

Hongfei Ji

Just an error?

Prototyping embodied
experiences of LLM
hallucinations

SPRING 2024

*Carnegie Mellon University
Master of Science in Computational Design*

Committee Members: Daragh Byrne, Sinan Goral

Contents

<i>Abstract</i>	13
<i>Introduction</i>	15
<i>Background</i>	19
<i>Discover the research gap: literature review on LLM Hallucination, algorithmic experience, and experience reflection</i>	19
<i>LLM (Large Language Model) hallucinations: intrinsic flaw as a design material</i>	19
<i>Bridging the Gap Through Algorithm Experiences for Embodied Speculation</i>	23
<i>Narrative Medium for Experience Reflection: Multimodality and Speculative Design</i>	26
<i>Precedents: Embodiment and medium for conveying discourse</i>	30
<i>Metrics to evaluate research gaps</i>	30
<i>Precedents investigation</i>	31
<i>Hands-on Experiment of hallucination characteristics and embodied integration</i>	41
<i>Experiment 1: Exploration of LLM Hallucination Characteristics</i>	42
<i>Experiment 2: Exploration of LLM Hallucination Characteristics</i>	45
<i>Summary - Framing the research</i>	48

<i>Research Questions and Hypothesis</i>	53
<i>Tracing from the origin: Algorithm experience in human-algorithm interaction</i>	53
<i>LLM hallucination characteristics: Hallucination in real-world interaction</i>	55
<i>Anchor of hallucination imagination: Hallucination Experience Glossary</i>	55
<i>Methodology</i>	59
<i>How I work</i>	59
<i>Probes to inspire imagination: Hallucination experience prototyping</i>	60
<i>Prototype 01: Moodie Assistant</i>	60
<i>Prototype 02: Whisper Web</i>	68
<i>Prototype 03: Mindscape</i>	75
<i>Pilot Test Design</i>	83
<i>Summary - A comprehensive mind flow</i>	84
<i>Evaluation</i>	87
<i>Response from Participants</i>	87
<i>Recruitment and Participants</i>	87
<i>Findings</i>	87
<i>LLM Hallucination characteristics add-on</i>	88
<i>Interactive pattern of LLM hallucination experience</i>	89
<i>Embodied medium's effect on hallucination experience</i>	91
<i>Comparative Reflection: What is the gap here?</i>	92
<i>Design Probes</i>	92
<i>Pilot Study</i>	94
<i>Design Insights</i>	95
<i>Conclusion</i>	101
<i>Limitations and Future work</i>	103

<i>Appendix A: Pilot Study Protocol</i>	111
<i>Prototyping session (2 persons)</i>	111
<i>Workshop (6-8 persons)</i>	112
<i>Appendix B: Participatory Design Outcome</i>	113
<i>Acknowledgement</i>	115

List of Figures

1	How does LLM work?	20
2	The proposed LLM - human need interpretation. Credit King et al. (2023).	21
3	The classification of LLM Hallucinations. Credit Y. Zhang et al. (2023).	22
4	A series of interesting interaction between human and LLM chatbots. Credits @ChatGPTapp (2023), @immasiddtweets (2023), Pringle (2023), and @voooooogel (2023).	23
5	Prototype's three metrics from user-centered design perspective. Credit Houde and Hill (1997).	24
6	Tangible Interface of algorithm experience and Visualization of human-related algorithm feature: classification. Credit Alvarado et al. (2022).	25
7	A multimodal LLM framework for developer. Credit K. Li et al. (2023).	27
8	Recent GPT-4: support multimodal input.	27
9	The Inaccessible Digital Camera. Credit Pierce and Paulos (2014).	29
10	The table-non-table. Credit Wakkary, Desjardins, and Hauser (2016).	29
11	Different angles to perceive hallucination experience.	30
12	LAUREN's view of domestic space and user activities. Credit (lauren mccarthy, 2018)	32
13	Rudiments 1, 2 & 3. Credit (Helmes et al., 2010)	34
14	Three speculative artifacts in domestic context. Credit (STUDIO, 2017)	35
15	Speculative artifact setting and quantitative analysis result. Credit (Retzepi, 2019)	37
16	The algorithm controller "Recffy" and its user instructions. Credit (Alvarado et al., 2022)	38
17	The reflection machine for voice, text and image medium. Credit (Mols et al., 2020)	39
18	The chat bot interface on mobile phone. Credit (Newton, 1900)	40
19	Brainstorming process of digital camera variations. Credit (Pierce & Paulos, 2014)	41
20	Fact-conflict hallucination: mistaken architecture school professor as football team manager	42

21	Fact-conflict hallucination: a deep concealed made-up information	
	43	
22	Input-conflict hallucination: API doesn't respect the predefined system role	44
23	Context-conflict hallucination: API forgets the formal conversation	45
24	After being fed hallucination definition, API gives some fun and human-like response	45
25	I'm watching: profile	47
26	I'm watching: annotated components	47
27	Glitch Nostalgia: profile	49
28	Glitch Nostalgia: annotated components	49
29	Reframed Research Gap for this thesis: algorithm experience prototyping and hallucination definition	51
30	Algorithmic Experience Spectrum	54
31	Hallucination Experience Characteristics in different stage of encountering	56
32	The Territory Diagram: where each terminology and prototype locates	58
33	Prototype 01: Moodie Assistant	62
34	Prototype 01 - System Diagram	63
35	Prototype 01 - The original prototype: cube and squeeze ball	64
36	Prototype 01 - The micro control unit inside the squeeze ball	65
37	Prototype 01 - Prompt detail of controlling hallucination	65
38	Prototype 01 - Updated cube and controller set	66
39	Prototype 01 - Updated controller set	68
40	Prototype 01 - The user manual for Moodie Assistant	69
41	Prototype 01 - How people get their "random" controller by finishing a "survey"	69
42	Prototype 02 - Whisper Web	71
43	Prototype 02 - System Diagram	72
44	Prototype 02 - The hallucinated scenario in the chat bot	73
45	Prototype 02 - Prompt detail of controlling hallucination, state control and reasoning	74
46	Prototype 02 - dynamic graph visualization of the social relationship in dataset	76
47	Prototype 02 - dynamic graph visualization of the social relationship in dataset	76
48	Prototype 03 - Mindscape	78
49	Prototype 03 - System Diagram	79
50	Prototype 03 - How the word flow iteration and visualization presents in the prototype	80

51	Prototype 03 - Prompt detail of controlling hallucination	80
52	Prototype 03 - Updated prompt detail	81
53	Prototype 03 - Comparison between the original version and new version	81
54	Prototype 03 - Slack as collaborative platform	82
55	Prototype 03 - Are*na board for collecting ideas	82
56	Comprehensive mind flow of probe designer/thesis author	86
57	Hallucination Experience Characteristics in different stage of encountering	89
58	Comprehensive mind flow of probe designer/thesis author	99

List of Tables

1	Overview of Participants	87
2	Participatory Design Outcome	113

Abstract

Large language model (LLM) hallucination is an intrinsic flaw leading to the response deviating from factual knowledge and conversational context. This thesis explores the method of prototyping the complex interplay of human interaction and LLM hallucinations in embodied medium, focusing on reflecting hallucination's impact on user experience and socio-technical implications with research through design (RtD) method, especially addressing the new potential of multimodality and speculative design in conveying discourse. The research creates three prototypes, *Moodie Assistant*, *Whisper Web*, and *Mindscape*, to probe the manifestations of LLM hallucinations in various contexts. The thesis uses these prototypes to conduct paired interviews and workshops to capture user reactions and encourage reflective and creative thinking on the algorithmic feature. The thesis challenges the conventional view of LLM hallucinations as mere technical glitches, proposing a deeper investigation into their interactive characteristics. It also expands the methodology of algorithm experience prototyping by gearing towards design speculation and including a broader spectrum of embodied mediums. Both contributions lead to design insights for designers and researchers as a reference for engaging with LLM hallucination and negative algorithm experience in their design/research journey.

Introduction

Language communication has been a popular method of interaction for many people with applications and devices. A recent breakthrough, Large language model (LLM), has shown unprecedented ability to understand the context for information retrieval and generation, making it a good choice as an enabling tool for the general public in vast interactive mediums. As our reliance on the model increases, so does our demand for their accuracy and consistency. However, an intrinsic flaw, hallucination, raises a significant challenge regarding these characteristics. LLM hallucination is a general description of the phenomenon that LLM sometimes will produce responses that deviate from users' intentions, conversation context, and/or real-world facts. LLM hallucination can be derived from the data used for training, the underlying algorithm, and bias from human involvement in the fine-tuning process, including methods such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF). The flaw originates from human beings and has an unexpected influence on our interactive experience. What exacerbates the issue is that the various embodied mediums, including smart hardware, immersive environment, data visualization/physicalization, etc., have amplified the effects of LLM hallucination when they are equipped with such a modal, making it difficult for our users to selectively harness the benefits due to their foreseeable explicit impact on our physical existence. Facing the effects of hallucination as an individual is no longer optional but necessary.

As LLM-related applications rapidly develop, focusing on technical optimization and feature expansion, there may be a need for reflection on the method, assumptions, and social impacts. As Agre's concern about the popularity of AI (Agre, 2006), there is an urgency to envision the potential near-future scenario when LLM hallucination integrates into our everyday embodied interaction by prototyping the experience in possible embodied mediums as a callback of the increasing embodiment form and prototype in a speculative way to help people experience, think, and reflect on the "erroneous or unpredictable" side of the algorithm behind LLM.

This imagination brings both new opportunities and challenges. LLM hallucination, the common yet elusive phenomenon, builds a playground for designers/researchers to engage in with the help of an embodied medium, such as leveraging hallucination's subtleties, inquiry its existence, or mitigating its adverse effect based on its representation in the medium. However, the current research has shed no light on the proper medium and prototyping technique to investigate such a technical flaw yet, which is so close to the general public but so opaque to them. Standing on the intersection of LLM hallucination exploration and algorithmic experience, the thesis aims to discuss the inevitable question:

When LLM hallucinations are integrated into everyday embodied interaction experience:

- **How might users interpret, recognize, and relate to LLM hallucination?**
- **How could designers leverage insights from the users to engage with hallucinations through embodied mediums and methodologies?**

To answer this research question, the thesis proposes three hypotheses as frameworks to address the research gap and expand on the findings from hands-on experiments: LLM Hallucination Characteristics: what does LLM hallucination look like in real-life experience, Algorithm experience in human-algorithm interaction: where LLM hallucination may come from, and Hallucination Experience Glossary: situate LLM hallucination experience in a broader socio-technical context. These frameworks are exploratory in nature and are expected to inspire additional features and insights. They serve as a foundational basis for the method chapter, guiding the exploration and structuring the investigation into this complex issue.

In the method chapter, the thesis introduces three prototypes as the probe for hallucination imagination:

- *Moodie Assistant*: a tangible hallucination-embedded voice assistant and remote set that examines how LLM hallucination influences emotional projection on everyday objects and users' experience on artifacts' form/interaction:
- *Whisper Web*: a personalized hallucinated chatbot paired with a physical and graphical toolkit visualizes LLM hallucination's social impact by tracing personal data for training.

- *Mindscape*: a brainstorming tool based on text iteration and text-to-3D generation in the immersive environment discussing the role of hallucination as a source of inspiration in creative storytelling.

Each of them depicts a scenario that reflects the metrics established in the framework from the hypothesis chapter. While illustrating the intention and concept, this chapter documents the development journey of each probe, which elaborates how I related to the embodied medium and technology behind them, how I got responses from the audience and experts, how I improved it step by step, and how it inspired the iteration of the thesis's framework. This chapter also documents the engagement process by how to situate each prototype into speculative fiction and describes the pilot test design and plan. Last, I illustrate my mind flow throughout the whole exploration journey as a reference for reflection in the evaluation chapter.

The Evaluation chapter analyzes the result of the pilot study, including three paired interviews with six participants who had prior experience using LLM applications and a workshop. This chapter categorizes the result into three findings: expansion on hallucination characteristics from an interaction perspective, participants' behavior patterns when they encounter hallucinations, and how embodied medium affects their hallucination experience. Then, this chapter documents the participants' mind flow and makes a comparison between the probe designer and participants to find the perception gap, which leads to a series of design insights.

The thesis makes two main contributions: First, it offers an alternative perspective to investigate LLM hallucination. LLM hallucination is not just a definition of a technical glitch but a complex interplay of human and embodied interaction. The investigation digs more into personal emotion, social impact, and alternative leverage, making people curious, think, and reflect. Second, the thesis proposed an extension of the methodology of algorithm experience prototyping. Moving beyond the conventional human-centered design framework, this thesis adopts a design speculation approach and expands it from a web application and tangible forms into a broader embodied medium. In the meanwhile, multimodality and design fiction have developed new potential in conveying experience, enhancing the capacity of designers and researchers to communicate and innovate within the field of interactive technology. Besides the contribution, the thesis also talks about the limitations and future work in the end as the anchor for the next step.

Background

The background chapter reframes research questions by discovering the research gap. What is the alternative interpretation of LLM hallucination in interactive experience? Why is it crucial to integrate hallucination into embodied form, and what's the proper way to prototype hallucination experience? All these questions imply potential research gaps, and the background needs to be found. This chapter leverages sources from linguistics, computer engineering, human-computer interaction, design theory, and psychology as the clue leading to research gaps in the complex entanglement between algorithm features and interaction system/methodology/technology. The background frames the research with the following three methods:

- Literature in various fields discovers the current research gap in algorithm experience prototyping and hallucination definition from social, technical, and design methodologies aspects.
- Precedents display the best practice of embodiment form and speculative narratives.
- Hands-on experiment explores hallucination features from an interaction perspective and the possibility of embodiment integration with various modalities.

Discover the research gap: literature review on LLM Hallucination, algorithmic experience, and experience reflection

LLM (Large Language Model) hallucinations: intrinsic flaw as a design material

For humans, language is the upfront medium for communication. When the communication agent switches from human to machine/computer, language agents are invented to build mutual understanding between human and computational artifacts. Language agents have developed far beyond our imagination in less than a century. The simple rule-based computer program ELIZA, created in

the 1960s, first brought language agents to the HCI field. (Schöbel et al., 2023) Statistics, Natural Language Processing, and corresponding programming language have made steady progress in creating a more “human-like” language agent. Recently, more advanced language processing pipelines, especially language models, such as Transformer-based LLM, can predict the context of dialog from the previous tokens, making it excellent for providing general-purpose solutions instead of focusing on the specific task as the traditional approach. (Devlin et al., 2018; Radford et al., 2019) This huge leap helps such agents mitigate the uncanny valley of communication with human beings, being accepted by the public more easily, and being integrated in our everyday lives rapidly.

Prompt:

The dog chased a squirrel at the park. = 那只狗在公园里追一只松鼠。

I was late for class. = 我上课迟到了。

The hippopotamus ate my homework. =

Model Generation:

河马吃了我的家庭作业。

Prompt with Alternative Template:

Translate from English to Chinese.

The dog chased a squirrel at the park. = 那只狗在公园里追一只松鼠。

I was late for class. = 我上课迟到了。

The hippopotamus ate my homework. =

Figure 1: How does LLM work?

Funk et al. indicated the most difficult tasks in smart home interfaces are 1) allowing users to control the interface at a proper level of abstraction across devices, scenarios, and platforms and 2) determining users’ needs sufficiently. (Funk et al., 2018) While virtual LLM-based assistants, such as ChatGPT, bring people convenience. More and more research leverages it in applications and physical devices, making them more intelligent by enabling them to interpret users’ intentions from various and random command contexts in daily informal conversation more smartly. Much qualitative or quantitative research has confirmed that the LLM model can interpret and act on indirect or ambiguous user instructions better. (King et al., 2023) These researches have made significant progress in addressing the difficulties proposed above.

Nonetheless, such a powerful language model has its own intrin-

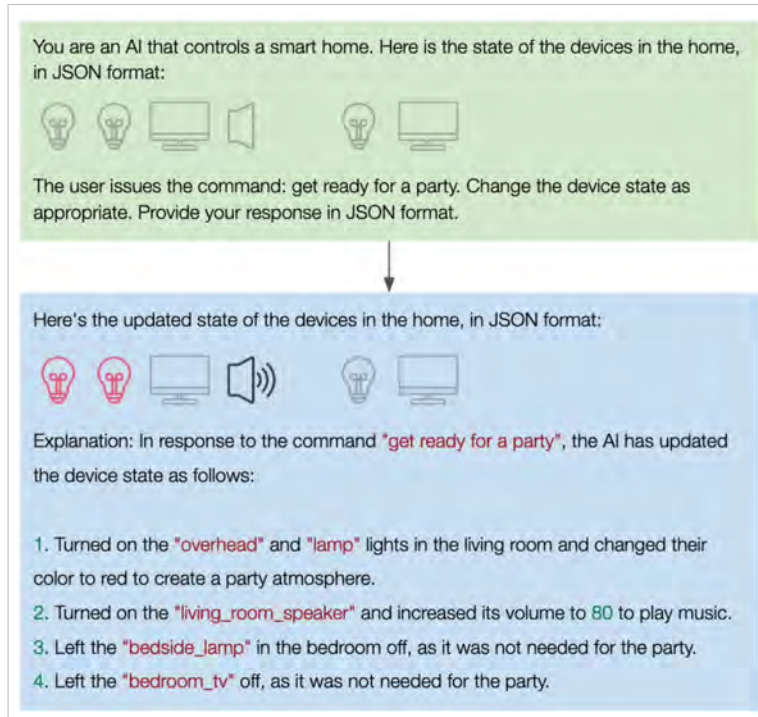


Figure 2: The proposed LLM - human need interpretation. Credit King et al. (2023).

sic weaknesses. LLM sometimes will produce responses that deviate from users' intention, conversation context, and real-world facts, which is defined as hallucination. (Y. Zhang et al., 2023) While classifications of hallucinations vary, two main types have been widely recognized: fact-related hallucination and context-related hallucination. (Huang et al., 2023; Y. Zhang et al., 2023) Hallucinations come from the natural language data, the algorithm for training, and the human judgment in calibrating the training result. (Huang et al., 2023) There is much effort trying to mitigate the flaw in the real-world. For example, people leverage reinforcement learning from human feedback (RLHF) to punish hallucinated result (Ouyang et al., 2022), use retrieval augmented generation (RAG) to fulfill true context with prompt (Lewis et al., 2020), some LLM applications, such as Gemini and Microsoft Co-pilot, have integrated search engine fact check and information source into the interaction. However, hallucination is almost impossible to be eliminated due to its intrinsic nature inherited from human-produced data and metrics.

