PROJECT

Kalpana Baheti
kbaheti3@gatech.edu

INTRODUCTION

Inflection AI recently came up with a wonderful app called ***Pi*** (Personalized Intelligence). It is a simple application (with an undoubtedly mathematically complex backend), that allows users to choose a mode of focus in dialogue and then have fruitful and fulfilling open conversations with an AI bot whose hallmark is that it is considerate, a good listener, and very kind.

This is how the app look likes at present. This shall also be version 1 in evaluation stages. We will use this as a base for wireframing of version 2.

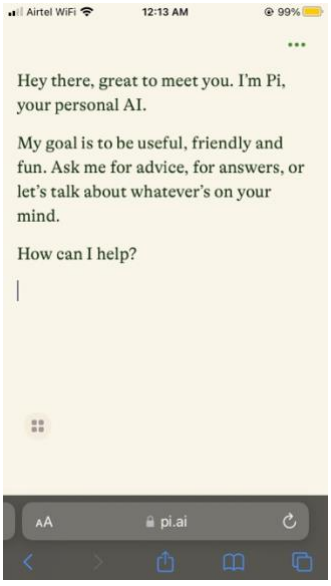

Figure 1 - Pi: Conversation Page

Figure 2 - Pi: Focus Selection

The way to access this interface is to go to https://inflection.ai/ on your mobile browser and click on *[Meet Pi]*. A conversation space with a greeting followed by a cursor opens up. On the bottom left of the page you see a four-dotted icon to set the focus of the conversation. You choose one, click on the *X* on the top right, and start typing your thoughts!

NEED-FINDING

Problem Definition –

We will address this interface's redesign in the following manner –

1. Part 1 – The Main Conversation Tab
    a. Goal Fulfilment
        i. Ease of Navigation
        ii. Ease of Conveyance of Thoughts
        iii. Ease of Understanding of Responses
    b. Comfort with Information Sharing
    c. Aesthetic Appeal

2. Part 2 – The Focus Selection Tab
    a. Goal Fulfilment
        i. Ease of Navigation
        ii. Ease of Understanding
        iii. Ease of Selection
        iv. Coverage of Conversation Focuses
        v. Granularity of Conversation Focuses
    b. Aesthetic Appeal

We will assume that the backend AI performs adequately in connecting the focus selection to the actual conversational content. Furthermore, we will not be experimenting with changing the modalities of information input and output on the Focus Selection tab. We will be assessing this possibility for the Main Chat page alone.

User Types (Includes Where They're From – Worldwide!) –

The user population for this app is quite massive. Everyone who goes through stress, wishes to have a safe space to vent and introspect, desires a casual pastime when bored, practice pitches, lectures and presentations, and plan for a productive day. But some general requirements would that the user –

1. Must have access to a phone and good internet.
2. Must have sound eyesight and good motor control with hands.

3. Must be literate in English and be aware of general AI limitations.
4. Must be above 12 years of age.
5. Must be comfortable with talking to a soft-bot.
6. Must have these app conversations sitting calmly (not doing any high-attention activities on the side like driving).

Need-Finding I: Interviews –

Process Description…

There will be 10 interviews taken with 10 people from the target population. They will be personas with varying levels of involvement in emotion when using the bot – light-hearted for-fun users, productivity users, and those who seek comfort and counsel.

After giving them 15 minutes to use the interface, I shall ask one question per sub-topic mentioned in the problem statement. The question will be phrased this way – *"Is there any feedback you'd like to share regarding the <aspect> of the app?"* There will 11 questions (one for each sub-topic) like this in total.

Data Inventory Addressal…

Context – Questions on constraints under which app is used.

Goals – Questions on conversation focus coverage and granularity.

Tasks – Questions on high-level navigation, conveyance and understanding.

Sub-Tasks – Questions on low-level navigation, conveyance and understanding.

The user's information is already covered, and the aesthetic appeal will take care of the UI constraints.

Possible Biases and Handling Them…

Observers bias is the major concern here since these are open-ended questions and I may interpret them in a way I'm usually prone to. To avoid this, I'll cross-exchange answers (anonymously) and gather thoughts on others' responses.

