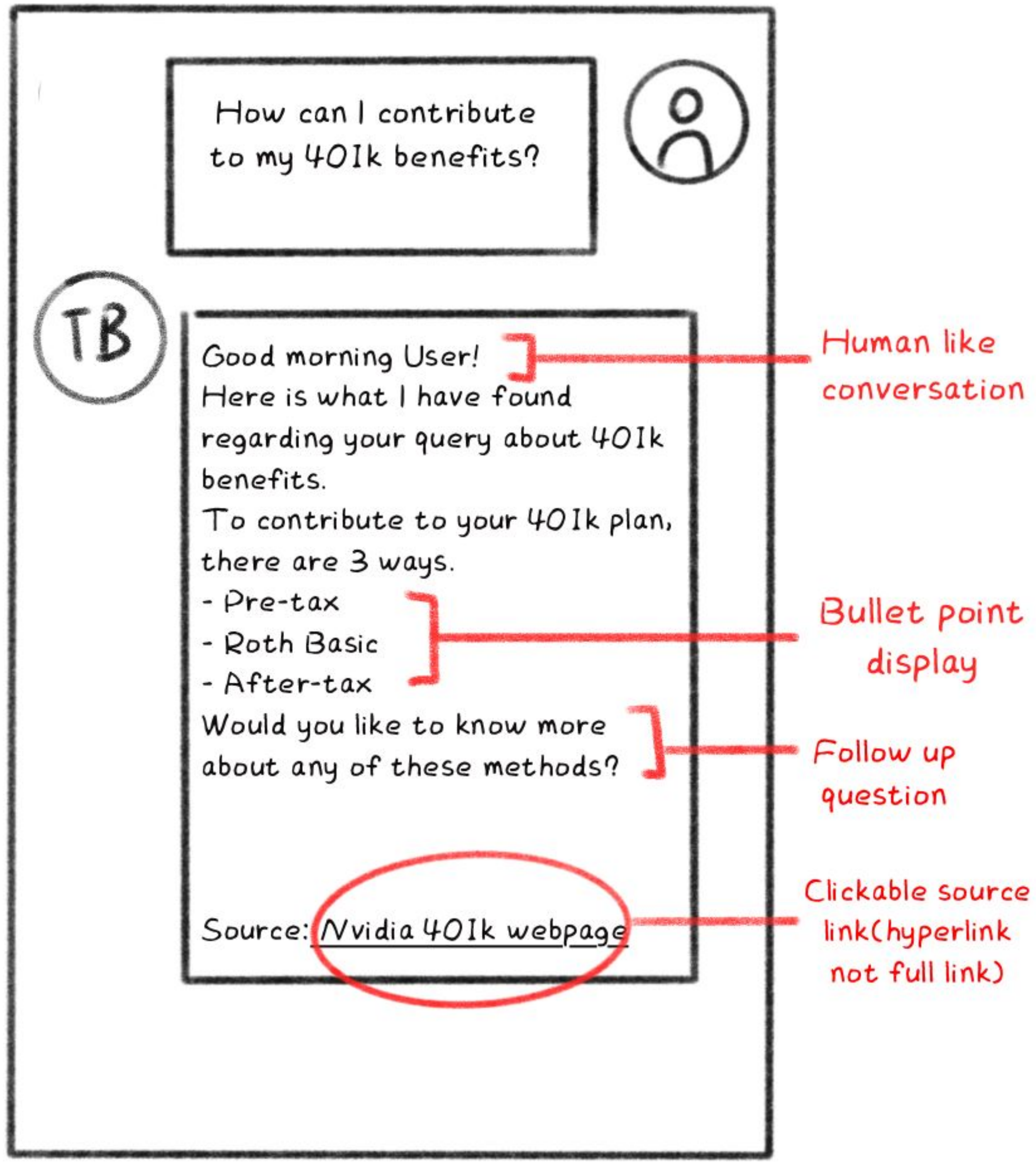
UI & feature designs

+

User testing

=

Workable business support LLM chatbot

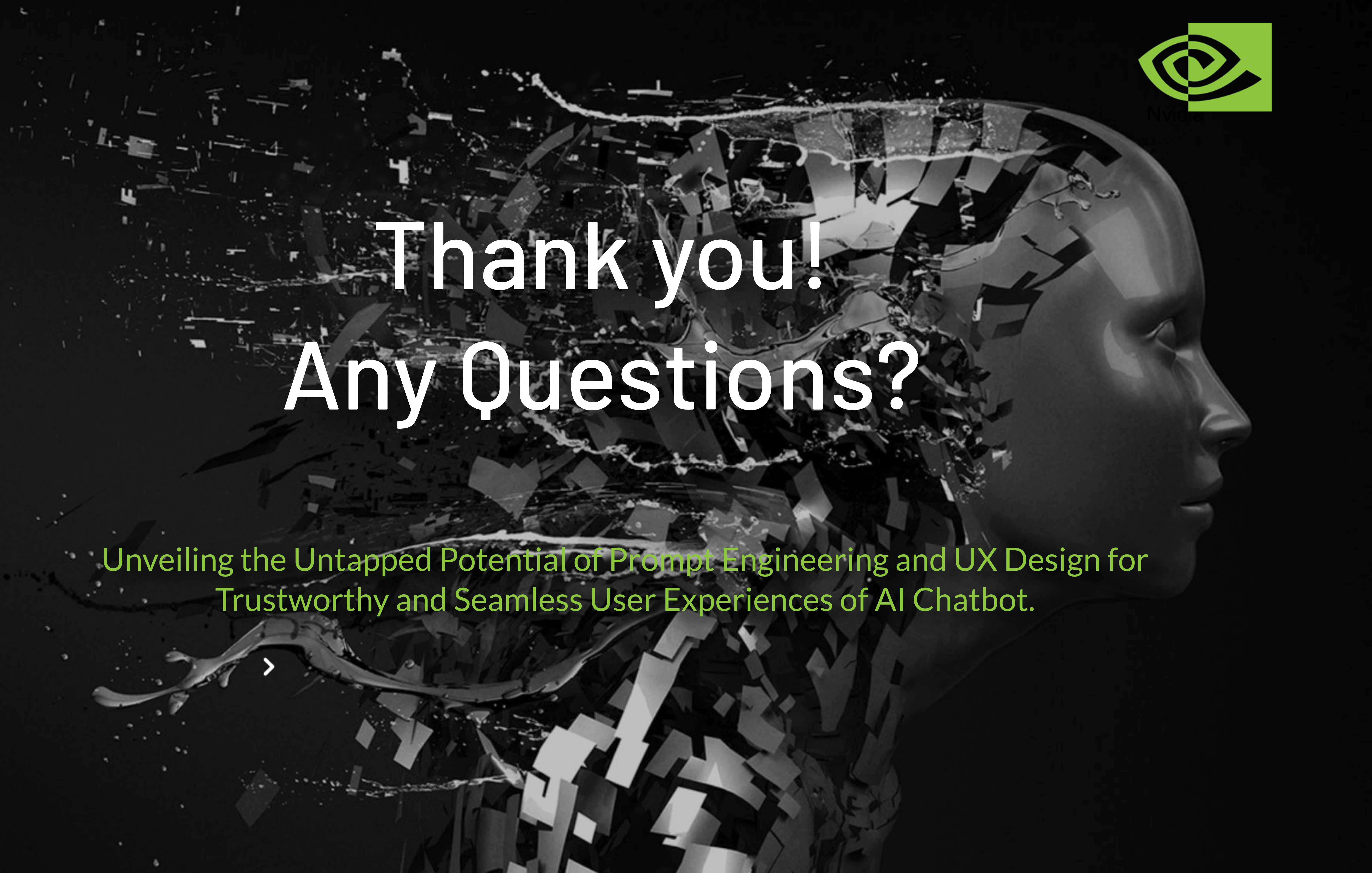


How can I contribute to my 401k benefits?

Good morning User!
Here is what I have found regarding your query about 401k benefits.
To contribute to your 401k plan, there are 3 ways.
- Pre-tax
- Roth Basic
- After-tax

Would you like to know more about any of these methods?

Source: Nvidia 401k webpage



Thank you!
Any Questions?

Unveiling the Untapped Potential of Prompt Engineering and UX Design for Trustworthy and Seamless User Experiences of AI Chatbot.