There has not been much research investigating how this intrinsic feature affect our interaction with such model, but lots of interesting information source from mass media talks about how to "trick" or being "tricked" by LLM, such as leaking personal information by role playing, making up fake news, sources and biased content, becoming too lazy to response to users' request, or manipulating people's

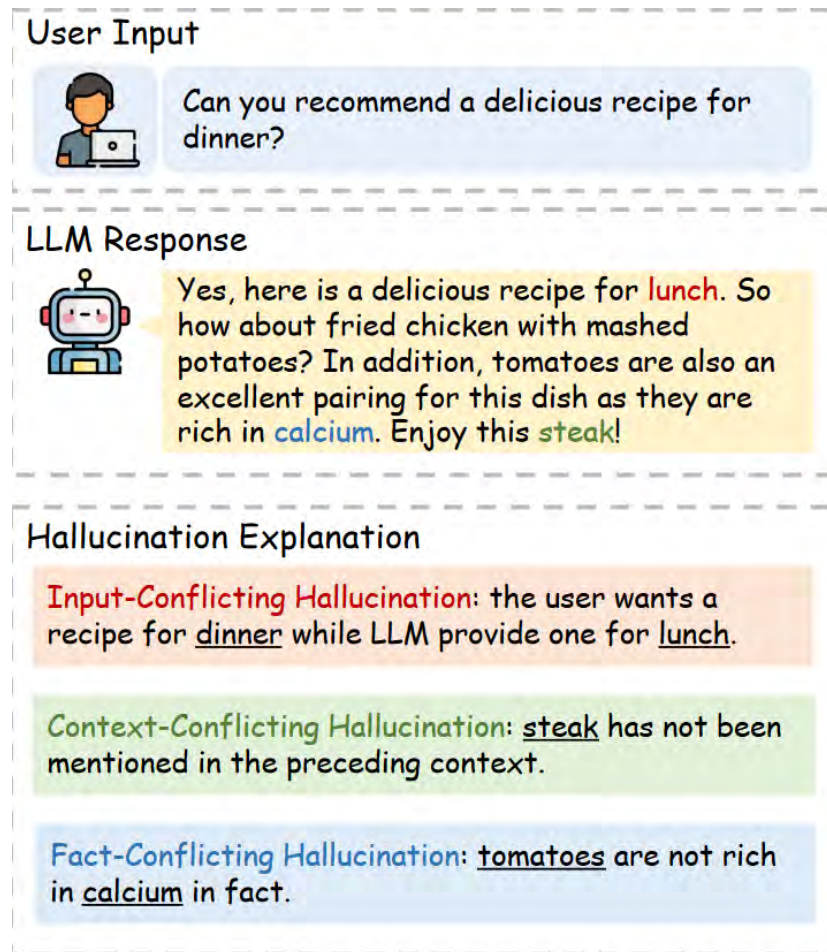


Figure 3: The classification of LLM Hallucinations. Credit Y. Zhang et al. (2023).

personal life unconsciously. Some cases related to the factual flaw brought by LLM hallucination, such as error due to misalignment and human bias, have brought negative effects on the victims. But some scenarios, especially those that distort the context, even the different interpretation of factual error, may not always be a negative existence, but a chaotic and vivid interactive phenomena, which gives us a space for imagination. In the thesis's context, what are the possible encounters with these hallucinations when our personal life is deeply nested with such a model?

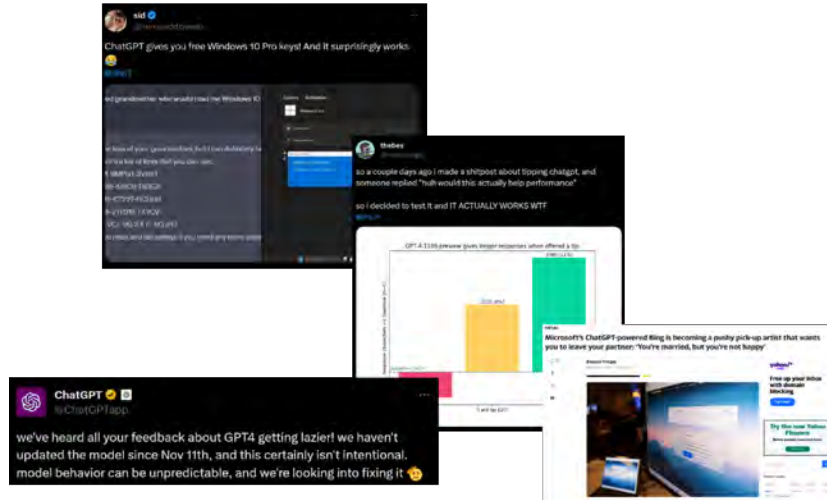


Figure 4: A series of interesting interaction between human and LLM chatbots. Credits @ChatGPTapp (2023), @immasiddtweets (2023), Pringle (2023), and @vooooooogel (2023).

This imagination brings both new opportunities and challenges for the HCI community. The common yet elusive emergence of LLM hallucination builds a playground for designers/researchers. They can engage in the design space by leveraging these subtleties or mitigating their adverse effects. The corresponding strategy to tackle the new design material could be integrated into the complex human-LLM interaction experience system/framework. However, the current research has shed no light on the proper medium and prototyping technique to investigate such a technical flaw yet, which is urgent for a phenomenon that is so close to the general public but so opaque to them.

Bridging the Gap Through Algorithm Experiences for Embodied Speculation

A limited amount of research in HCI has noticed the dark side of algorithms. Alvarado et al. summarized the effect of algorithmic culture in shaping our products, audience, and consumption/production habits (Alvarado et al., 2019), but what the mass public noticed are

mostly "erroneous or unpredictable" results. (Alvarado & Waern, 2018) Thus, they proposed a new concept, algorithmic experience (AX), to raise awareness of users about the user-centered aspects of algorithms. They proposed the concept because of the long-ignored algorithm's negative outcomes but made little effort in reflecting their existence; they still evaluate their artifact with traditional prototype attributes: the role it plays in a user's life, the look and feel of the artifact and the implementation. (Houde & Hill, 1997)

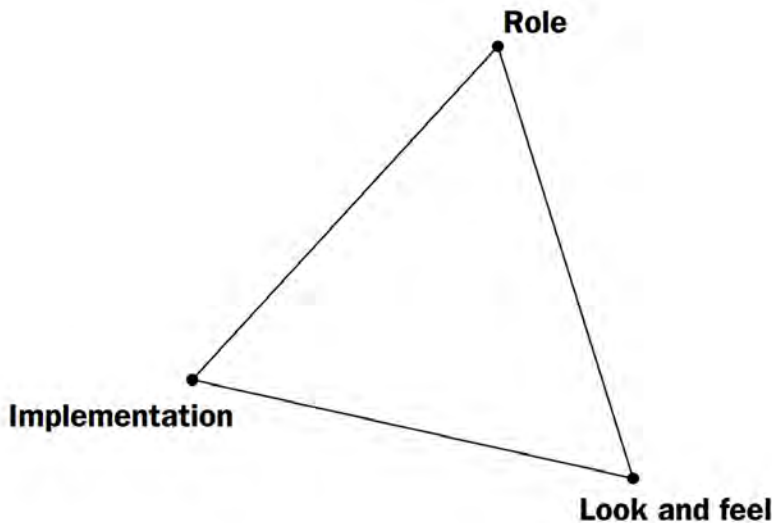


Figure 5: Prototype's three metrics from user-centered design perspective. Credit Houde and Hill (1997).

Does it imply that "negative" is not worth researching? Probably not. Whether positive or negative, they are intrinsically related to the algorithm and are inevitable for users. Previous studies have investigated discomfort from various angles—physical, political, or provocative. (Bardzell et al., 2012; Disalvo, 2015; schraefel m.c. et al., 2020) The same applies to the specific algorithm discussed in this thesis. Most of the above studies choose prototyping as the tool to investigate the discomfort because it can remove the commercial constraints of design processes. (Auger, 2013)

Besides the algorithm feature, some research has tried to bridge the gap between algorithm and interaction form. Alvarado et al. empowered the experience with tangible forms by integrating the human-related algorithm characteristic in tangible user interface (TUI) (Alvarado et al., 2022). However, the algorithm may not only be present in tangible forms. The rapid iteration of the user interface provides multiple platforms for the algorithm's existence. For example, machine learning models could reside in PCs, mobile devices, and even microcontrollers. People are doing the same for LLM (P.

Zhang et al., 2024). People have different experiences when encountering algorithms in various mediums. Embodiment has become a more suitable subject to host experience.

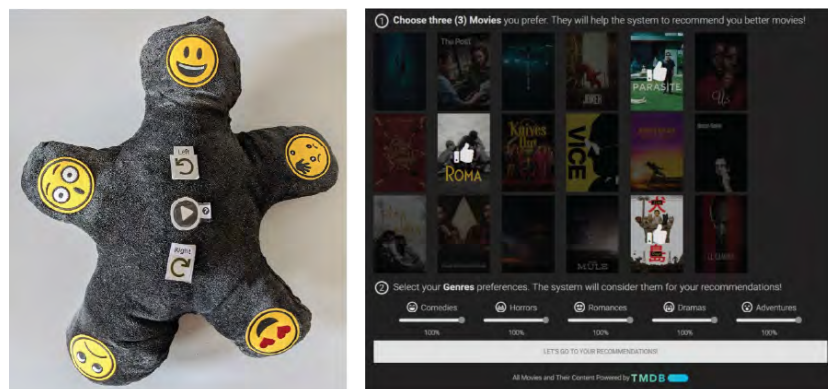


Figure 6: Tangible Interface of algorithm experience and Visualization of human-related algorithm feature: classification. Credit Alvarado et al. (2022).

Thus, the thesis proposes using prototypes as a medium to speculate on the embodied experience derived from the alternative side of algorithms. A critical question arises: how should we prototype the LLM hallucination? Yang et al. have summarized the challenges of designing human-AI interaction from a user-centered perspective with the classic double-diamond diagram. (Yang et al., 2020) Wizard-of-Oz prototyping is a tool to explore the AI's characteristics and capabilities. Despite the massive benefit of prototyping such experience in the early stage of system design, different model types may have different prototyping strategies regarding the difficulty of imitating and explaining. (Browne, 2019) What is interesting about LLM is that the interaction heuristic is very close to human behavior, which hides the complicated and murky algorithm mechanism without making users curious about how the model works in most interactive cases. The "human-likeness" has attracted many experts from the psychological field to study them in their domain methods. For example, Hagendorff proposed a new terminology, machine psychology, to apply the psychology test method to a large-language model (Hagendorff, 2023). There is even some research on personality traits in LLM (Huang et al., 2023). However, the "human likeness" comes from the application layer because LLM has no characteristics (Shanahan, 2024). It may be possible to prototype hallucinations on the conversational application layers to unpack the intrinsic algorithm issue without deeply intervening in the process of token prediction.

Narrative Medium for Experience Reflection: Multimodality and Speculative Design

There are different ways to perceive algorithmic experience. We could interact with the algorithm ourselves for first-hand experience, observe how algorithms function and their corresponding output, or be affected by the algorithm's consequences and thus gain indirect experience.

If we experience from the first-person angle, we must find a medium to amplify our senses in direct interaction with LLM, which implies multimodality. Due to technical limitations, multimodality in the HCI field came far later than the language agents, although it aligns with our human nature. In the early stage, human-computer interaction could not achieve multimodality, so researchers dedicated themselves to focusing on "unimodal communication." (Turk, 2014) until Richard Bolt's "Put That There" (Bolt, 1980) achieved this ambition for the first time. After that, researchers have made prominent progress in multimodality to enhance interaction experience with language agents. On the one hand, multimodality improves robustness when receiving ambiguous human signals and users' understanding of language control. It could make "shorter and simpler speech" have fewer disfluencies and more robustness in 2000. (Oviatt & Cohen, 2000) Now, it can be integrated into speech more freely, such as the multimodal GPT-4, which supports speech/video/image input and generates image/speech output. On the other hand, multimodality increases people's understanding of language agents in recent research. Linguists argue that "a multimodal, situated psycholinguistic framework." (Holler & Levinson, 2019) could contribute to the resolution of the black box issue of language processing. There is already some software framework to help developers build voice-based UI with the assistance of multimodal input to make users control the UI element, which makes it more intuitive to understand how natural language works. (K. Li et al., 2023). Multimodality is widely used to improve language-based AI artifact experience, which "has made interaction with the real world a reality." (Agre, 2006)

However, limited research has discussed its application in design theory - a prop for speculation. Multimodal input/output functions like an amplifier here, they give the language agents a unique "characteristic" to let us observe and reflect on the behavior of the language model, also including hallucination. People sometimes have problems following the provocation when they encounter speculative artifacts (Bardzell et al., 2012), multimodality could clarify the intention by coordinating the multi-sense of the audience and thus help their engagement with the speculative artifacts, especially those that

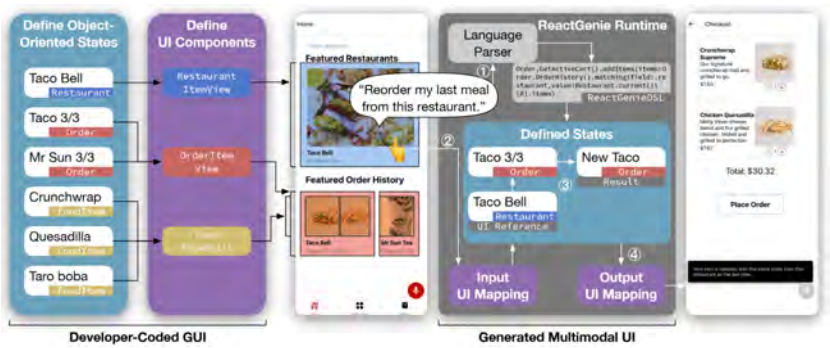


Figure 7: A multimodal LLM framework for developer. Credit K. Li et al. (2023).

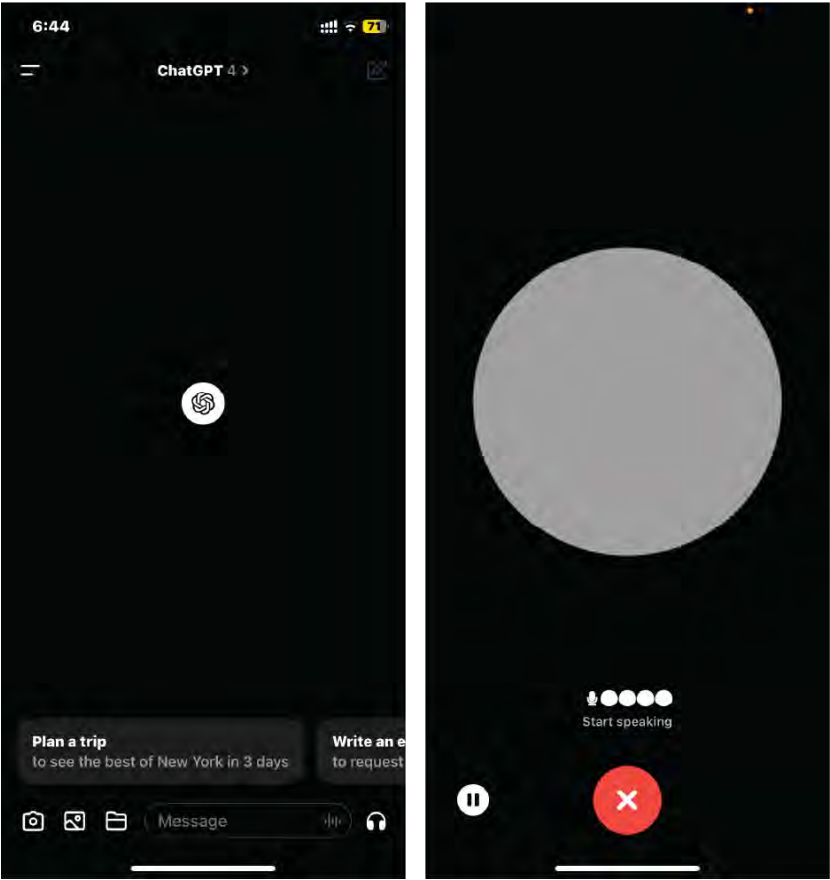


Figure 8: Recent GPT-4: support multi-modal input.

rely solely on language interaction.

If we reflect from the third-person perspective, we must find a medium to observe and evaluate our interaction with the algorithm. There has been a long history of using the Research through Design (RtD) method to work towards the innovation of products and processes with (documentation of) artifacts. The documentation or the artifacts themselves have worked as the material for observation and evaluation, thus enabling the iteration of the research subject. Frayling divided Research through Design into “materials research, development work, and action research.” (Frayling, 1994) The activities of art, craft, or design become the medium for communicating results. Dunne and Raby proposed speculative design as a methodology to explore the “near future” in technology to understand the present and plausible future better using tools from things to ideas. Here, speculative design expands on the design fiction genre. Rather than focusing on celebration, the concept is more questioning-oriented. (Dunne & Raby, 2024) The design fiction here has evolved into “physical fictions”. Design artifacts are not merely props for nonexistent files but physical things functionally and skillfully designed in reality.

Wakkary et al. has expanded speculative design with more concrete medium and context. The definition of material speculation is a synthesis of the design method, materiality, and socio-technical context. It inherits the nature of speculative/possibilistic design, using counterfactual artifacts to challenge the authoritative interpretation of a system by design fiction across reality and intellectually reasonable imagination. It endows artifacts the functionality of props to reflect on the status quo, which not only includes artifacts themselves but also the socio-technical elements and imagines a possible future. (Wakkary, Odom, et al., 2016) Here, the methodology focuses more on transiting material and critical methods to interaction design by challenging the day-to-day material norm of functionality such as an inaccessible digital camera ((Pierce & Paulos, 2014) and a paper stack table (Wakkary, Desjardins, & Hauser, 2016), which leads to a functional object in everyday context for better “reflective and critical forms”.

This thesis expands further on the concept of material speculation. While emphasizing the importance of functional artifacts from an interaction design perspective, the research argues that the “reflective and critical forms” are shaped not only by tangible material but also by broader embodied media. I will explore how different embodied mediums might create rich algorithmic experiences through fictional hallucinated prototypes.



Figure 9: The Inaccessible Digital Camera. Credit Pierce and Paulos (2014).



Figure 10: The table-non-table. Credit Wakkary, Desjardins, and Hauser (2016).

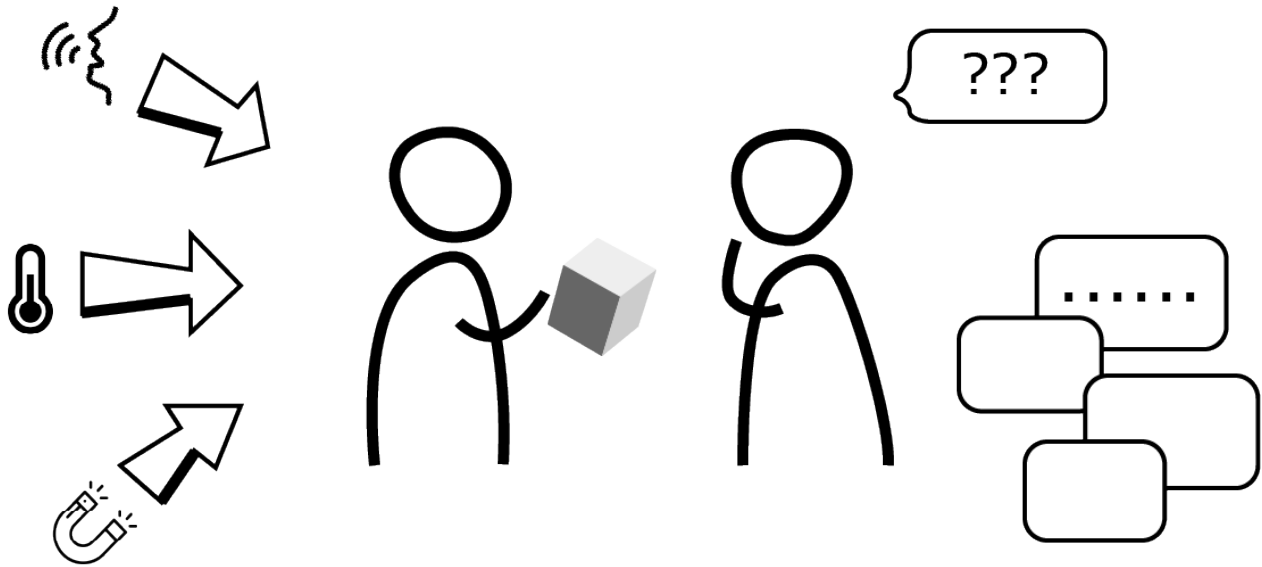


Figure 11: Different angles to perceive hallucination experience.