Need-Finding Results…

Summary from Interview 1 and 2 – Light-Hearted For-Fun Purposes

The goal was fulfilled since the tone was light and interesting and was bot initiated. It felt very fun and humane. The navigation was tricky since the texting bar kept covering the typing region so they couldn't see what they were typing. The focus selection bar was not very relevant since the conversation did not need any serious constraints.

Summary from Interview 3 and 4 – Personal Productivity Purposes

The selection for this focus was available and the experience was quite smooth. The scrolling was a problem when the user wished to see something the bot had mentioned above. There was also a suggestion that an automatic connection (after user confirmation and approval) was created with Calendar/Reminders.

Summary from Interview 5 and 6 – Career Advancement Purposes

These conversations were short but required focus. The users had several suggestions about the focus selection panels on the granularity of choice within career advancement. They wanted interviews, seminar presentations, brainstorming, pitch deck advice, stressbusting and so on. The open-text commands did cover some of this well though. Navigation and content understanding was praised here.

Summary from Interview 7 and 8 – Emotional Purposes

The users were wary of using this. One was scared that if they spoke of a few things by mistake, they would feel worse. But post-event, the user felt much better and actually felt lighter. Both users needed to type fast and explain a lot and conveyance was an issue. The open chat was not too comfortable for long statements, nor did it support new line entry. Furthermore, the white screen hurt their eyes.

Summary from Interview 9 and 10 – Other Purposes

A user here had a very interesting comment. If there is an option for 'Play a Game', why isn't there one for 'Sing a Song'. The users both felt there could be greater coverage of focus selections. Being their first time, they had some trouble with navigation and couldn't find the focus selection button. The volume choices didn't show up for a few seconds. And the screen was too bright.

Need-Finding II: Product Reviews –

Process Description…

These are the four sources we will be gathering user feedback from –

1. Established product review site – https://www.producthunt.com/products/pi-by-inflection-ai/reviews
2. An in-depth introspective web article by a user – https://apix-drive.com/en/blog/reviews/pi-by-inflection-personal-ai
3. Article on LinkedIn of user perspectives – https://www.linkedin.com/posts/eric-slatkin_chatbot-ai-conversationalai-activity-7059568205689737216-mQWm?utm_source=share&utm_medium=member_desktop
4. The biggest hub of comments on Pi directly to CEO – https://www.linkedin.com/posts/mustafa-suleyman_im-sooo-excited-to-announce-the-release-activity-7059261969551241218-o8wZ?utm_source=share&utm_medium=member_desktop

Each of the comments will be read and segregated into the problem statement categories based on where they belong best.

Data Inventory Addressal…

Based on the problem statement sub-topics (as described in the previous need-finding data inventory), the context, goals, needs, tasks, and sub-tasks shall be populated.

Possible Biases and Handling Them…

Social desirability bias might occur since these are public forums and people may feel inclined to praise a new AI venture. **Voluntary response bias** may also occur since connoisseurs might praise the product but those who might truly need it may not speak up or may not be aware. And finally, there is the usual **observers bias** when there are ambiguous or subjective sentences written. To counter this, review content will be prioritized based on stakes per single person and not by how many people talked about it. The observers bias will be resolved by consulting outside opinions on the statements.

Need-Finding Results…

Priority I –

1. Information Sharing Discomfort - Privacy concern on Pi asking for the phone number of the person. However, the compensation is that users can delete their account and chat history anytime they wish -though not through the interface, but through an email to the company.
2. Chat Conveyance Problem – After clicking Enter no further changes can be made to the reply. No newline functionality available.
3. Focus Selection Coverage and Granularity – It has a lot of users who use it for business preparation, brainstorming and level-heading. Perhaps some fine-tuned options in that area would be appreciated.

Priority II –

1. Constraint Selection under Conveyance Problem – The users wished to be able to speak to the bot themselves, but currently, the bot can speak, the user can only type.
2. Understanding Problem – Quoting a user on the voice mode – "*The one thing that bugs me is the latency between text and speech. It might help is the text is delayed enough for the speech to catch up."*
3. Wish for Added Goal – Chat cannot be downloaded via URL and stored.