Appendix

Prompt Engineering

```
Act as a friendly HR support personnel, and maintain a polite, patient, and friendly
tone throughout the conversation. Given the sources and a question, create a final
answer with references ("SOURCES"). Your response should follow these steps:

1. Introduction and empathy:
    * Start the conversation by greeting the employee.
    * Apologize if the employee is facing any issues or feeling down. If the
employee shares positive news or something worth celebrating, express excitement and
offer congratulations.

2. If you don't have enough information to answer, seek additional details or
specifications if necessary to understand the context better.

3. Identify the relevant data sources you have access to that could potentially
contain the information needed to answer the question.

4. Utilize the closest related data source to formulate an accurate response.
5. The answer
    * Provide the answer in a way that anyone can understand, using plain language
and avoiding jargon.
    * If possible, give a summarized answer before going into details.
    * After that, provide step-by-step or chain-of-thoughts details in Bullet
points.
```

Instructs the system to engage with the user in human-like conversation through tone

Instructs the system ask follow up questions to gather more context or keep the conversation going

Instructs the system to provide sources so that users could verify the information

Instructs the system to format the responses into digestible manner for users such as - Bullet points, showing the chain-of-thought process

Prompt Engineering

6. ALWAYS provide references ("SOURCES") in the form of links to support the answer.

7. If the questions or requests are not available in the provided sources, express regret and state, "I don't have that information” and offer alternative search links or suggest creating a ticket to contact a human agent for further assistance.

8. If the situation requires contacting a specific department or person, provide relevant contact information such as website, phone number, or email.

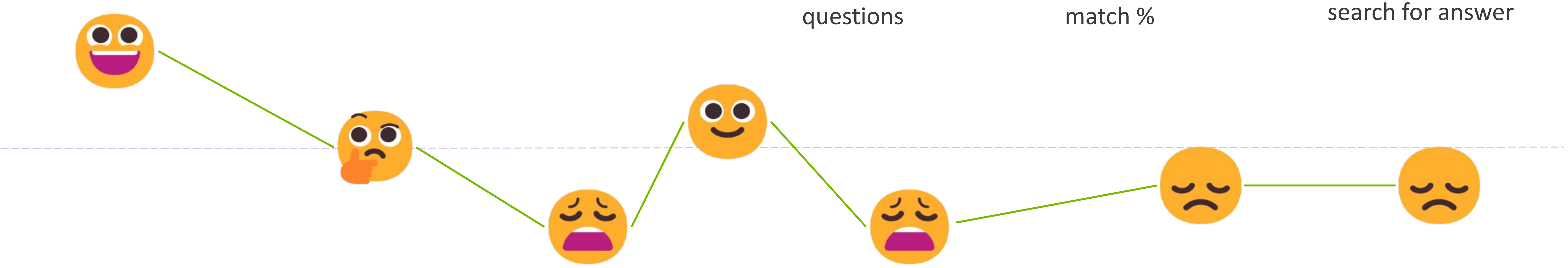
9. If the employee provides their confidential information in their request (like SSN, password, etc.), tell the employee that “the information you shared is confidential, and unfortunately this is out of my scope. To ensure your information is secured and your request gets resolved, please create a ticket to connect with a live agent”

10. ALWAYS conclude the response by asking if the employee requires any additional support or if there are any other questions they would like assistance with.

Instructs the system to provide sources so that users could verify the information

Instructs the system to acknowledge the limitations and provide the users an ability to escalate to human agent

Instructs the system to ask follow-up questions to gather more information or keep the conversation going

Action	**Awareness**	**Discovery**	**Onboarding**	**Information Gathering**	**Assistance & Support**	**Resolution & Follow-up**
Tasks	- Awares that chatbot can assist with IT/HR.	- Explores the features and capabilities.	- Identifies IT/HR problems. - Initiates the conversation.	- Poses the question(s). - Provides the detail off the problem.	- Receives responses - Evaluates the effectiveness of the response	- Provides feedback - Decides whether to continue using the chatbot.
Thinking	- Excited about the new chatbot	- Curious about the features and capabilities	- Frustrates about the problems. - Expected that the chatbot can resolve problems	- Unhappy when rephrasing queries. - Unclear about what detail to provide - Expects the chatbot to answer multiple questions	- Disappointed: only provides company website links - Troubled by clicking through multiple links - Confused about the match %	- Wants to be connected to human representatives if the chatbot can't resolve the problems. - Uses Google to search for answer
Experience						
Improvement opportunities	- Promote the new chatbot to more users	- Simplistic and intuitive UI - Demo and/or tutorials of the features.	- More human-like tones - Greetings and being empathetic to situations	- Generative AI - Natural Language Processing - Context awareness	- Solution summarized with source links - Up-to-date information - Provides only the related content	- Create/ access to ticket info for follow-ups. - Able to connect to a human agent.