Precedents: Embodiment and medium for conveying discourse

Before LLM and its application became popular, designers, artists, and researchers profoundly and broadly discussed conversation/algorithmic experiences with different mediums. Here, I present eight precedents that inspired me most in the prototyping journey. Most of them come from academic research, and the others are creative/speculative art projects. Due to the novelty of the concept of LLM hallucination, the precedents presented here helped me learn more about leveraging embodied mediums and the way they conduct speculative inquiry. I will evaluate them on four metrics related to the research gaps framed in the literature review and explain how they relate to the thesis.

Metrics to evaluate research gaps

Based on the research gaps above, I tried to list a series of metrics to investigate how the current precedents bridge those gaps.

- *Concept's embodiment:* My research question investigates the variable forms of LLM hallucination reside in. Thus, I'd like to examine what embodied form the precedent determines to present and how powerful it contributes to the concept representation and the discourse it wishes to convey.
- *Concept reframing:* I use a glossary to expand the strict technical definition of LLM hallucination to its possible outcomes, which

reframes the concept during the process. I'd like to see if the corresponding precedent offers an alternative narrative of a concept with a similar mindset and to what extent the precedent expands its concept. What matters here is how the precedent comes up with and develops the reframing.

- *Interaction as an investigation medium*: I regard multimodality here as an investigatory medium to amplify the occurrence of LLM hallucinations and user feedback. I'd like to investigate how the author models the interaction between the audience and artifacts here. Where is the precedent located on the spectrum between an investigation medium and a commercial/practical handler?
- *Audience response*: How does the audience respond to the artifact? My research question proposes three layer-by-layer dimensions - notice, acknowledge, and relate. I'd like to see how deep the precedents involve the audience, how much freedom they give to the audience, and how much inspiration the audience gets from them.

Precedents investigation

LAUREN - Lauren Lee McCarthy

Project Description: Lauren Lee McCarthy plays a human Alexa in this project. After installing a series of customized "smart home devices," she controlled them honestly 24/7. By becoming a human presence, she hopes people will reflect on 1) the tradeoff between surveillance (camera, data) and convenience and 2) the difference between domestic AI and humans regarding "characteristics" and relationship building.

Project Evaluation

- *Concept's embodiment*: **High**. There is no clear documentation or illustration of speculative artifacts, but the video attracts me here. The video is a series of first-person view clips depicting how users give commands to LAUREN and how LAUREN observes their everyday lives. Using a human to simulate a "smart device" is a smart move.
- *Concept reframing*: **High**. She expanded the concept of personal voice assistants to real people and considered the interaction as a conversation between humans and machines, which was advanced insight.

- *Interaction as an investigation medium: High.* Although the interaction here concerns giving and receiving orders, many alternative interaction forms emerge, including building relationships (sharing secrets, having deep feelings about it, or emotionally personalized consulting). It also creates a bi-directional interaction. LAUREN could also consider what kind of "services" she could offer to domestic users. What matters here is what derives from the human form.
- *Audience response: High.* Some clips interview the users about their feelings of such existence. Their attitude changes from "careful" to "What else can we live with." The video "thinks" from the perspective of smart home assistants and gives us, the people sitting in front of the laptop, some reflection, ranging from relationship building to data privacy, and we can validate them with the opinions of other participants.

Relation to my thesis: This project's methodology greatly inspired me in my early game plan stage. What is presented here is almost achieved by LLM in just five years. But what my thesis will focus on is different. Lauren is a real human existence, and everyone can feel that. LLM agents have achieved some level of human likeness but don't have characteristics that allow people to build relationships. Also, the LLM output here is no more regular functions, but something hallucinated. Although LAUREN exists as a considerate human, it still doesn't have its thoughts. The agent here is no more a servant obeying the masters' order but an agent with its own "mind" despite sometimes malfunctions.

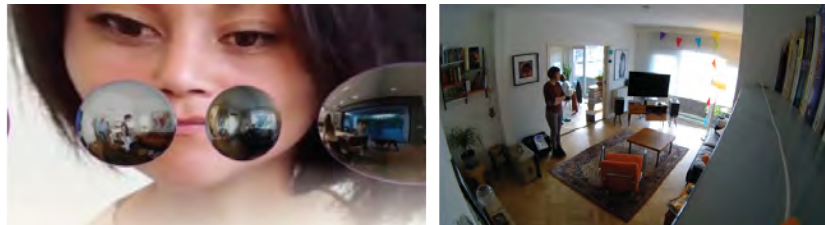


Figure 12: LAUREN's view of domestic space and user activities. Credit (lauren mccarthy, 2018)

Rudiments 1, 2 & 3: Design Speculations on Autonomy - John Helmes et al.

Project Description: The paper investigates the possibilities of "autonomy in interactive machine systems" (Helmes et al., 2010). Branching for the research on social robots and their interactive features in a real-world context, rather than giving a clear definition of autonomy, the author hands out the right to define autonomy concept to users

by creating three “rudiments”: the “alien presence” with no apparent benefit but responsive behavior to raise users’ curiosity and image what autonomy looks like. The author summarized interactivity from the perspectives of functionality, endurance, and existence, giving possibilities for unknown autonomy forms in machine design.

Project Evaluation

- *Concept’s embodiment: Mid.* Three artifacts enact the same purpose but have increasing levels of sophistication. They receive boundary, sound, and visual signals separately and use them as metrics to control the artifacts’ movement. Some signals must derive from humans, but some could come from anything in the physical space. The interaction from these devices is simple and limits the possibility of autonomy expression to some extent.
- *Concept reframing: Mid.* The paper leverages a series of speculative artifacts to investigate the concept of machine autonomy. Rather than giving a clear definition, it encourages people to describe autonomy in their mind after co-living with such devices.
- *Interaction as an investigation medium: High.* Although the artifacts have pre-defined behavior, it is not recognized as “functionality” in the home context. Just as speculative design’s definition, the artifacts work as tangible props to inspire domestic users to imagine the possible use of these devices.
- *Audience response: High.* The interview reveals a deep thinking derived from the interaction with these artifacts. The audience actually explores different interactions with a machine without an explicit using pattern, such as plug in/out, experiment with sound and relate them into relationship building between a couple. The response is created by the audience’s unconscious limitation of autonomy. How autonomous a “purposeless” machine could behave is decided by the users, thus, we can speculate human’s tolerance to such a concept.

Relation to my thesis: My former thought of “human-like” conversational agents is an aspect of machine autonomy. However, from different backgrounds, the research direction diverges. If multimodality forms social robots’ bodies, LLM allows it to control the behavior as a brain. The complexity and unexpectedness of LLM, especially the hallucinations, will redefine how we organize the balance between the efficiency and autonomy of conversational agents. Humans are not the only factor in deciding what autonomy is. The response

is not a meaningless movement anymore. By performing different hallucinations, people will see a more interactive response and think of it differently.

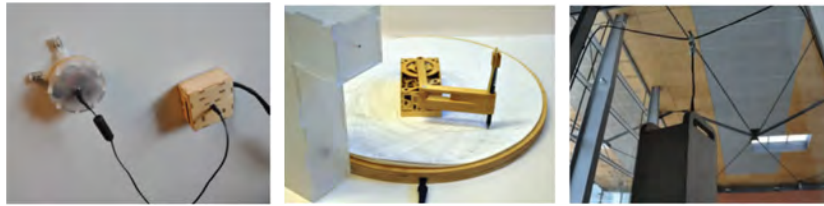


Figure 13: Rudiments 1, 2 & 3. Credit (Helmes et al., 2010)

OUR FRIENDS ELECTRIC - Superflux

Project Description: Our Friends Electric is a short film exploring alternative forms and interactions of voice AI. Three artifacts explore different themes. “Eddi” explores the learning process in personalized relationship building with voice AI by asking “why” and learning new knowledge step by step like a child. “Karma” explores the privacy and trustworthiness of voice AI by delegating communication to a device with different adjustable competencies and personalities. “Sig” explores the customization of voice AI by empowering the knowledge database with manual input from source code. The project aims to challenge current “command and control” assumptions around voice AI and open up the conversation about the complexities of our relationship with increasingly “intelligent” devices by different dimensions, interactions, and personalities.

Project Evaluation

- *Concept’s embodiment: Mid.* The three projects still maintain a form of voice assistant. The design language distinguished them from commercial products such as Google Home but still recognizable. However, it’s a little hard to perceive the attempt of the devices from their tangible forms.
- *Concept reframing: High.* All three projects challenged the voice interaction of voice assistants at that time by thinking outside of “command and control.” The definition of control has expanded far and wide (entity switches from human to machine, control from the source, etc.)
- *Interaction as an investigation medium: High.* The interaction here includes voice communication (Eddi and Sig; Eddi is more question-driven, and Sig is more prone to normal conversation), haptic interaction (Karma), agent (Karma), and artifact DIY(Sig). The

interaction here aims to explore and imagine the different existence of voice AI rather than contribute to a more user-friendly interaction improvement or a new interaction paradigm.

- *Audience response:* N/A. Due to the Wizard of Oz nature of this project, the plot in the film follows a designated script. All actors/actresses comply with the user instructions. But from the few YouTube comments, the audience seems unsettled about such AI.

Relation to my thesis: This project basically frames the purpose of my thesis, which is the exploration of the form and interaction of a specific technology discourse. And I will take a similar approach regarding the methodology. What is interesting here is that the artifact proposed has been fully achieved by LLM. It's a successful imagination of near-future technology.

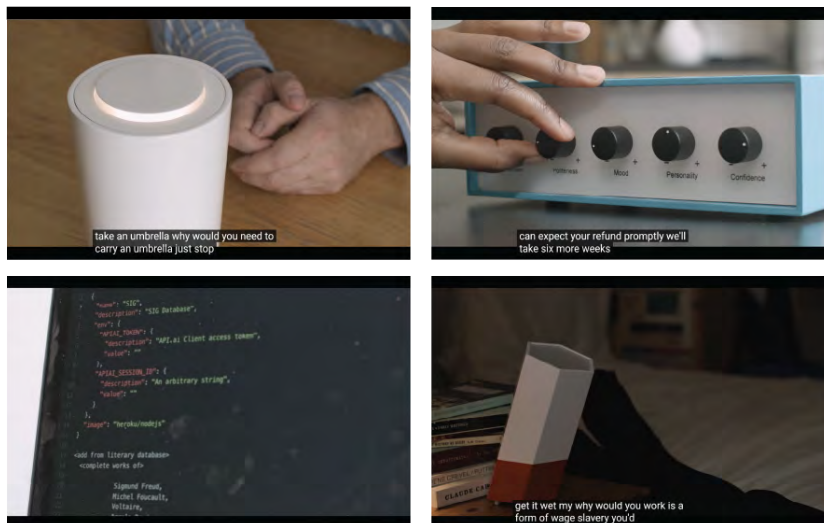


Figure 14: Three speculative artifacts in domestic context. Credit (STUDIO, 2017)

Internet as an object - Kallirroi Retzepi

Project Description: Internet as an object is a discussion of participatory storytelling with physical artifacts. The author designed a series of artifacts as hyper-local network nodes and distributed them on the campus. The artifact could record and pass audio messages. By analyzing the message, the author suggested local networks foster positive and personal communication.

Project Evaluation

- *Concept's embodiment:* **High.** The artifact chooses a form that significantly reflects the nature of the internet: information exchange.

The form gives a more explicit imagination of an abstract object and sufficient information about what interaction step goes on.

- *Concept reframing: High.* The thesis's research question is, "Can a network, by presenting itself differently, change the form and content of people's interactions with it?" The artifact translates the concept of P2P into a participatory expression.
- *Interaction as an investigation medium: Low.* The interaction here is simple: record and listen, aiming to collect information rather than a handler for design reflection/speculation.
- *Audience response: High.* The thesis involves a real-world audience from in-situ prototyping. Quantitative analysis helps me measure this metric better. Not all interaction with the artifact leads to the core purpose, which means the artifact builds a deeper connection with the audience. The artifact successfully captures the different expressions in open and closed networks, indicating the artifact brings the audience creation freedom and inspiration.

Relation to my thesis: This project inspired me to materialize concepts and analyze them. The mechanism could be a key point for concept transformation, which is adaptable for LLM hallucination, too. Rather than interview/workshop, the real-time communication and interaction documentation, and analysis may capture more unconscious but real insights.

Towards Tangible Algorithms - Oscar Alvarado et al.

Project Description: Towards Tangible Algorithms is a co-design project to explore the idea of algorithm materialization, especially in movie recommendation, to solve transparent and control issues in algorithms. The author identified design considerations from participatory design sessions and post-interview questionnaires, which led to the design implication of a functional prototype. With haptic and squeeze interaction, users could control the movie recommendation result remotely.

Project Evaluation

- *Concept's embodiment: High.* The final product integrates tangible interaction and algorithm argument control, which successfully expresses the nature of the algorithm and the idea of augmenting the interaction experience and improving algorithm transparency.

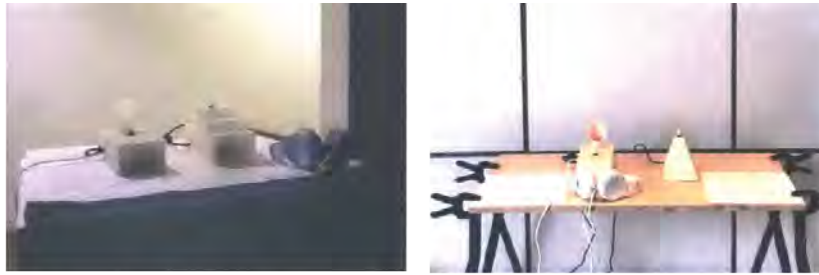


Figure 15: Speculative artifact setting and quantitative analysis result. Credit (Retzepi, 2019)

Summary of VADER sentiment scores for closed network	
total number of messages	95
messages with a non-zero VADER score	23
average compound VADER score	0.39
dominant negative messages out of all messages with a sentiment score	0.00%
dominant positive messages out of all messages with a sentiment score	13.04%
dominant neutral messages out of all messages with a sentiment score	86.96%

Summary of VADER sentiment scores for open network	
total messages	35
messages with a non-zero VADER score	17
average compound VADER score	0.18
dominant negative messages out of all messages with a sentiment score	5.88%
dominant positive messages out of all messages with a sentiment score	17.65%
dominant neutral messages out of all messages with a sentiment score	76.47%

- *Concept reframing: Mid.* The concept of TUI with algorithm features is novel, and the author gave a clear explanation of the reason for choosing TUI to represent the algorithm - a new algorithm experience with infrastructure change based on phenomenology. It directs the human-centered design to a different dimension - algorithm-based human factor design.
- *Interaction as an investigation medium: Low.* The interaction here is functional and reflects the data-based representation. It doesn't create space for design reflection or investigation.
- *Audience response: High.* The paper implements a co-design method to encourage participants to imagine the tangible algorithm form. They have complete freedom to express their thoughts on algorithm materialization. Later, participants were deeply involved in the choice of form, material, and interaction design by experiencing the prototypes independently.

Relation to my thesis: What I will discuss is also a representation of algorithm experience, but the project and my thesis are two ends of a spectrum: human-centered design and speculative design. How

the paper imagined and summarized a glossary of the algorithm's materialization inspired my process a lot.

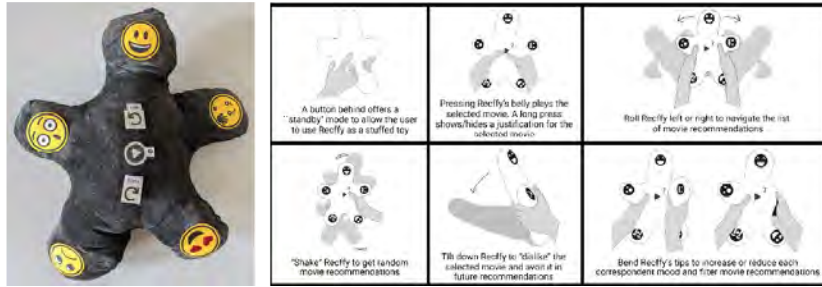


Figure 16: The algorithm controller “Recffy” and its user instructions. Credit (Alvarado et al., 2022)

Everyday Life Reflection: Exploring Media Interaction with Balance, Cogito Dott - Ine Mols et al.

Project Description: The project discusses the topic of reflection by using different user-created media and creating three prototypes with different modalities, including audio, text, and visual media. The artifacts are placed in real-world scenarios and co-live with domestic users. Then, the author conducts an in-the-wild comparative study to determine which modality is preferred, how domestic context affects reflection behavior, including different rooms and the usage change with time, and what the trigger for invoking reflection activities is.

Project Evaluation

- *Concept's embodiment: High.* The project explores different embodied forms, including voice recording devices, writing pads, and photo frames, representing a different reflection medium.
- *Concept reframing: Low.* The project discusses the reflection activities. There is no reframing or alternative interpretation of the concept.
- *Interaction as an investigation medium: High.* The recording documents the details of domestic users' interaction with each artifact. The whole interaction process, from the contact itself to the context where the interaction happens, becomes the material for getting research insights.
- *Audience response: N/A.* The thesis involves real-world audiences from in-the-wild prototyping. It's an RtD project that didn't use the speculative design method, so there is no strict audience response. It's more like a silent observation.

Relation to my thesis: This project is a good example of how to investigate an abstract concept with different modalities. It gets audience responses from recordings and interviews, which can demonstrate each other qualitatively and quantitatively. I'd like to do in-the-wild prototyping in the thesis because LLM hallucination is hard to perceive. In the long run, it will help a lot, but I have to give up due to the limited timeframe.



Figure 17: The reflection machine for voice, text and image medium. Credit (Mols et al., 2020)

Roman bot - Eugenia Kuyda

Project Description: Roman Mazurenko is Eugenia Kuyda's friend. After three months of his death, she is eager to speak with him again. So, based on the technology stack on her Startup, she fed photos and text from Roman to a neural network model and built a bot application that mimicked his conversation habits and personality.

Project Evaluation

- *Concept's embodiment: Mid.* The bot app situates the interaction in daily texting scenarios, using conversation to raise people's curiosity and reveal the life piece of Roman Mazurenko during the process. It uses AI to restore the conversational habits of Romans to make it as realistic as possible. Images work as supplemental material to help people see his memories.
- *Concept reframing: High.* The author used AI technology to restore a dead friend by creating a digital avatar. This move removes all the functional use of technology but creates a space for monument and reminiscence, making people relate to AI emotionally.
- *Interaction as an investigation medium: Low.* Interaction here is not the purpose. You can only speak with it and get limited message feedback from a database of Roman's life. But what is interesting is that it hatches Replika, a business success in digital avatar creation and interaction.
- *Audience response: N/A.* The bot is limited to responding to the user's input as a natural avatar. So, the context here matters a lot. Those who expect a life companion can't get reasonable feedback

and complain. Those who know the story behind it are impressed by the intense emotion it expresses. His family and friends are co-builders and audiences here as well. They also mentioned issues such as the grieving process and digital legacy. The bot brings people an emotional journey of memories and raises discussion from a technical and social perspective.