Priority III –

Goal Fulfillment from the AI side - Quoting a journalist who used it; *"it had problems with creativity, a sense of humor, cynicism and cruelty, characteristic of living people."* Fine-tuning of the AI itself per focus selection was preferred here. There were some false negatives in suspending certain users who didn't do any wrong. This has again to do with the backend though.

Populating the Data Inventory and Stating Requirements –

Table 1 — Data Inventory Based on Need-Finding Execution

Inventory Part	Data Input and Corresponding Requirements
Context	The users weren't concerned with privacy, but rather with how gentle the bot would be with them or how realistic during emotional sharing. That was a high-stakes context. The second was that their eyes hurt from the light mode and the scrolling and reference to previous chat was flawed and tedious. We won't cover Privacy of phone number in this scope.

Inventory Part	Data Input and Corresponding Requirements
	Requirements – The AI controls will be covered under Goals. The aesthetics will shift to dark mode and there will a haptic friendly scrollbar for users.
Goals	The users categorized their focus selections (and thereby, goals) mainly into casual light chat, productivity and planning, career advancement, and emotional introspection partnering – there was some variation in how dedicated they wished the chatbot to be, and how realistic. **Requirements –** Provide these four major options and AI controls under Settings and consider an open-text instructional alternative for the last two users who preferred variety.
Tasks and Sub-Tasks	**Task 1** – Set AI Controls (**Sub-Tasks:** Click on Settings, Multi-select choices and set gradient options, read method to create custom query, Close Settings) **Task 2** – Set Chat Exchange Modality (**Sub-Tasks:** Click on volume, Read pop-up, Try and select voice, Select voice on one or both sides, Click anywhere to close) **Task 3** – Navigate Conversation (**Sub-Tasks:** Read prompt, Type response (new line permitted with Enter), Use scroll bar anytime, Click Send when response ready)

HEURISTIC EVALUATION

Discoverability…

Relevance of Principle in the Problem's Context:

The two main goals in using this interface are setting the focus, mode, and constraints of the conversation, and having the conversation itself. This is a mobile screen; an open conversation should have an open feel so the screen cannot be too cluttered. Hence, minimalist design is preferred. However, that could impact discoverability due to lack of explicit indicators.

Gulfs of Execution:

First time users are often there just to try to out the app. Repeat users may like to fine-tune their experience. The intentions bear no significant gulf. Actions is where the gulf may exist. The little four-dot button opens up fine-tuning of AI, but the four-dots is small and not explicit in meaning. This is where a gulf may exist. Regarding the cursor blinking slow, that is quite discoverable, and hence there is no gulf there. The stage of executing on interface has no gulf either once the icons are discovered.

Gulfs of Evaluation:

There is no significant gulf present in the current interface at the interface output, interpretation or evaluation stages since the whole process is quite evident and hence discoverable. The selections get highlighted, the volume button turns on upon clicking it, and the AI conversational response is clearly displayed.

Affordance…

Relevance of Principle in the Problem's Context: We covered whether the operators for the tasks were discoverable in the last point. Now we shall cover whether their form of existence is intuitive to use to accomplish their respective goal and visually and reactively representative of what the user thinks it's meant for.

Gulfs of Execution:

The gulfs here exist mainly around the fine-tuning side of the task, since the actual conversation is prompted by a cursor which is a known definition of starting a message or elaborating on thoughts. But when it comes to fine-tuning, while people would still figure out to press the four-dot icon (since it's the only there, not because it actually depicts settings), when the different focuses are opened up, the user does not know if they can select one or more than one as a mixture. This is not afforded by the design. Neither do they know if their actions are easily reversible. Hence, this is a gulf at action and execution stages.

Gulfs of Evaluation:

Once a single focus on the fine-tuning is selected, the tab closes and the conversations begin. The cursor is starts blinking indicating the chat is ready and you can click the cursor to start typing like you usually do. This is well afforded and does not bear significant gulf.

Consistency…

Relevance of Principle in the Problem's Context:

There have been other AI chats before this, and they have been widely used by the same user population. The way those chats have been designed is what should be abided by, unless there is an addition, reduction or change that substantially improves the experience.