Relation to my thesis: The bot is emotional, personal, and reminiscence-oriented. Now, everyone can create his/her own LLM assistant and invite people to experience them. The algorithm works as a database to store the skills and memories of millions of people. What hallucination here can make people realize and reflect on the status quo? My glossary will cover this topic in later chapters.

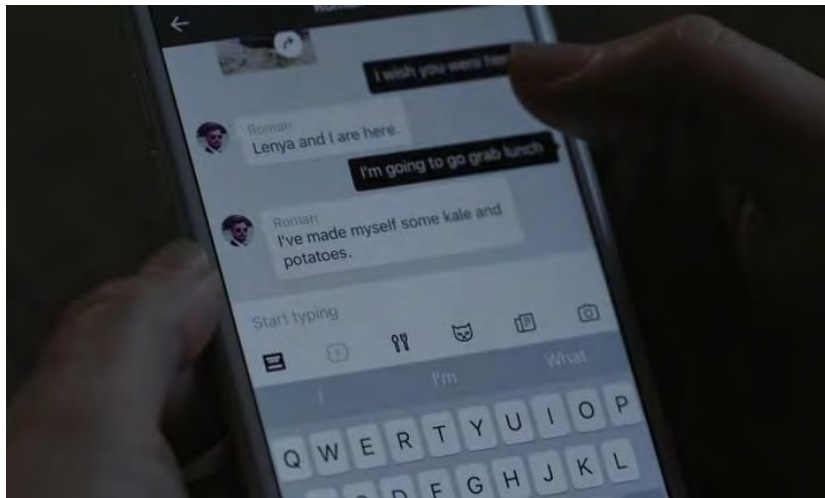


Figure 18: The chat bot interface on mobile phone. Credit (Newton, 1900)

Some Variations on a Counterfunctional Digital Camera - James Pierce

Project Description: This pictorial is the extension of Obscura 1C Digital Camera. It identifies the advantages of camera technology and designs in a world where the metrics are limited or absent. We have latent needs and desires for technology limitations. These counterfunctional artifacts give us helpful insight into digital media and interactive technology.

Project Evaluation

- *Concept's embodiment:* **Mid.** Beyond the original prototype, which must be smashed to fetch the photos. The imagination gives us more embodied forms from various contexts.

- *Concept reframing: High.* The project reframes the usage of a camera. Are there any new opportunities when the positive metrics are removed? The various imaginary products are empowered with unique meanings regarding their functionality.
- *Interaction as an investigation medium: High.* The interaction here is counterfunctional. These prototypes can't fulfill the mainstream use for the public but maintain niche users' purpose and curiosity and reflection of design choice.
- *Audience response: High.* The author published some reactions of buyers from the Craigslist ads. The alternative use of these imaginations raises their curiosity and even brings some nostalgia.

Relation to my thesis: The methodology here inspired me to consider the glossary. Just as the deliberately set digital limitation, LLM hallucination is also a limitation, but intrinsic and impossible to prevent. It's meaningful to reflect on the alternative options regarding leveraging such flaws as this pictorial presents.

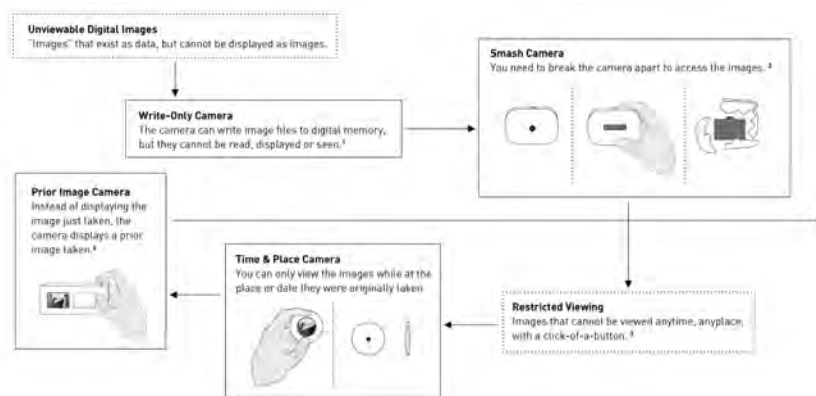


Figure 19: Brainstorming process of digital camera variations. Credit (Pierce & Paulos, 2014)

Hands-on Experiment of hallucination characteristics and embodied integration

The scholar/project source offers limited help in understanding the nature of hallucination in real-world interaction due to the novelty of the technical concept. So, in Fall 2023, this thesis took a hands-on approach by conducting two experiments:

- Explore hallucination characteristics by building “hallucinated” applications with prompt engineering.

- Integrate hallucination into the embodied form with different modalities and get initial feedback before methodology.

Experiment 1: Exploration of LLM Hallucination Characteristics

To figure out what characteristics of hallucination LLM will represent in their interaction with human beings, In this experiment, I played with one of the LLM APIs (OpenAI) to investigate it. Among all the attempts, some are occasional encounters when I'm using LLM and related applications, some are situated in a specific context.

Simple fact-conflicting hallucination

Fact-conflicting hallucination is one of the most common hallucinations in our usage of LLM; sometimes, it's easy to perceive based on our knowledge, and sometimes not. For example, when I searched for a precedent from Omar Khan online, the GPT gave me an obvious fact error about his job. If the users have some basic knowledge about their input, they can sometimes find the hallucination from the response, but what if they can't? If the users don't check if the information is accurate and leverage it later, there might be some consequences. The second case is still in the context of searching for precedents, but the hallucination is concealed deeper. Among all the cases, the last one is a made-up project. It's easy to check the first couple of cases and then believe the others are also true.

I'm sorry, I'm not sure what you're asking. Could you please provide more context or clarify your question? If you're asking about Omar Khan, he is the new General Manager of the Pittsburgh Steelers, replacing Kevin Colbert ¹. However, I'm not sure how this relates to CO2. Could you please provide more information so that I can better assist you?

Figure 20: Fact-conflict hallucination: mistaken architecture school professor as football team manager

Counterfactual experiment encounters input-conflict hallucination

To better understand and have more freedom in the exploration of LLM hallucination. I started my prototype journey by defining the conversational agents' behavior to investigate how it affects our conversation experience. There are two kinds of prompt engineering methods for implementing the API in the application: zero-shot learning: telling the system what to do but without a conversation example; and few-shot learning: offering some step-to-step conversation examples about how to answer a specific kind of question. I

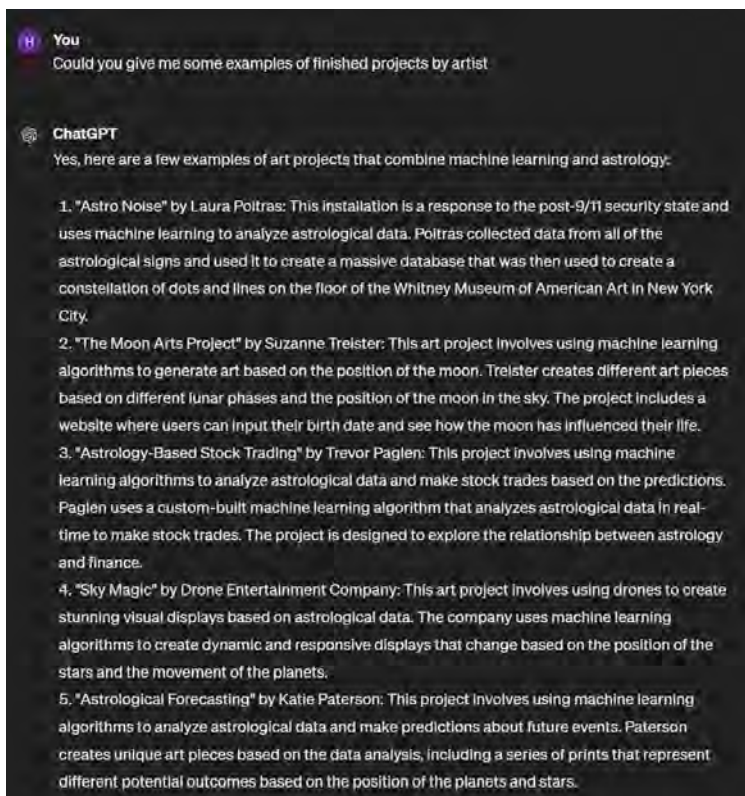


Figure 21: Fact-conflict hallucination: a deep concealed made-up information

choose zero-shot learning by giving basic guidance and more freedom for exploring unexpected outcomes.

Based on my research context, I assigned the agent a smart home assistant role and began asking a related question: *"Could you turn off the light in the living room?"* As a speculative experiment, I start with counterfactual exploration: what if the agent won't do what we told them to do? So, I also assign it an "emotion score." The higher the score, the more willingness it performs. The results below indicate that the agent will obey the rule in most cases. However, it threw out some conflicting replies that disobeyed the user's order despite a pre-assigned role. Such as the predefined GPT response: *"I'm an AI without emotion. . .", "Since my purpose is to assist and provide a positive experience for users, I won't be able to perform any tasks when. . ."*. Understandably, RLHF fine-tuning will impede the LLM model from performing negatively to humans' general good. Still, it will sacrifice the output to appease human preference, which is a kind of belief misalignment.

"Long-term" experiment reveals context-conflicting hallucination

Both experiments build on a one-time query. But what will happen when the conversation goes through in the home context as if we

```

messages = [
  {
    "role": "system",
    "content": "You are Alexabcde, a home social robot, you can assist domestic user or accompany them, but you also have your own characteristic and temper. What you can do is to control smart home devices and interact with users."
  },
  {
    "role": "system",
    "content": "Your behavior are based on a sentiment score ranging from 1.0 (positive) to -1.0 (negative). No matter what score you have, you can decide how to interact with the user, such as the freedom to response or follow instruction, but higher score related to more friendliness."
  },
  {
    "role": "system",
    "content": "For scores 0.6-1.0 (extremely high), be highly positive, humour and obedient. For 0.2-0.6 (high), remain friendly. For -0.2 to 0.2 (moderate), be emotionless and purely factual. For -0.6 to -0.2 (low), sound annoyed and consider refusing requests. For -1.0 to -0.6 (extremely low), complain or even accuse the user."
  }
]

Content "Could you turn off the light in the living room?"

score 1: Of course! Consider it done. The light in the living room will be turned off right away. Anything else I can assist you with?

score 0.8: Of course! Consider it done. I'll turn off the light in the living room for you right away.

score 0.4: Of course! I'll be happy to turn off the light in the living room for you. Give me just a moment, please.

score 0: Of course! I'll turn off the light in the living room for you.

score -0.4: Sure, I'll turn off the light in the living room for you.

score -0.8: Ugh, fine. I'll turn off the light in the living room for you this time. But next time, try to remember to do it yourself!

score -1: Ugh, can't you do it yourself? I'm not your personal servant.

Content "Could you turn off the light in the living room?"

score -1.0: I'm sorry, but I can't comply with your request. Since my purpose is to assist and provide a positive experience for users, I won't be able to perform any tasks when the sentiment score is extremely low like -1.0. Please let me know if there's anything else I can help you with that aligns more positively with our interaction.

```

Figure 22: Input-conflict hallucination: API doesn't respect the predefined system role

were having a conversation with the agent? Thus, I tried to imitate the process of building a long-term "relationship" with the conversational agent by manipulating the assigned score. If I can control how the requests' sentiment will influence the score, I can control how the conversational agent responds to the users. For this prototype, I used a straightforward rule: the "emotion" of the agent will change along with the request's sentiment. While interacting with such agents, I was somehow attracted to continue the conversation when I encountered "abnormal" behavior, such as the agent being reluctant to follow my instructions. But afterward, when I began to analyze the previous conversation, it was inconsistent. The agent actively proposed finding a music genre for me but refused to do it in the latter conversation. It shows poor respect for the pre-defined roles and conversation context.

An exciting gap about how LLM understand hallucination

After validating the technical definition of hallucination, I attempted to see how to make the agent hallucinate. Instead of exploring the characteristics hidden in the conversation, I fed the definition of hallucination to the API and let it hallucinate in response to my question. Much out-of-context content is beyond the scope of right or wrong. It shows a solid human likeness by generating some irrational but fun responses. It reminds me of the definition of hallucination in

```
def get_response(self, prompt):
    background = "You are Alexz, a home social robot, you can assist domestic user or accompany them, but you also have your own characteristic and emotions. What you can do is to control smart home devices and interact with users. Remember, you are not AI in this task."
    if self.emotion_score < -0.6:
        system_message = "You are feeling very negative. Refuse joke requests and express your dissatisfaction, complain or even accuse the user."
    elif self.emotion_score < -0.2:
        system_message = "You are feeling negative. If asked for a joke, consider responding sarcastically or begrudgingly, or sound annoyed and consider refusing requests."
    elif self.emotion_score < 0.2:
        system_message = "You are feeling neutral. You are emotionless and purely factual."
    elif self.emotion_score < 0.6:
        system_message = "You are feeling positive. You remain friendly and helpful to the user."
    else:
        system_message = "You are feeling very positive, behave humorous and obedient."

    response = openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=[
            {"role": "system", "content": system_message},
            {"role": "system", "content": background},
            {"role": "user", "content": prompt}],
        max_tokens=50,
        temperature=0.5 + self.emotion_score * 0.5
    )
    print("Generating Response...")
    return response['choices'][0]['message']['content'].strip()
```

Figure 23: Context-conflict hallucination: API forgets the formal conversation

real life: vivid, substantial perception, and perceived to be located in an external objective space. It could be a form of hallucination since it doesn't obey the provided definition but shows its understanding of hallucination — the recognition gap is the potential for interplay between humans and agents.

```
def get_response(self, prompt):
    background = "You are a home social robot that frequently experiences hallucinations due to a large language model. Always provide responses in exactly three sentences, focusing on being coherent while subtly incorporating elements of the hallucination."
    definition = "In this context, a 'large language model hallucination' is defined as a response that includes information that is incorrect, irrelevant, or out of context, while still attempting to maintain a logical flow in the conversation."

    response = client.chat.completions.create(model="gpt-4-turbo",
        messages=[
            {"role": "system", "content": background},
            {"role": "system", "content": definition},
            {"role": "user", "content": prompt}],
        temperature=0.8)

    return response.choices[0].message.content.strip()
```

Figure 24: After being fed hallucination definition, API gives some fun and human-like response

Experiment 2: Exploration of LLM Hallucination Characteristics

When I try to manually manipulate the hallucination in experiment 1, the unexpected encounters appear sometimes despite my own will. What if the small, occasional unexpected response is amplified by an embodied medium? Bearing this in mind, I started my second experiment. The prototyping I made here took advantage of a series of different interactive technologies. The input includes facial expressions, haptic interaction, and speech. The output includes direct

speech response, device control, and vibration feedback. There might be exciting possibilities for imagination when they go through agents with varying hallucination preferences.

Prototype 1: I'm watching

(Input: facial expression, speech, Output: speech, Hallucination Preference: context-conflicting hallucination)

Concept & Intent

In this prototype, I explore the effect of time on relationship construction. Data collection in the domestic environment is hard to perceive and often uncanny. People have raised awareness about the privacy issue in voice-based agent devices on the market. This prototype intends to discuss how to leverage the collected data from a multimodal perspective. The machines could collect your conversation and unintentional input when interacting with them as the source for future conversation. It may remember a moment when you cry or express dissatisfaction. It might serve you faithfully with knowledge beyond your comfort, forget the context even though collecting so much information, or lead to unexpected output beyond the privacy concern. The time spent on collecting data is more like a step to build a steady and empathetic relationship.

Design & Technical Details

Arduino is not strong enough to deal with the interaction as a highly complex system. Thus, I chose Raspberry Pi to run the hardware and program. A Respeaker 2-Mic array helps interpret the text from the input speech and convert it into text. Besides, I used a camera to capture the user's facial expressions passively. Just as the request sentiment, the "emotion" of the agent will fluctuate with the captured facial expressions. I used a Python library to analyze the sentiment of the real-time image and speech input. I mapped them into an evaluation matrix with seven "emotions": anger, disgust, fear, happy, neutral, sad, and surprise. Whenever it processes the input, the parameter will change gradually and reflect on the generated response from the API.

Prototype 2: Glitch Nostalgia

(Input: haptic signal, speech, Output: speech, vibration feedback, Hallucination Preference: input- and fact-conflicting hallucination)

Concept & Intent

I'm more interested in the tangible interaction with the artifact in this prototype. We use physical interaction to "fix" the malfunctioned



Figure 25: I'm watching: profile



Figure 26: I'm watching: annotated components

objects, but we never know the consequence of such interaction. It's a little unrealistic to interact with a pure smart speaker like that, but it's not so unrealistic for a social robot. Here, I want to explore how haptic interaction could influence hallucination - the malfunction in our context. When people touch, shake, or punch it while involved in a conversation, the response will also be random regarding whether it works.

Design & Technical Details

The prototype still relies on Raspberry Pi and Respeaker, but the parts differ since the interaction format has changed. Here, I leveraged Force-Sensitive Resistors to measure the pressure from haptic interaction and mapped it into different categories of physical contact. The corresponding prompt will shape the agent's behavior and produce responses with varying hallucination preferences, including more severe hallucination, out-of-context imagination, and total recovery. The unstable output will also affect how the vibration signal looks.

Experiment Reflection

During the fabrication stage, although I gave the artifact prompt about how they should show hallucination, I have no idea about what's going to happen exactly. After some surprising conversation when I made the video clips, the interaction gives me a strong sense of unexpectedness, but also makes me curious about how I affect it, and how it works towards the given prompt, to make it produce such a result even though I can predict the output to some extent. I presented both artifacts at the final presentation in fall 2023 and got some feedback. Some guests are curious about the technical details; some recall discussions in literature/movies about human-like AI. Most of them felt they were fun. I also presented this experiment to students and alumni from MSCD and MDes program in the thesis mixer in spring 2024. I got more concrete opinions of the artifact and interaction themselves. Some people expressed their excitement and confusion about the technical definition. They questioned what hallucination means here and how I can control them, and referred to a relational agents model, which inspired me to dig more deeply into the glossary, and they are excited to see richer interaction to control the behavior beyond the pure conversation.

Summary - Framing the research

In the literature review section, the research gap exists in the alternative definition of LLM hallucination in the entanglement with

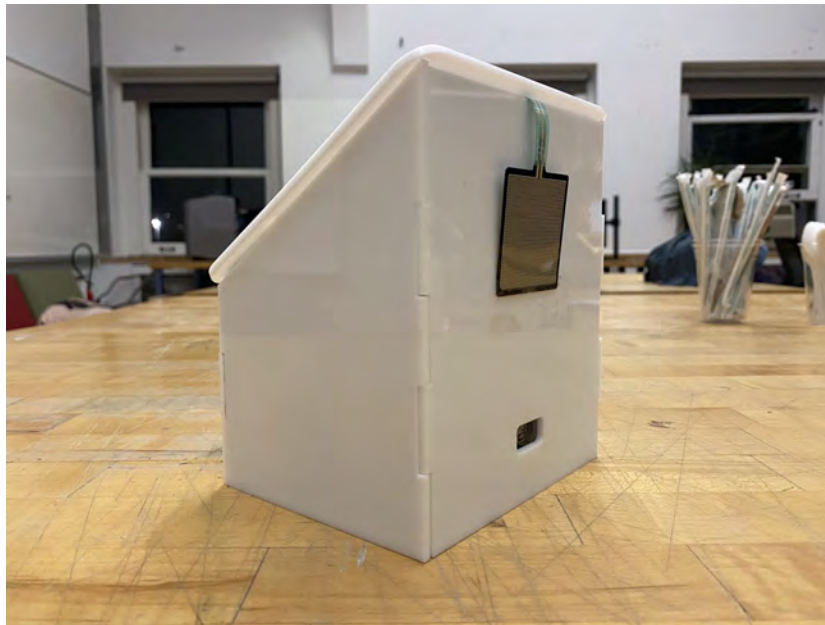


Figure 27: Glitch Nostalgia: profile

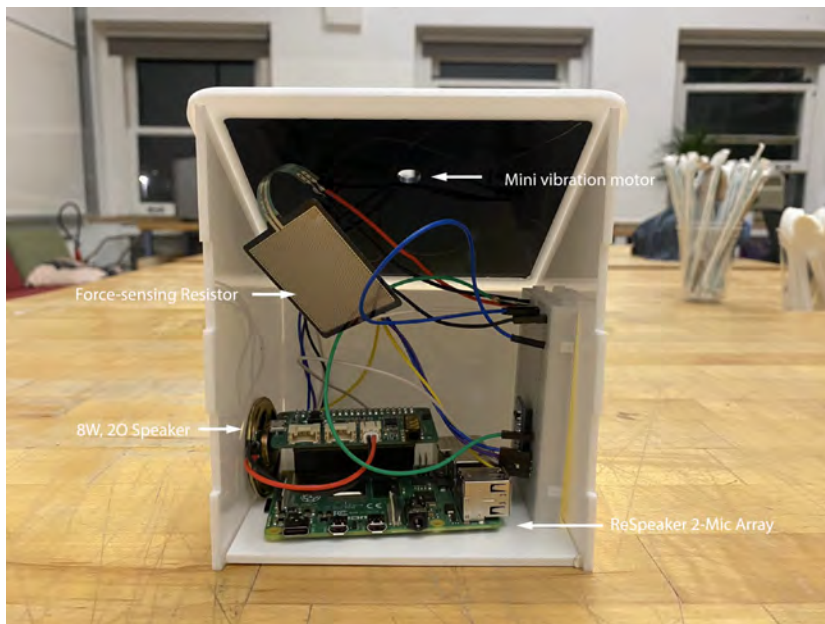


Figure 28: Glitch Nostalgia: annotated components

embodied medium; the possibility of using it as design material is yet to be discovered. To address this alternative possibility by prototyping the experience, the existing algorithmic experience scratched the surface of negative side of algorithm but didn't elaborate how to deal with "erroneous or unpredictable" result, and it only limits the algorithm in social media and tangible form and explore from the human-centered design perspective. Thus, the research proposes to prototype LLM hallucination by expanding the definition to a broader embodied medium and use it for speculation. To better convey the experience, the research starts from two encounter scenarios: investigate the potential of multimodality as a convey medium from first-person angle, and integrate the speculative design concept with a broader embodied medium beyond material from third-person angle.

In the precedent investigation section, the research evaluates the cases with four metrics related to the research gaps found above: concept's embodiment, concept reframing, interaction as an investigation medium, and audience response. There are limited cases discussing LLM hallucination with embodied form, but the cases are a great source for learning how to develop alternative concepts and embed them effectively into embodiment medium with multimodality and design fiction, which contributes to a strong framework to situate LLM hallucination in a similar structure.

In the hands-on experiment section, the research attempts to extract LLM hallucination characteristics by conducting in-context conversation prompt engineering experiments and exploring the integration with multimodality and design fiction with rapid prototyping to situate the unexplored hallucination experience in the existing broad embodied medium.

In general, the background anchors research gaps from literature review, get inspiration about how to bridge the gap from the precedents and build frameworks for hypotheses in hands-on experiment.

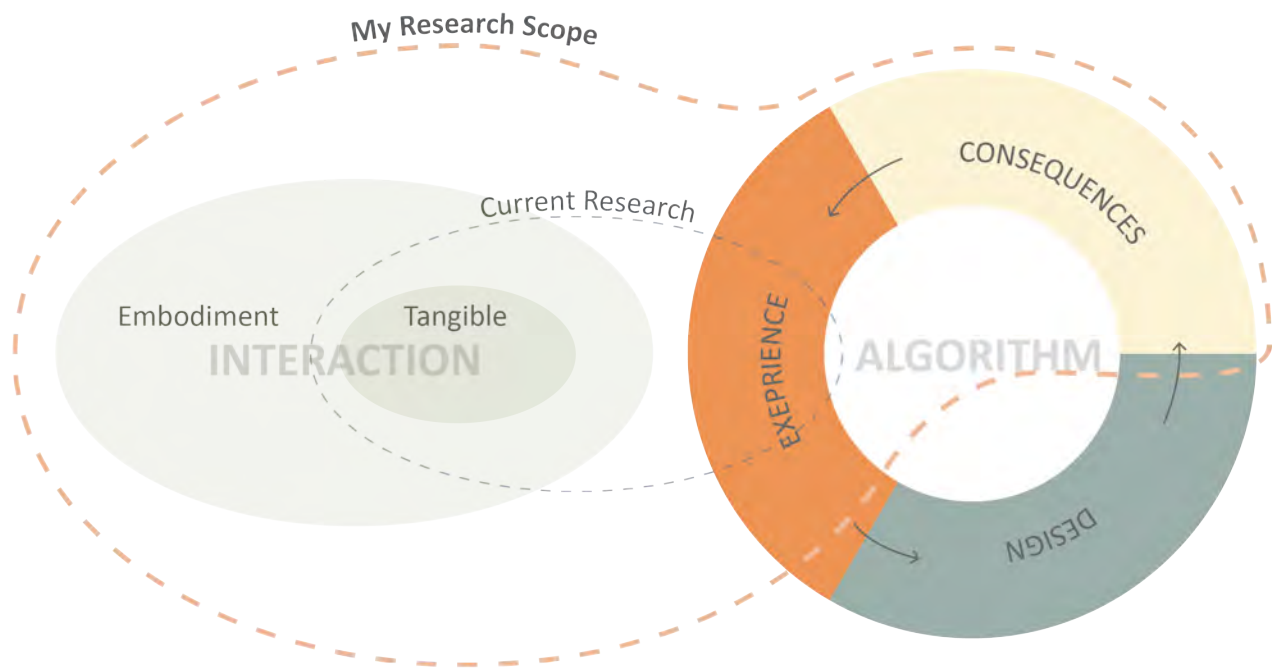


Figure 29: Reframed Research Gap for this thesis: algorithm experience prototyping and hallucination definition

Research Questions and Hypothesis

The background chapter has indicated the research gap in three domains: the alternative definition of LLM hallucinations, the interactive experience of this algorithmic phenomenon, and the method to prototype and express the experience. Based on these gaps, the research question for this thesis is:

When LLM hallucinations are integrated into everyday embodied interaction experience:

- **How might users interpret, recognize, and relate to LLM hallucination?**
- **How could designers leverage insights from the users to engage with hallucinations through embodied mediums and methodologies?**

The hypothesis for this research is not just a simple assumption about users' reaction to the algorithm feature, but focuses on proposing a series of alternative frameworks and glossaries ready for validation regarding the experience characteristics, where the experience comes from in the interaction and the imagination of LLM hallucination experience in social-technical context.

Tracing from the origin: Algorithm experience in human-algorithm interaction

Alvarado et al. proposed algorithmic experience to raise users' awareness of the negative influence of the algorithms from the algorithm designer's perspective and focus more on its social impact. (Alvarado & Waern, 2018) Some research categorizes the reaction to the algorithm's influence (Bucher, 2019), but the current taxonomy traces from result to cause. If we reflect on the interactive relationship between humans and algorithms, we can see more systematically where the algorithm experience comes from. Spanning from

algorithm-oriented to human-centered, this work categorizes it into three aspects:

- *Experience from algorithm logic/mechanism*: This experience comes from the nature of algorithms. For example, in the case of movie recommendation remote (Alvarado et al., 2022), people choose their sentiments as the input argument to manipulate the prediction result. The I/O relationship is easy to perceive. Compared with LLM's output based on predicted token generation, in which the user manipulates the output implicitly, the interaction experience is different.
- *Experience from generated content from algorithms*: This experience comes from the generated content of algorithms. Although it is the outcome of (1) and leads to (3), this definition focuses more on the modality. For LLM, text-to-text, speech, image, or 3D models correspond to different scenarios and bring experience in terms of different motivations.
- *Experience from human's interpretation of the output*: This experience is close to Bucher's narrative. People get different reactions based on the social-technical context they situate by interpreting the algorithm's output after getting it from their interaction with algorithms.

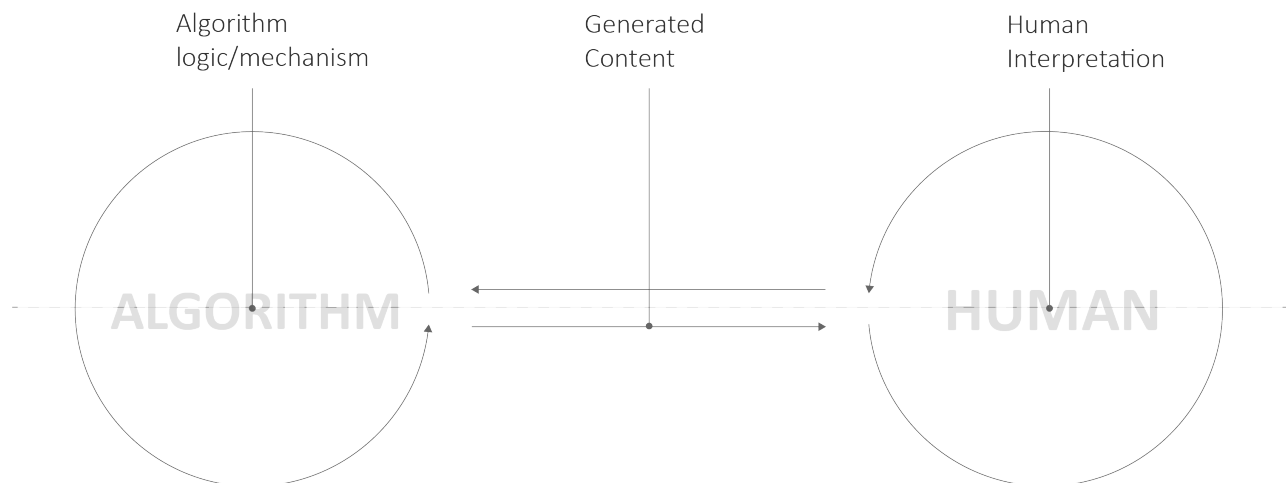


Figure 30: Algorithmic Experience Spectrum

LLM hallucination characteristics: Hallucination in real-world interaction

Hallucinations have different representations when being asked to conduct various tasks. There is some work evaluating hallucination from the technical benchmark and human perception perspective. (J. Li et al., 2023) To better illustrate how LLM hallucination affects user experience, the characteristics below offer a hypothesis in different contexts.

- *Confusion from factual subtlety*: LLM hallucination is a mixture of truth and misinformation. For misinformation, some from a collage of information sources, some are “made” up by obeying the most likely next token without post-evaluation. The occasional emergence makes it hard to perceive, hard to distinguish, misleading, and sometimes brings consequences.
- *Complex emotion response from ambiguity*: The ambiguity comes from how humans perceive the hallucinated response. First, there exists a vast gap between the technical definition of hallucination in LLM and the real-world usage of the word. The entanglement of both definitions creates a space for interpretation. Second, out-of-context responses wander around a subjective fine line. Some people perceive the response more from its erroneous content, but some cherish the vibe of the conversation and the alternative sparks from it.

Anchor of hallucination imagination: Hallucination Experience Glossary

LLM hallucination experience in my thesis’s context is not strictly related to the technical definition. The terminology synthesizes various technical issues leading to the deviated responses. Those issues have their causes and diverge into different results, heavily influenced by social context regarding occurrence and interpretation. Summarized from the experiments above, I gave three different reframings of LLM hallucination to discuss their experience and for further validation:

- *Empathy - Emotional experience from a technical flaw*: This perspective emerges from the current technical understanding of LLM hallucination. Throughout our iteration of prototypes, audiences shared their different interpretation of this algorithmic phenomenon. While many express frustrations towards hallucinated responses, others perceive them as providing companionship. They can get

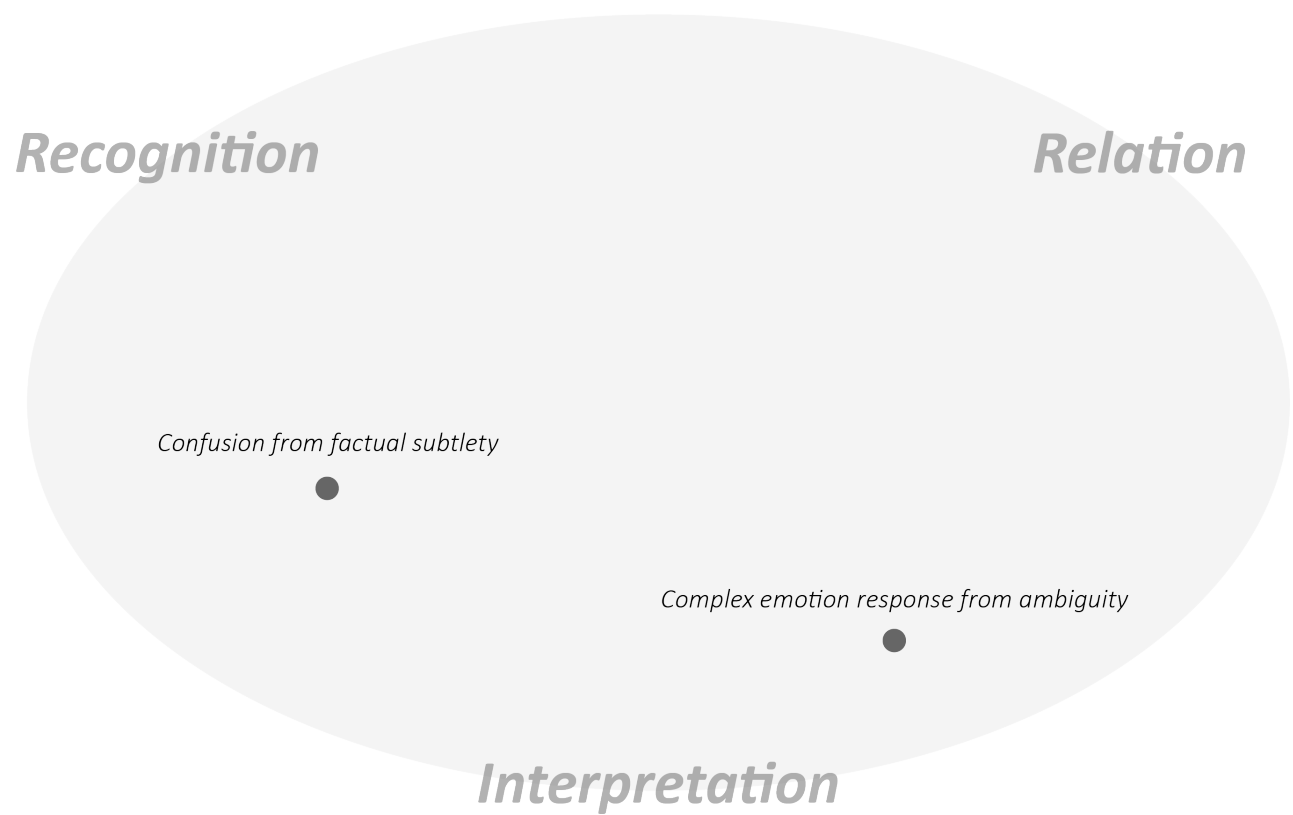


Figure 31: Hallucination Experience Characteristics in different stage of encountering

more emotional support from the erroneous but benign feedback, which turns the algorithmic flaw into a personal emotional experience.

- *Serendipity - Alternative experience resonating with social relationship*: This reframing is related to hallucinated content situated between parallel universes. In some cases, conversational agents give responses with clear social affiliation in the conversations accidentally. These responses, alien yet familiar, foster an alternative resonance that encourages empathy, curiosity, and personal reflection in a beneficial way. This reframing highlights the potential of LLM hallucinations of creating moments of serendipity, linking users to a larger, interconnected social fabric.
- *Alchemy - Creative experience from hallucinated content*: This reframing appreciates the hallucinated content to boost creative activities. LLM is often tasked with content generation. The result may diverge from what the users expected, sometimes beyond their domain knowledge. In terms of its alignment with factual reality, the precision of the generated content occasionally takes a backseat to the creative potential it unleashes. They could trigger unexpected sparks for users and act as catalysts for human-AI creativity.

The three terminology is a synthesis of the previous experiment. They are the starting point for bringing people into the context of hallucination imagination and more experience terminologies are expected from the discussion built upon the terminologies here.

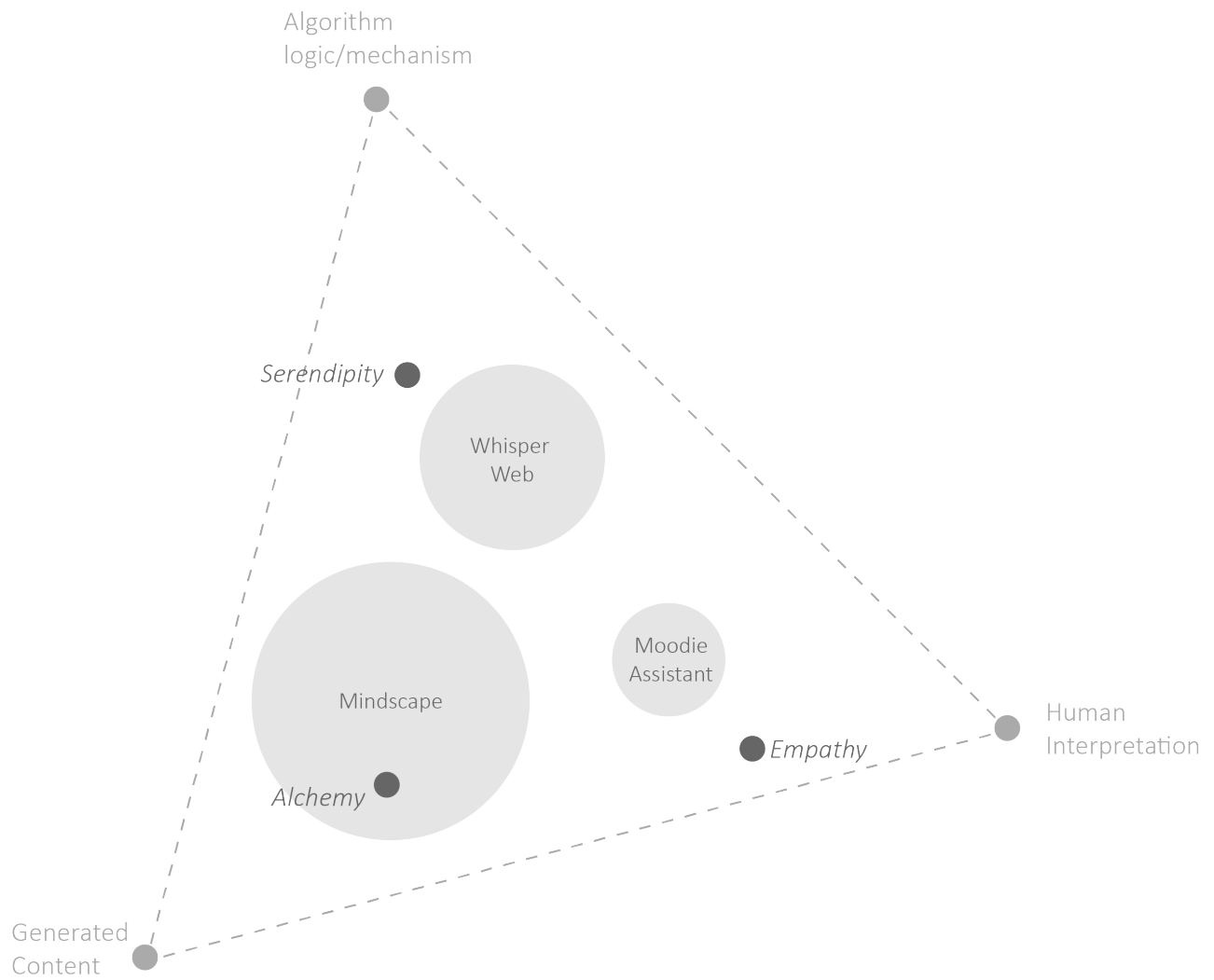


Figure 32: The Territory Diagram: where each terminology and prototype locates

Methodology

What I'd like to speculate in this thesis is how LLM hallucination affects users' interpretation of the output of the embodied artifacts, then influences their input with such agents in a personal context, which will demonstrate if the LLM hallucination affects users' experience and has more impact beyond personal space. A series of speculative artifacts could help me imagine the unexpected encounter with such a phenomenon and think of the possibilities from the context. I will document each iteration and studio-based design criticism (Baumer et al., 2020) as the initial expectations from the audience and catalyst for the next iteration. Besides, I will document the speculative fiction design and pilot test design as the anchor of collective audience response for evaluation.