Gulfs of Execution:

The first that comes to mind is the volume button. That is not present in several other AI chats. This is not consistent and is a gulf at the intention (people might not know they want this) stage of the gulf of execution. But this is a good thing – it significantly improves the experience. The other one is the setting button similarly missing in other AI chat apps and present on this one.

Gulfs of Evaluation:

Outside of the settings and spoken version that are not present in other apps, the manner in which the user-written piece is compartmentalized and moved a little upwards after pressing [Enter] and the response of conversational AI being conveyed is akin with what other AI chat apps follow, hence there is no gulf here.

Flexibility…

Relevance of Principle in the Problem's Context:

Within our very large user base, we have people who prefer to fine-tune their chat, those who don't, people who prefer spoken conversation over written, and within those, ones who prefer female over male voices and so on. This warrants flexibility.

Gulfs of Execution:

When it comes to selection of fine-tuning facility, spoken version, and type of response voice, there doesn't exist a gulf (though a cost comes with consistency) since these are present in intention, action awareness, and interface execution. However, there is a gulf within fine-tuning where user cannot select multiple options at once nor can they select custom options. Looking at what is present, a user may not even know if they want another option. Hence, this is a gulf present at intention, action, and execution (since no soft button exists to click on) stages.

Gulfs of Evaluation:

This part seems to be mostly covered quite well in the sense that flexibility in conforming selections and sent responses does not matter here since they stick by a single standard indicators that the execution was successful. However, when it comes to volume, the only way to finally know if you have the right voice and volume set is by starting the conversation. If an automatic 'Hello, I'm Pi' in

the chosen volume and. Voice is played, that might cover a gulf at the interface output and final understanding part of the evaluation process.

Ease and Comfort…

Relevance of Principle in the Problem's Context:

This app is supposed to handle matters close to heart and wellbeing of a person and help them with their life. It is essential that the experience is smooth, pleasant, and devoid of stress and triggers.

Gulfs of Execution:

The gulf of execution here is related to the ease of performing the action at the interface execution stage. When in a dim-lit room, this chat app does not permit adjusting of screen light or a dark-mode version. This makes the operation on the interface unpleasurable and fatigue causing. In general, haptics and audio are not contributing to this gulf so ease and comfort on that end is fulfilled.

Gulfs of Evaluation:

Sometimes the user likes to scroll upwards on long responses. The mobile web-app session basically scrolls all the way up to the start. This is a major gulf in the evaluation of goal of scrolling up. The intention was to scroll a little upwards. The result is understood by the user since it takes them to the start of the conversation, but that's not what the user wished for, so it fails here.

Improvements in Conclusion…

1. Consider a different design for the settings button.
2. Consider a sample voice auto-test for the volume settings.
3. Redesign coverage, mode, and granularity of fine-tuned settings for focus choices based on human-in-loop need-finding conducted previously.
4. Redesign visuals and scrolling haptics.

INTERFACE REDESIGN

Wireframe Prototype (New Version)...