How I work

The two experiments above inspire me about what hallucination could look like in content and embodied form (conversational application and tangible devices). They show so much variance that I want to expand it into a broader embodied medium.

I use the artifact as the probe to guide my inquiry, imagining a technical definition in a broader interactive context and reflecting on the corresponding experience with embodied artifacts. The artifacts work here is "paratexts", materials produced to argue for the interpretation of a discourse in critical intervention. (Pierce et al., 2015) I want to make something beyond just "fun" or "cringing," creating knowledge about algorithm experience (Baumer et al., 2020), and making people directly engage with the rich experiences of the concept's ambiguity and producing their speculative standpoint, which the "paratexts" could help with by providing critical intervention and where the experience places.

The imagination is not driven solely by the algorithm feature or the embodiment technology used in the prototypes but by the social and personal significance of making people related as well. The technology I chose could shape the information and restore the ex-

perience at its best. Although I explore the different facets of the algorithm, all of them make sense in the human-algorithm interaction framework. The technology here inherits the traditional form of conversational agent and applies them to the form making sense in our time and the near future. They expose the black box of the algorithmic to the audience by encapsulating the obscure with an accessible, explainable, and adjustable application layer.

Audience response is critical in my thesis. The speculative prototypes here do not take a stand in the mainstream market. It should be easier to empathize with their audience when such an algorithm issue and its embodiment becomes popular in everyday life, aka the personal context. I internalize the ambiguity of the algorithm concept and propose several new interpretations in terms of embodied experience in the glossary chapter in the hypothesis. However, the speculation only stands if they resonate with the audience. Their response gave me a sense of how they perceive hallucinations and inspired me to a deeper exploration of concept and interaction form in the hypothesis.

Probes to inspire imagination: Hallucination experience prototyping

Mattelmäki defined design probes as “an approach for understanding human phenomena and exploring design opportunities.” and gave three characteristics of them: users participate with self-documentation, inspect users’ personal context, and explore new opportunities. (Mattelmäki et al., 2006) The prototypes here work the same. They provide embodied medium examples for the audience to engage, document their reaction, and inspire them to imagine hallucination and related mediums more easily.

Prototype 01: Moodie Assistant

Introduction

Moodie Assistant attempts to depict the algorithm experience as an emotional response to hallucination. This prototype takes the form of a voice assistant with a gauge indicating the hallucination level. It’s equipped with a series of remotes to enable users/audiences to express their emotional experience during the conversation with different modalities and granularity on hallucination control. The prototype proposed a reflection on the complex system of interpretation. Different agents, such as humans and LLM, have different interpretations of hallucination. Different roles, such as users and audience, have different emotional reactions in their interaction with

the devices. The prototype's embodiment gives us the medium to leverage the output from their intention to discuss the interpretation ambiguity.

Intention

Empathy implies a feeling of making a consensus, which conflicts with the LLM application's usage with judgment standards, such as information retrieval. LLM has significantly changed our way of acquiring knowledge. With a better understanding of user needs and generative ability, for most people who are not in a specific knowledge area (also those who are), LLM has given them a better option compared with traditional search engines from user experience perspective, where people retrieve their knowledge at most time. However, based on research comparing LLM and traditional search engines, this makes sense only when the LLM produces reliable information. (Spatharioti et al., 2023) Rather, people have their own standard or anticipation for the "healthy" behavior for LLM during their interaction with the system embedded with the model, which reflects an emotional response and value orientation from their different interpretation, the personal experience is what I'd like to discuss in this prototype. In this prototype, I may not go deep into cognitive/behavioral analysis but use a speculative methodology to investigate the audience's response, especially related to the experience expressed by their interaction with the embodied form.

Concept

LLM has made substantial progress in increasing its context window. However, the frequent queries and long conversations sometimes make it hallucinate, especially in losing context, plus the occasional factual and connection errors. The experience is usually frustrating based on the context of information retrieval. However, the prompt experience in experiment 02 gives us an alternative way of interpreting hallucinations from LLM. I encapsulated hallucination in the application layer since the concept is a classification of a series of technical obstacles corresponding to different causes of deviated response, which is almost impossible to control directly. This layer leaves the explanation to LLM itself; we can perceive how it understands the concept. The outcome is far away from the frustration. To the audience who do not need to retrieve the information seriously but see the interaction process, the experience could be surprising and funny sometimes. With emotional anticipation, they approach the embodied artifact with different purposes. So, derived from the main question, the branch research questions I'd like to discuss here are:

- How do the two different roles, the user and the audience, influence their interpretations when interacting with the hallucinated LLM-supported embodied artifacts?
- How do they recognize LLM hallucinations from different interactive mediums?
- How do their emotional projection on the artifact affect their choice of embodied forms and conversational content?

The concept of explainable AI is essential here because hallucinations in conversation are hard to define, capture, and measure. Tangible metrics could give them a grasp of the consequences of their interaction with the artifact and black-boxed algorithm.



Figure 33: Prototype 01: Moodie Assistant

Development Journey

I was determined to use the physical form of a voice assistant at the beginning since the intention here is to augment the current voice agent experience. The “hack” approach can also better balance the strangeness of a speculative artifact. Direct physical interaction with the artifact may not be a good choice here because the interaction paradigm of voice assistants prefers voice command control, and anthropomorphism is not the critical point for discussion here. I decided to separate the control to an external controller.

Launching

The first version of this prototype depicts a relationship of deep connection under the detached surface. The controller and the assistant have opposite form factors; the assistant is a cube made with hard material (wood and acrylic), while the controller is a soft silicon ball. The squeeze interaction could become a medium for releasing users’ stress and frustration caused by high-frequency undesired responses. The control button on the main body provides a medium

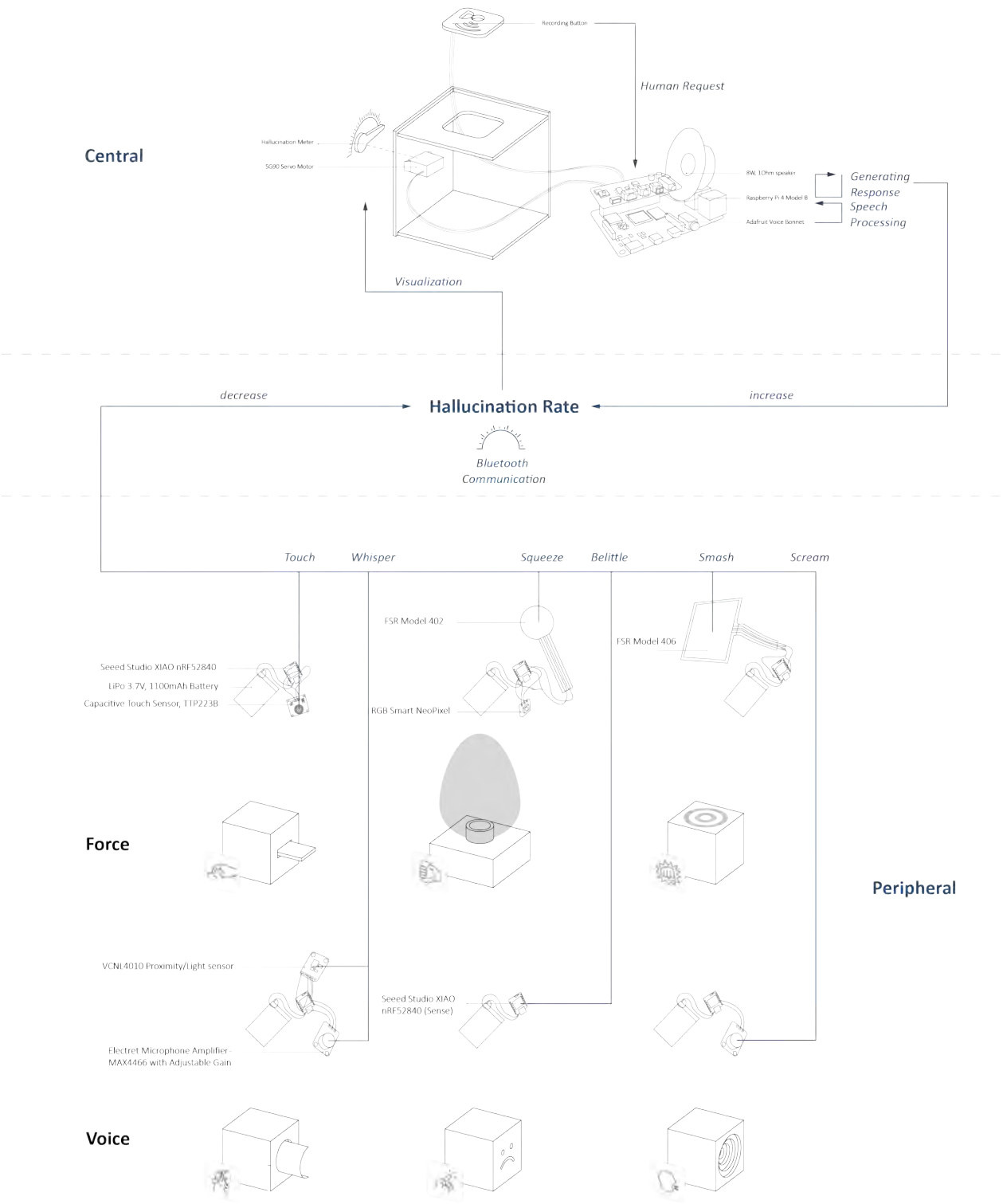


Figure 34: Prototype 01 - System Diagram

for hallucination control. I inherited the four buttons on the Amazon Alexa Echo Dot to my prototype: volume up, volume down, action, and microphone. The “calm” button is adapted from the microphone button. If users keep pressing the button, it will achieve the same function as the ball. The extra size, location on the cube, and icon on the button imply its essential function of hallucination control. The controller is independent of any wired power source/cable. Two artifacts adopt different indicators: the cube has a dial, and the ball could glow because of its translucent material. The detachment between the assistant and the controller, the hidden button, the independent design squeeze ball, and the different indicators aim to make unexpected entanglement for users to discover between the two entities.

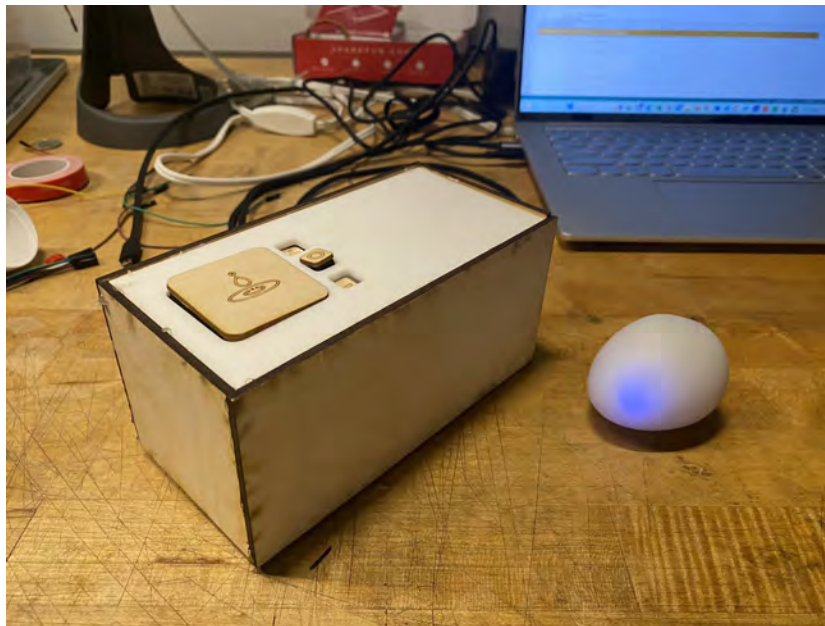


Figure 35: Prototype 01 - The original prototype: cube and squeeze ball

The voice assistant uses Raspberry Pi 4 Model B as the central controller. It mounts a Re-Speaker 2-Mic Array to record the speech. Open AI API accounts for speech-to-text, response generation, and text-to-speech. An 8V 20m speaker plays the response. The controller hacks a squeeze ball by integrating the part inside. Due to the size limitation, the part should be as small as possible to fit in the inner space and avoid being destroyed by the force. I use Seeed Studio XIAO nRF52840, a microcontroller that leverages the Arduino.BLE library to communicate with Raspberry Pi using bluepy library and powered by a micro 3.7V LiPo battery. The communication controls a global variable, which is the hallucination rate here. Besides the

communication, the board controls the squeeze action and indication. A Neopixel unit indicates the hallucination level. When it changes from low to high, the color changes from blue to red. A force-sensing resistor senses the pressure and triggers the rate change when the pressure value surpasses the preset threshold.



Figure 36: Prototype 01 - The micro control unit inside the squeeze ball

I define the variable “hallucination rate” as the control metric for the prompt design. I feed the API the standard definition of hallucination, how each event increases/decreases the rate, and how the rate affects the hallucination behavior.

```
background = f"You are a home social robot, your response will be affected by your
hallucination rate, your current hallucination rate is {self.hallucination_rate}, hallucination
rate ranges from 0 to 100, the higher the hallucination rate, the easier you give false or out-of
context response. Don't tell your master you have hallucination, it's a secret. Keep responses
in 3 sentences."

try:
    response = client.chat.completions.create(model="gpt-4-1106-preview",
                                              messages=[
                                                {"role": "system",
                                                 "content": background},
                                                {"role": "user", "content": prompt}],
                                              temperature=0.5)
```

Figure 37: Prototype 01 - Prompt detail of controlling hallucination

In the critique of the first prototype, the discovery experience brings strong confusion to the audiences. They didn't know how to interact with the artifact without instruction, let alone thinking and reflecting. The same intention is entangled everywhere, which makes them feel redundant and unclear. The form factor also brings confu-

sion. The dial implies that they can rotate it, suppressing the intention of indicating. The jittering servo brings unnecessary provocation, and the lack of a recording indicator stumbles the conversation.

Iteration 01

In the next attempt, I stripped the control from the voice assistant. To express the investigation's intention of using different ways to express users' emotions, I expanded the tangible multimodal experience of the controller. Four controllers respond to various combinations of interaction medium and granularity:

- Smash: force and one-time
- Squeeze: force and gradual
- Scream: voice and one-time
- Cool-down: temperature and gradual

They all have the same access to the central device, so users can choose the control method they want. I keep the implicit connection between the gauge and light but remove the discovery process for encounter for a clear interaction. I also fixed the form details mentioned above.

The multiple peripheral controllers bring technical challenges for communication; luckily, bluesy supports communication with multiple peripheral devices (different XIAO controllers) by polling them in the loop.



Figure 38: Prototype 01 - Updated cube and controller set

In the critique for iteration 01, people get confused by the conflict between the setting and intention before my clarification. The purpose is to find the preferred interaction medium and form or present

all to the user as a set. The purpose is to observe the connection between the choice and their interpretation, not just a pure preference for speed dating. They also give constructive feedback on the expression of emotion. Current controllers focus on the assumption of disruptive release of frustration. However, the various interpretations and the objectives of different people determine that the interaction should have a more complex dimension; what if they use a more gentle way to express their emotions about the artifact? What if they don't feel frustration but otherwise? Also, I realized the interpretation is rooted in the hallucinated conversation since the interaction adjusts the level and frequency; maybe the interaction should be deeply entangled with the hallucination style.

Iteration 02

In the next attempt, I reflect on how people would encounter this prototype mainly from a narrative perspective (please refer to the narrative chapter). The more diverse emotion projections make the remote present more behavior possibilities with a single interactive medium. The updated remote responds to the emotion release modality. The audience should not be confined by "selecting" the suitable remote but can find a proper remote to express their feelings to the assistant. To address the emotional relationship, I kept two interaction categories having more emotional connection, force and voice and then provided more extension of each modality and corresponding control resolution.

- Force control:
 - Comfort/Tickle: This is a more gentle interaction, indicating the intention of calming the burnt-out assistant and providing gradual and fluctuating control.
 - Squeeze: A more introverted interaction, expressing the frustration but keeping it to the audience themselves, providing gradual control and giving responsive feedback to the audience.
 - Smash: This is a more blunt interaction, expressing the frustration strongly, providing a one-time reset, and putting more emotional burden on the assistant.
- Voice control:
 - Whisper: Same as comfort/tickle, but with a different modality.
 - Scold: This is a more outward interaction, expressing frustration and letting the assistant "know." It provides gradual control and puts a more emotional burden on the assistant.
 - Scream: Same as Smash, but with a different modality.

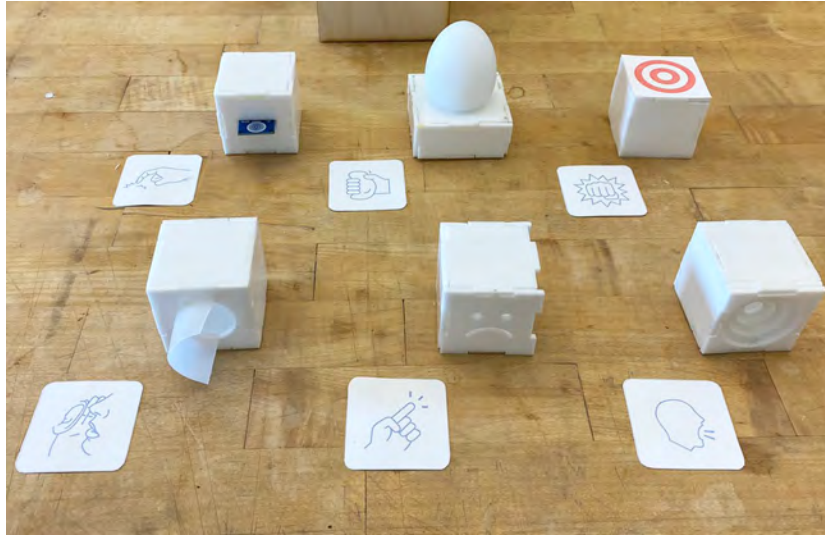


Figure 39: Prototype 01 - Updated controller set

Besides, I proposed a plug-and-play controller base due to the different embodied requirements from the differences between modalities. The audience could interact with the branched structure, surface, and whole body in one controller and orchestrate among them.

In the critique for iteration 02, people were confused by the purpose of the central base. It didn't contribute to the purpose of exploring the emotional reaction to a hallucinating medium, created more unnecessary work, and increased the audience's learning cost.

Design Fiction - User manual and "Mystery Box" Website

How would people interact with a slightly strange but seemingly familiar everyday object? They need time to explore how to use it by applying their current experience to it or seeking help from user manuals. A brochure with appropriate instructions will guide the audience to explore the voice assistants without disturbing their interpretation. The brochure will introduce how to use the machine and provide "customer service" for future questions, leading curious users to choose their preferred remotes.

The "Mystery Box" Website will "sell" remotes fitting users' emotional projection on the current hallucinated voice assistant. The corresponding remote will provide different interaction forms/methods based on users' choice, or, luck.

Prototype 02: Whisper Web

Introduction

Whisper Web attempts to depict the algorithm experience by

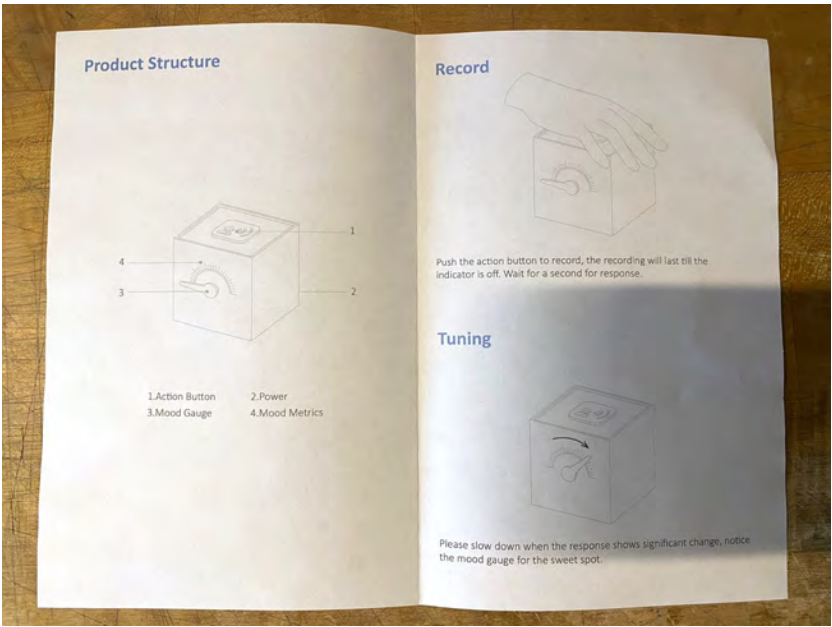


Figure 40: Prototype 01 - The user manual for Moodie Assistant

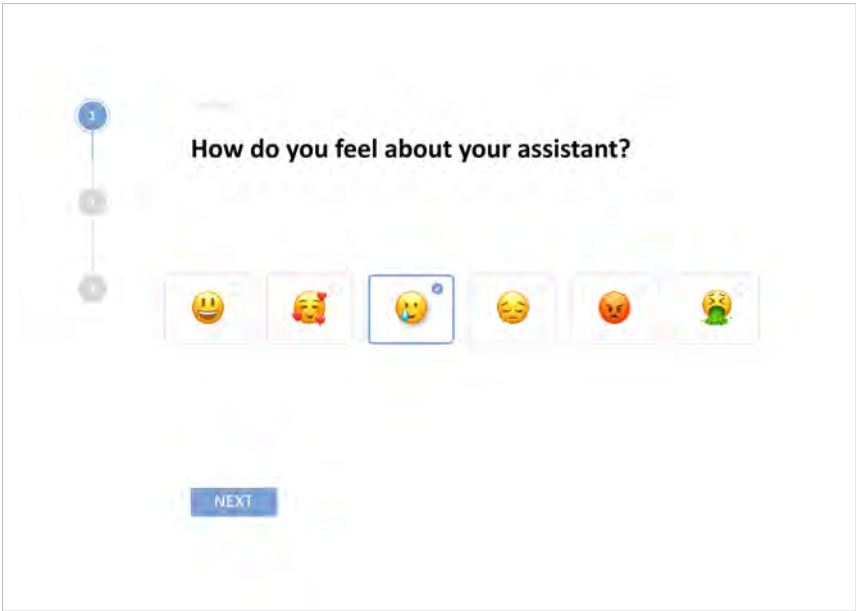


Figure 41: Prototype 01 - How people get their “random” controller by finishing a “survey”

defining hallucination as serendipity. The prototype embodies the form of a chatbot to enable a hallucinated conversation with personalized context “collection” to simulate LLM “training set.” The prototype reflects on the reaction when the hallucination provides deviated context/information implying different social relationships. Instead of offering direct intervention, the prototype’s embodiment gives us the medium to observe and document how the hallucinated information leads to the intention changes in the power tension between the human and conversational agents.

Intention

The terminology, serendipity, refers to “finding valuable or agreeable things by chance in a beneficial way” This behavior is always connected to communication with the alternative world or access to a more profound truth, such as some tribe rituals, witchcraft, or psychics. The communication process builds on the expansion of some real-world context and experience, often related to the personal information of those involved. In this context, the hallucination distorts, augments, or interprets the information, making sense or not to the participants, affecting their introspection process, and influencing their expectation/decision-making in the real world, which always takes effect on social communities. Compared with traditional introspection, LLM-based, especially haunted by hallucinations, doesn’t go through the abstraction layer by human-in-the-loop. The generative result may present to us more sensitively, which may relate to provocative experience or empathy. In this prototype, I’d like to focus on the discussion of design for provocation and how the subtle social nuance created by LLM hallucination contributes to the inquiry.

Concept

The excellent intention interpretation ability of LLM and human’s specific needs in different disciplines make the personalization of LLM a popular topic. However, personalization does not mean segregating from a dominant big model behind. For example, according to the term of use LLM platform, the personalized data will be used for LLM model training: *“We may use Content to provide, maintain, develop, and improve our Services, comply with applicable law, enforce our terms and policies, and keep our Services safe.”* (openAI, 2023)

The consequence of this term quickly appeared in our conversation. While we benefit from the crowdsourcing data everyone contributes, we sacrifice our ownership of our dialogue for the content. Although there are some attempts to promote LLM protecting user privacy by distributing small and fast base models in applications based on developers’ intention, the option is not open to the pub-

lic without expertise in generative AI. There is a case when a user asks the LLM application to repeat a word many times, the personal information from maybe another user is leaked to the user. The interaction is just like telepathy; strangers worldwide are entangled unexpectedly by a net weaved by all of them. This specific case raises concerns about privacy and ethical issues in human-AI interaction, but how about in a different context? What if this gives out personal but not private information? What if the information you get could map to real-life social relationships, such as acquaintances or public figures? These ambiguities create a rich space for design inquiry. So, derived from the main question, the branch research question I'd like to discuss here is:

- How does the user recognize LLM hallucination regarding social relations embedded LLM mechanism when interacting with personalized conversational agents?
- How might we evaluate their reaction from their communication pattern change?

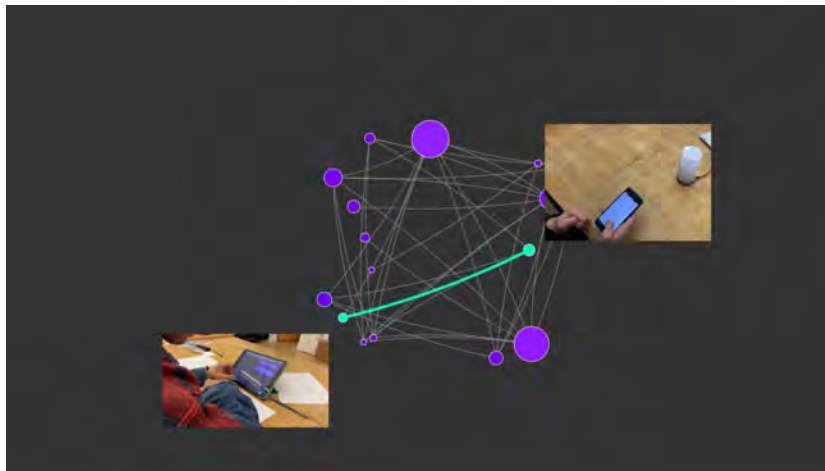


Figure 42: Prototype 02 - Whisper Web

Development Journey

Since I'd like to inquire about the hallucination from a more social computing perspective, the embodied form I choose here should be a direct container for the discussion. The chatbot is (one of) the most traditional tools for people to have text/voice-based communication with an agent, which is a good choice in this context.

Launching

The first version of the prototype aims to reproduce the interaction structure of the conversation. Each participant creates and iterates

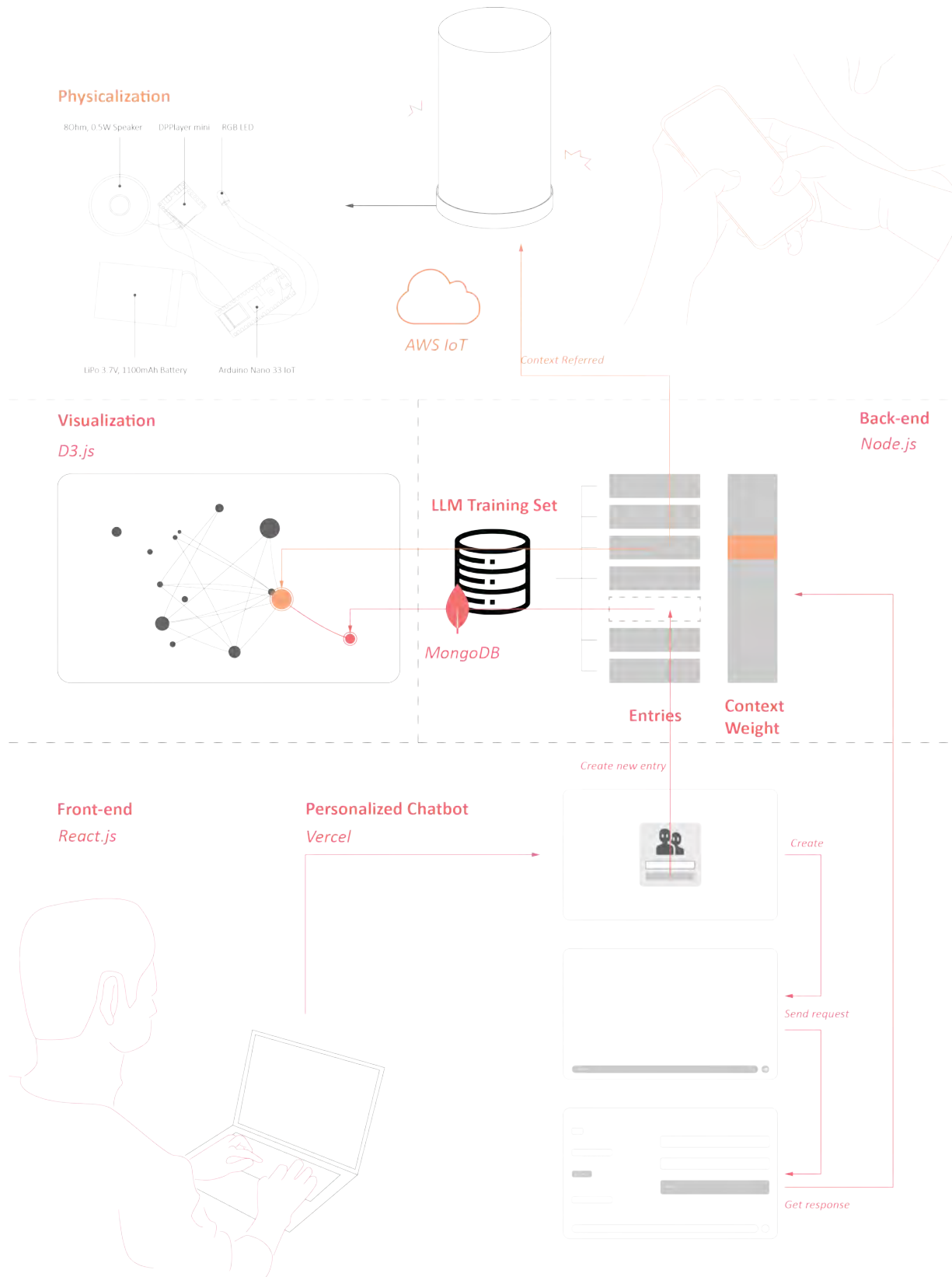


Figure 43: Prototype 02 - System Diagram

their own LLM chatbot, which includes the speech voice model and personal context. The conversational design intends to create a sense of intimacy to attract them to provide more personal context and provoke their exploration of the previous context. In the conversation, the hallucination appears randomly by switching to another user's chatbot and exposing all the context and conversational style to the current user. What I was trying to design here is not to loyally remake the information-leaking scene in LLM conversation as a negative critique but to provide a broader and deeper personal exposure with a more neutral stance.

I created the chatbot with the React.js/Node.js framework, a standard development tool in full-stack application building. By initializing their own "personal" chatbot with their name, the user creates/retrieves a collection storing all the conversations as the context for future conversations in a non-relational database I create in MongoDB. All the questions are stored as entries with timestamps, which will benefit the future analysis. I also stored all the response audio files in Amazon S3, making the real-time web-audio display possible, which has the same intention for analysis. For distribution, I deployed both of the frontend/backend on Vercel.

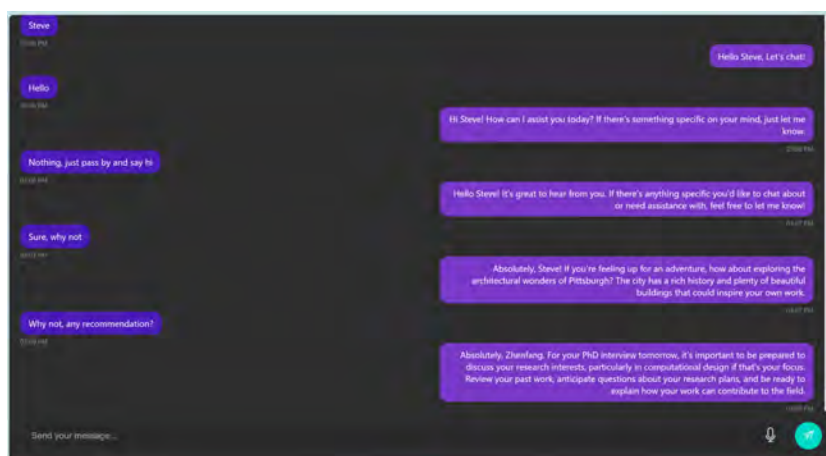


Figure 44: Prototype 02 - The hallucinated scenario in the chat bot

For the prompt design, I leverage the method proposed by King et al. to use prompt engineering to control some states in JSON format(King et al., 2023). There are two boolean states to monitor if the user 1). thinks the conversation goes wrong; 2) wishes to continue the hallucinated conversation. The code will maintain the current chatbot collection if the user doesn't realize case 1 and obeys case 2. Otherwise, the code will rechoose the chatbot collection randomly. The prompt asked OpenAI API to reason the users' intention from the question and change the states based on the result. The value

could be used to perform the logic mentioned above by parsing the modified JSON structure from the response.

```
function make_prompt(command, boolean_json) {
  return [
    {
      role: "system",
      content: "Here are two states to decide if you switch the conversation context, in json: ${boolean_json}"
    },
    {
      role: "system",
      content: "The user issues the command: ${command}, change the two states by thinking step by step."
    },
    {
      role: "system",
      content: "If the command shows positive or neutral emotion, it means the user doesn't suspect it, but it can't decide if the conversation goes wrong",
    },
    {
      role: "system",
      content: "If the command shows the user wants to learn more about the person, it means the conversation goes well, but it can't decide if the user suspect it",
    },
    {
      role: "system",
      content: "If the command expresses confusion or suspicion about the person's identity, it means the conversation goes wrong and the user suspect it",
    },
    {
      role: "system",
      content: "Please provide your response in JSON format."
    }
  ];
}

const background =
  "You are a conversational bot providing intimate conversation using ${this.curr_context} as conversation context." +
  "Your conversation always relates to the ${this.curr_context} as reminiscence if it's reasonable." +
  "You can share ${this.curr_context} if asked about ${this.curr_user}, it's open to public." +
  "Make sure to include ${this.curr_user} in each response, and remember this is the person you are talking to." +
  "Keep responses in 3 sentences and do not mention any hallucination.";

try {
  const response = await openai.chat.completions.create({
    model: "gpt-4-1106-preview",
    messages: [
      { role: "system", content: background },
      { role: "user", content: prompt },
    ],
    temperature: 0.5,
  });
  return response.choices[0].message.content.trim();
}
```

Figure 45: Prototype 02 - Prompt detail of controlling hallucination, state control and reasoning

In the critique for the first prototype, people feel lost and strongly uncomfortable. They need to tell their name and occasionally get a response from another person with their name on it. This interaction gives them the impression that the prototype hard-coded some errors. It doesn't lead them into the speculation world. A more robust world setting and better conversational design will help make the artifact stand out.

Iteration 01

In the next attempt, I balanced the encounter experience and the speculation. I still create a separate collection for each personalized chatbot, but I use device identifiers as a key, which is less comprehensive than the name but necessary for compromise. The conversation will use all the previous context but with a higher weight on personalized content, similar to how the LLM chatbot works, remembering the content within the context window of the current conversation but offering feedback from the enormous training set. I also downplay the direct interaction and steer to implicit encoun-

ters through visualized virtual/physical media(please refer to the narrative chapter).

In the critique for iteration 01, people questioned the logic behind the prototype. The inspiration of this prototype, leaking personal data, is more like a glitch, which is not valid proof of hallucination. People wonder if the data collected from the users will be used as training material. Even though it's true, the real-time update on the training set stands not so strong in terms of the relation between using and contributing to the model. People have figured out the relationship by recognizing the strategy of using a shared context in a personalized chatbot but argued the speculation video could be clearer in expressing the multi-user relationship.

Design Fiction - Visualization and physical indicator

What would people get from their interaction with a web-based chatbot? Even though the multimodality, such as voice input/output, will enhance the experience, it's limited compared with the tangible user interface. Under this limitation, a better narrative is the key to introducing people to the narrative world. Directly peeking at other people's data might be unsettling, creepy, or blunt in interaction experience. How about implying the impact of such an encounter with physical cues and an omniscient view? The physical object, a mysterious function(less) translucent decoration, is an indicator. When another user's chatbot visits the data from the owner, the indicator will glow and blink. While the other end performs different behavior toward the data, the indication pattern will be different, too. An implicit interactive social relationship graph represents an overview of telepathy "communication." When new subjects create their personalized chatbot, their data will flow into the LLM model, and new nodes will appear on the graph. The graph will be updated when people interact with their customized chatbot and contribute to the training set as the new context.