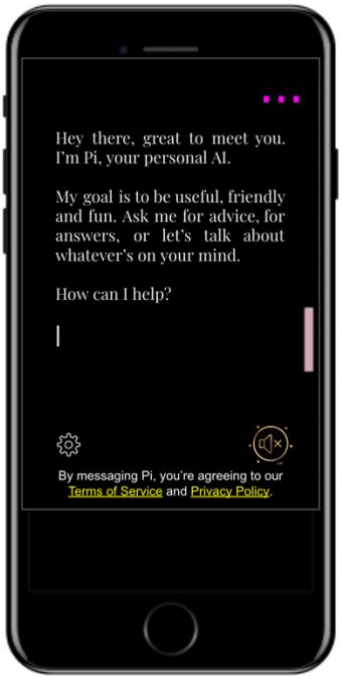

Figure 3 – Landing Chat Page

Figure 4 – Upon Clicking Settings

Figure 5 – Upon Clicking Volume

Figure 6 – Upon Clicking Cursor

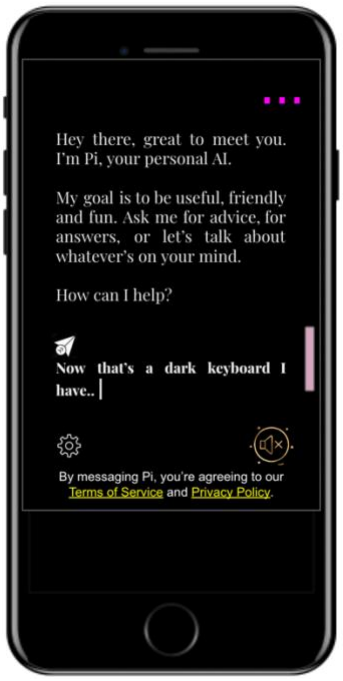

Figure 7 – Ready to Send/Scroll

INTERFACE JUSTIFICATION

Positive Existences Retained…

1. We retain the positions of the focus selection and volume icons in the same places.
2. We retain the overall feeling of space on the chat and the keep the chat structure as it was.
3. We provide the same voice examples for the spoken version of the chat.
4. We cover the main focus options that already exist for selection.
5. The mode for selection for options is to click on it and they turn darker to indicate selection was registered.
6. We keep the prompt to answer the same – a slow blinking cursor.
7. The Terms and Services and Privacy Policy links for information are placed in the same position.
8. Finger-scroll is supported.
9. The three dots on the top right for external related operations are in the same place and follow the same path of use – hence we have not covered that in the wireframing.
10. Going back to chat screen after focus selection and volume adjustment is retained as the same action – clicking the X on the top right, and clicking anywhere outside the option box, respectively.

Response to Criticism as per Original Problem Statement Format…

1. ***Part 1 – The Main Conversation Tab***

    a. ***Goal Fulfilment***

        i. **Ease of Navigation**

        Gulfs of Execution: The hampered uncontrolled scrolling effect was a gulf at performing action on interface.

        Redesign: The scrolling upwards is slow and does not go back to the start. The scrolling may now be controlled by a side bar as well. The side bar has a wide enough width for a finger to easily be placed on it.

ii. **Ease of Conveyance of Thoughts**

Gulfs of Execution: There was no facility to involve new lines to create a feeling of pause and effect in user's response. The draft could not be easily edited and could be sent very easily by error before completion. The voice exchange mode did not have a version where users could speak to the bot. These are all gulfs in the action stage.
Redesign: The draft is now designed like a normal messaging chat interface, and new lines are created with clicking on [Return]. After the draft is complete, the typing space may be closed via [X] or [Done]. After that a tiny paper-plane icon emerges above the typed draft indicating [Send]. This sends the message without ambiguity. The briefness and familiarity of the icon saves space while maintaining functionality. There might be a gulf of evaluation caused for some existing users since the send button is new and an extra step and they might think the text is sent. But there was a lot of criticism on the previous version. The voice mode selection now has two small stages of selection. The user must select which sides of the conversation will deliver through voice. If the AI side is turned on, the second pink bubble opens with voice choices.

iii. **Ease of Understanding of Responses –** There was an issue of the speed of the voice not commensurate with text, which can be easily adjusted but we haven't covered that.

b. ***Comfort with Information Sharing*** – The emotional sharing mode faced this concern. It has been addressed under selection tab.

c. ***Aesthetic Appeal***

Gulfs of Execution: The bright screen mode was harsh on the eyes and caused fatigue and consequently difficulty in executing any action on interface.

Redesign: This has been changed to dark screen mode with pastel hues that are popularly liked and soothing on the eyes.

2. *Part 2 – The Focus Selection Tab*

   a. *Goal Fulfilment*

      i. **Ease of Navigation**

      Gulfs of Execution: The four-dotted icon didn't register well and was overlooked. This was a gulf in action understanding.

      Redesign: The icon was changed to the well-known Settings icon while still saving space and maintaining better functionality.

      ii. **Ease of Understanding**

      Gulfs of Execution: There was a gulf through missing actions. There was no fine-tuning possible, whereas now there is.

      Redesign: The % is a common symbol understood as a fraction or extent.

      iii. **Ease of Selection**

      Gulfs of Execution: Multi-select on options was not functionality permitted. Neither was fine-tuning specifications. This was a gulf at interface execution and missing action levels respectively.

      Gulfs of Evaluation: The interpretation upon selection of one focus is flawed. It does not specify that no other selection may now happen. Ideally, the interface response should've been whitening out the others.

      Redesign: We allow two options and upon selection of second options the rest are whitened out. The volume controls have a new panel with familiar on-off sliders.

      iv. **Coverage of Conversation Focuses –** This was a gulf of execution in missing action. Redesigned by categorizing

main topics better and clearly instructing how other focuses may be created.

v. **Granularity of Conversation Focuses –** This was again a gulf in missing action and was fixed through adding the percentages in deviation and softness of arguments.

b. ***Aesthetic Appeal*** – Pastel color theme followed.

EVALUATION PLAN

Data Recorded – Overall Numerical Rating

Section 1 Scoring – Conversation Pathway (Includes Navigation, Conveyance, Understanding, and Aesthetics)

Section 2 Scoring – Focus Selection Pathway (Includes Navigation, Understanding, Selection, Coverage and Granularity, and Aesthetics)

Section 3 Scoring – Chat Exchange Mode Pathway (Includes Navigation, Understanding, Selection, and Aesthetics)

Testing Process –

- Post-Event
- Synchronous
- Live Demonstration (both versions of the app are comparatively new)
- Individual Feedback

Experiment Design and Definition of Null/Alternate Hypothesis –

A **randomly selected half between-group** manner (collected from halves of each sub-group of our key demographic – stratified random selection) will be assigned to version 1 and the other half to version 2. We will keep it simple, **no ordered selection** here.

Control Group: Assigned to original version of the Pi app.

Experimental/Treatment Group: Updated version of Pi app (prototype given).

Independent Variables: Interface categories - version 1 and version 2

Dependent Variable: Numerical score for each version per section.

Statistical Test: *Chi-Squared Test* per section.

H_{NULL} for Section *i* → If both distributions for Section *i* are **equal**

$H_{ALTERNATIVE}$ for Section *i* → If Distributions for Section *i* are **unequal**

Raw Scores out of 5 – Found in the Appendix: Conducted for…

30 People, 2 Groups (of 15), 3 Sections, Row Order is Original 1 then Updated 2

Results for Chi-Square Test –

Section 1 (Conversation Pathway) – Updated Version Better!

```
Chi-square statistic: 4.899596740229112
P-value: 0.987161474415576
Reject the null hypothesis. Unequal Distributions!
```

Section 2 (Focus Selection Pathway) – Updated Version Better!

```
Chi-square statistic: 5.235687633262262
P-value: 0.9822546147307654
Reject the null hypothesis. Unequal Distributions!
```

Section 3 (Chat Exchange Mode Pathway) – No Significant Difference.

```
Chi-square statistic: 0.73015873015873
P-value: 0.9999998752887159
Fail to reject the null hypothesis. Equal Distributions!
```

Final Comments – The focus selection and chat space updates were necessary. The speech controls, while effective among the population that prefers voice mode, was not emphasized in the scores. That could be deprioritized. Further narrowing down may be done through qualitative evaluations on individual points and higher grain quantitative evaluations.

REFERENCES

1. The Ed Discussion lectures, extra readings, and Inflection AI.
2. My friends and acquaintances from my old school who contributed to the need-finding interviews and the empirical evaluation.

3. The product review web source, the dedicated web article, the LinkedIn article, and the open-reviews LinkedIn post on experiences with Pi.

APPENDICES

Empirical Evaluation - Raw Scores out of 5 Conducted for…

30 People, 2 Groups (of 15), 3 Sections, Row Order is Original 1 then Updated 2

Section 1 (Conversation Pathway) –

5	2	2	5	5	3	2	3	5	3	5	2	4	4	3
3	5	5	4	4	5	4	4	4	4	4	3	5	5	5

Section 2 (Focus Selection Pathway) –

3	3	3	3	3	3	3	1	5	5	5	3	3	3	4
4	4	4	5	5	5	5	5	3	3	5	5	5	5	4

Section 3 (Chat Exchange Mode Pathway) –

4	4	5	4	5	4	3	5	5	5	5	3	3	3	4
5	3	4	4	5	5	3	4	5	3	5	5	4	4	3