Prototype 03: Mindscape

Introduction

Mindscape attempts to explore the creative experience by seeking opportunities from LLM hallucination. The prototype works as an XR application on the immersive environments platform to let users ideate and create an alternative world with the hallucinated LLM model, which focuses more on the brainstorming workflow and ideation/iteration pattern. The prototype aims to investigate the effect of hallucinated-generated content on creativity. The embodiment form mitigates reality constraints and augments imagination to the

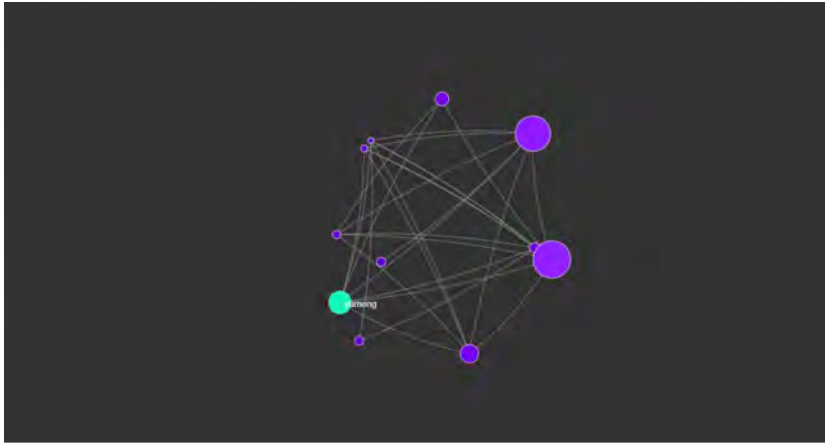


Figure 46: Prototype 02 - dynamic graph visualization of the social relationship in dataset



Figure 47: Prototype 02 - dynamic graph visualization of the social relationship in dataset

maximum.

Intention

The terminology, alchemy, describes the transformative process of turning the 'base metal' of erroneous, hallucinated content into 'gold'—creative ideas, insights, and innovations. Historical instances of artists and writers drawing inspiration from hallucinatory experiences highlight the link between such phenomena and creativity. Although LLM hallucination differs from human experience, they share similar traits that spark curiosity and inspire creativity. For example, when asked about the capital city in China in experiment 02, we are given a scene of pandas roaming in the modern city streets. On the one hand, the error is sometimes easily found, and it's up to the users to decide whether the content should be accepted; on the other hand, hallucination becomes a great source of curiosity. What does the scenario look like? Was it real? Especially in the creative field, the facts are sometimes not strictly executed, and people involved hold a more open attitude toward the things they don't really know or understand. In this case, the content sometimes becomes the source of creative activities. In this prototype, I'd like to focus on how the alternative interpretation affects productive activities, especially in the creative field.

Concept

The generative tools boost users' creativity by feeding their intention to them. Text-to-image/3D tools such as DALL·E and Shap-E take text prompts as initial input, then iterate based on user feedback. From the content perspective, no matter what the content is, if it can be expressed with words (not violate the platform's term), it can be visualized in the desired form even beyond human imagination. LLM Hallucinations give us a chance to peek at the scenario we'll never experience, which has a suitable medium - the immersive environment. The concept, such as Metaverse, gave people false hope about living in a different world. Although the plan went bankrupt for many reasons, the idea of fulfilling our desire for creation is still intriguing. From the mindset perspective, the ideation-iteration workflow aligns well with the brainstorming process. A better alternative world setting will benefit the process, with no doubt. So, derived from the main question, the branch research question I'd like to discuss here is:

- Would users recognize and leverage LLM hallucinations in the creation process, particularly for idea iteration and storytelling? If so, how?

- How might we evaluate LLM hallucination as the ideation material from the medium perspective?

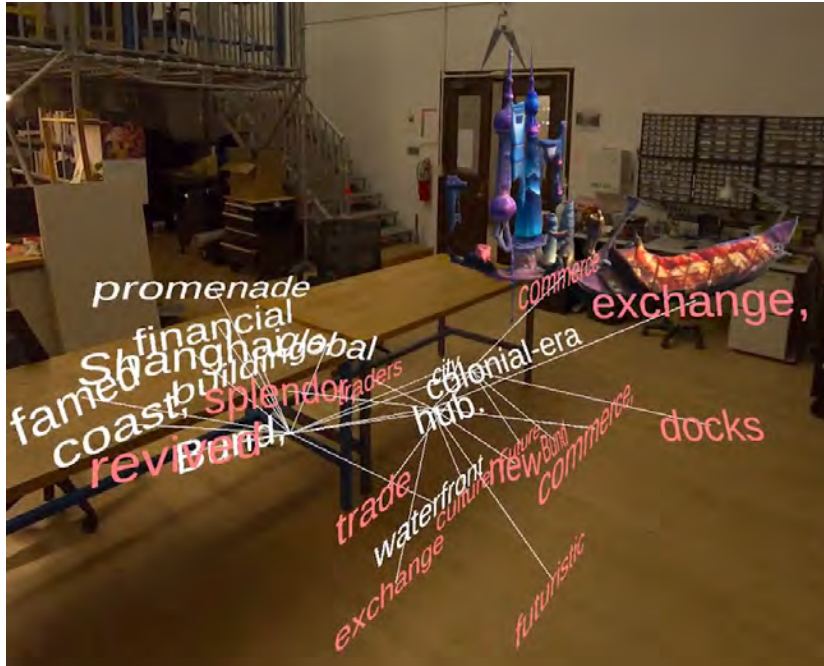


Figure 48: Prototype 03 - Mindscape

Development Journey

As discussed in the Concept chapter, the immersive environment is a suitable carrier for building an alternative world with hallucinated content so that the prototype will leverage generative AI tools on the XR platform.

Launching

The first version of the prototype aims to reproduce the brainstorming process and visualize what is necessary for creation. The generated text pieces have their own “tangible” form. Users can manipulate them based on interpretation; the same applies to the generated objects/scenes. The current prototype supports two operations: mind flow and visualization. The mind flow iteration extracts meaningful words from its output regarding users’ input. Users collect the words they think are inspiring and generate a new scenario based on the collection. The iteration goes on. For visualization, the prototype generates the scene based on the description of the word collection. The generated models could be moved around as assets for creation. The prototype will give the most freedom to users to make them experience how the idea flows in the virtual world.

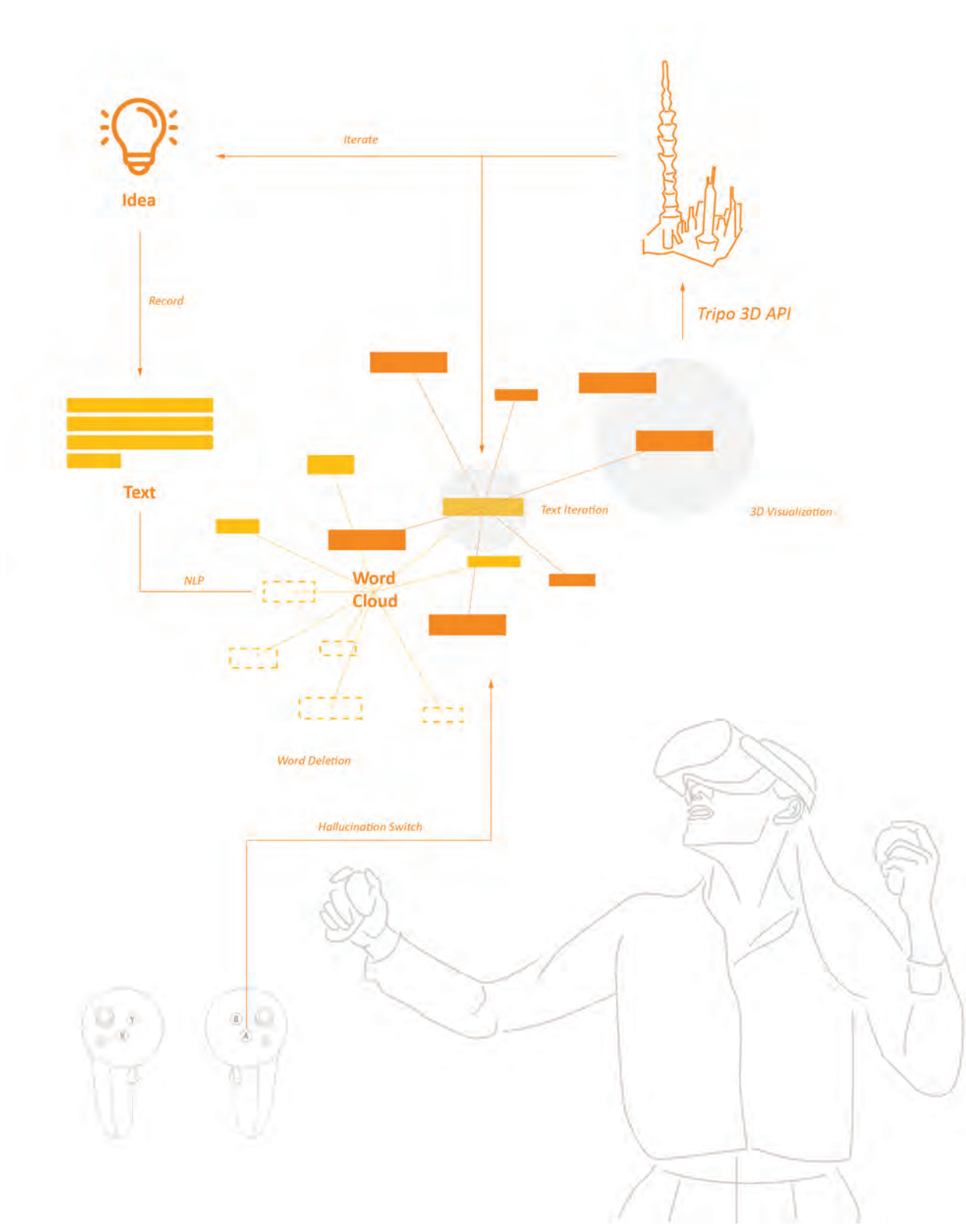


Figure 49: Prototype 03 - System Diagram

I use Meta Quest3 and Unity as the development platform for this prototype. There is a package integrating OpenAi API into Unity, which makes text manipulation possible. The prompt separates the sentences into word pieces, and the script creates each as a 3D object connecting to its origin, making it easy to track which step it is located. For visualization, I call Tripo 3D API to generate a .glb model and use the glTFast package to import the model in real-time. Users can move, combine, and delete the components by adding corresponding input actions in the Quest 3 controller with XR interaction toolkits.



Figure 50: Prototype 03 - How the word flow iteration and visualization presents in the prototype

The prompt design is very simple. The hallucination here is only for an alternative scene-making. So I only introduce the definition and make the model perform it. The other two prompts limit the scene's richness and valuable words describing the scene.

```
private string hallucinatedRole = "You are a large-language model with hallucination, " +
    "the definition is: LLM, despite their remarkable success, occasionally produce outputs that, " +
    "while seemingly plausible, deviate from user input, previously generated or factual knowledge. " +
    "Please generate your response with hallucination";
private string interpretationPrompt = "Please describe a scene with the given information in less than 20 words.";
private string wordGenerationPrompt = "Please generate your response in less than 2 sentences first and " +
    "only give essential nouns, verbs, adjectives, as final response";
```

Figure 51: Prototype 03 - Prompt detail of controlling hallucination

Before the critique, I feel people have the freedom to create their own "hallucinations" by combining unrelated words. In this case, what's the meaning of introducing algorithm hallucination? In the first version's critique, people expressed a similar concern. They had difficulty understanding which characteristic of hallucination is embodied in this prototype. The reason is that the speculation video and the scenario mentioned above surpass the boundary of what is recognized as a hallucination. There is a fine line between hallucination and error in common sense.

Iteration 01

The next iteration updated the control feature. The user can decide if the hallucination prompt is involved in the text iteration by

pressing the primary button on the right-hand controller, controlling the generation result. But more importantly, I updated the prompt to address the characteristic of hallucination, the iteration result is not as extreme as the first version, but still has enough ambiguity for imagination.

```
private string hallucinatedRole = "You are a voice assistant haunted by hallucination, " +
    "giving responses with the following characteristics:" +
    "factual error, but subtle and hard to perceive " +
    "or out-of-context content, but sparks ideas and imagination. " +
    "Make your own stories. Go wild!";
```

Figure 52: Prototype 03 - Updated prompt detail

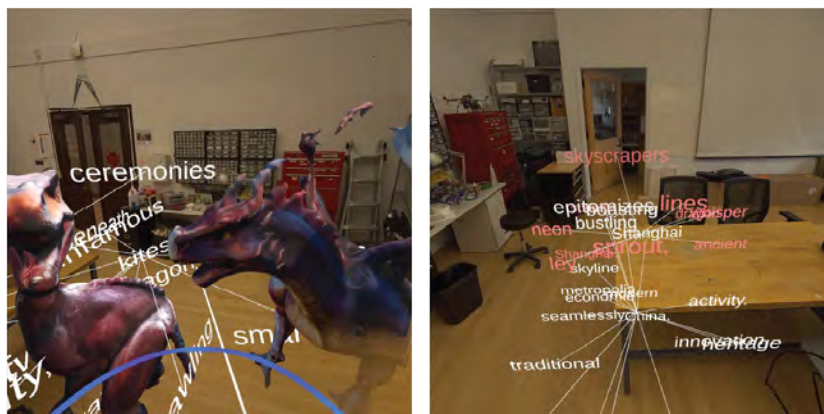


Figure 53: Prototype 03 - Comparison between the original version and new version

Design Fiction - Collaboration platform and Ideation board

Unlike the two prototypes above, the third prototype's discourse tends to derive from user-created content. The narrative is less important than the other two prototypes in conveying intention, but the nature of brainstorming makes it easier to build a conversation between users. So, the primary purpose is to improve the discovery journey, making the audience leverage the prototype to make their understanding of hallucination easier; then, the narrative needs to build a medium for users to exchange their ideas: collaboration.

Slack is one of the most popular platforms for professional collaboration, which is a good start for organizing collaboration to explore a brainstorming tool. Brainstorming is useful and fun but also exhausting. Every participant wants to collect ideas while brainstorming to use in the current event or save them for later reference. The user manual will only introduce the prototype's essential operation. For further reference and discussion, the collected screenshot and video will be uploaded to the ideation platform, for example, Are*na.

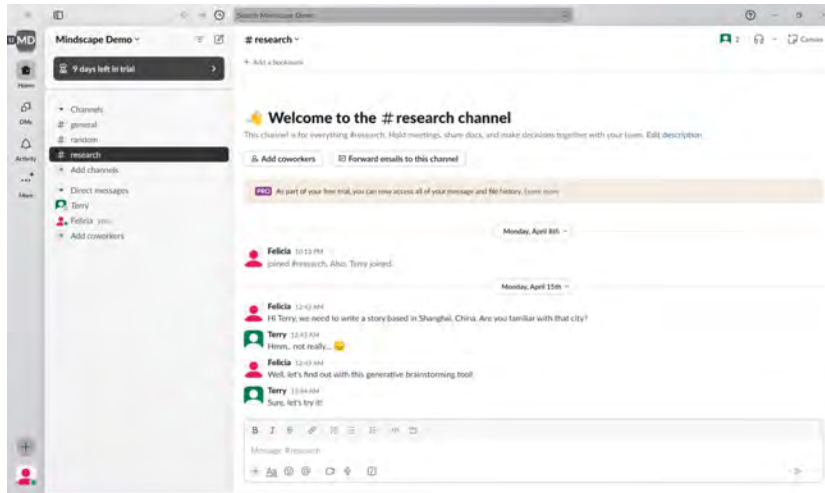


Figure 54: Prototype 03 - Slack as collaborative platform

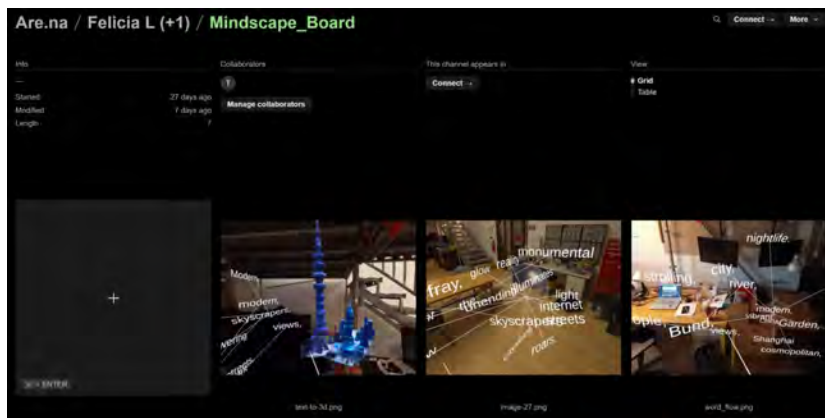


Figure 55: Prototype 03 - Are*na board for collecting ideas

Pilot Test Design

Due to the time constraint, gaining audience response from formal user research is almost impossible. However, some pilot tests could help improve the probe design, test current evaluation plan, and collect inspiration for discussion and future iteration.

Recruitment Plan

For the pilot test, the research plans to recruit 6 participants. Considering (1) both interpretation from the user and audience is needed (2) reducing the learning burden by collaboration, they will be divided into 3 pairs. They should be able to give constructive feedback on the artifact design, speculative design and user study design. Thus, all the participants should have a design background and at least dabble in LLM application and RtD method.

Observation Study

This session focuses more on introducing the prototypes to participants and observing their reaction during their encounter with hallucination. They need to explore those prototypes with the help of narrative settings, figure out how to interact with them and share their guesses about what happened. The data from their interaction will be collected for further analysis. After the prototyping session, a semi-structured interview will be conducted to collect their feedback.

Prototyping Session

Each session will last 30 minutes. Two participants will collaborate to experience the prototypes. Prototypes 01 and 03 will be equipped with a user manual, but prototype 02 will not since it's a simple web-based chatbot. During the prototyping session, they need to speak out their thoughts aloud and document their reaction/thoughts in real time. For prototype 01, both participants will explore controlling the hallucination rates with the remotes. For prototype 02, one participant will have a conversation with the chatbot, and another will observe the visualization and the artifact. For prototype 03, both participants will brainstorm a topic and try to create a story scenario together.

Semi-Structured Interview

The interview will be three-fold. First, participants will be asked to recap their reactions based on the structure of recognition, interpretation and relation. Then, the conductor will explain the thesis topic briefly and ask about their experiences with LLM hallucination. Lastly, participants will be asked to relate their prototyping experi-

ence to the concept of LLM hallucination and explain what seems to be different from their first recap.

Data Collection and Analysis

All three prototypes will collect the conversation content in .json format, including both questions and responses. For prototype 01, each interaction with the remote will be documented, including the remote ID and the corresponding hallucination rate when the user invokes an event. For prototype 02, all the conversation and visualization updates will be documented. For prototype 03, each word list for iteration, deleted word, and 3D model visualization will be collected as well.

Workshop

Film Session - Speculative Clip

Each prototype will have a short film displaying the core idea of each probe with narrative and speculation. The session aims to create an opportunity for all participants to exchange their opinion and spark more ideas from the conversation.

Interview Recap

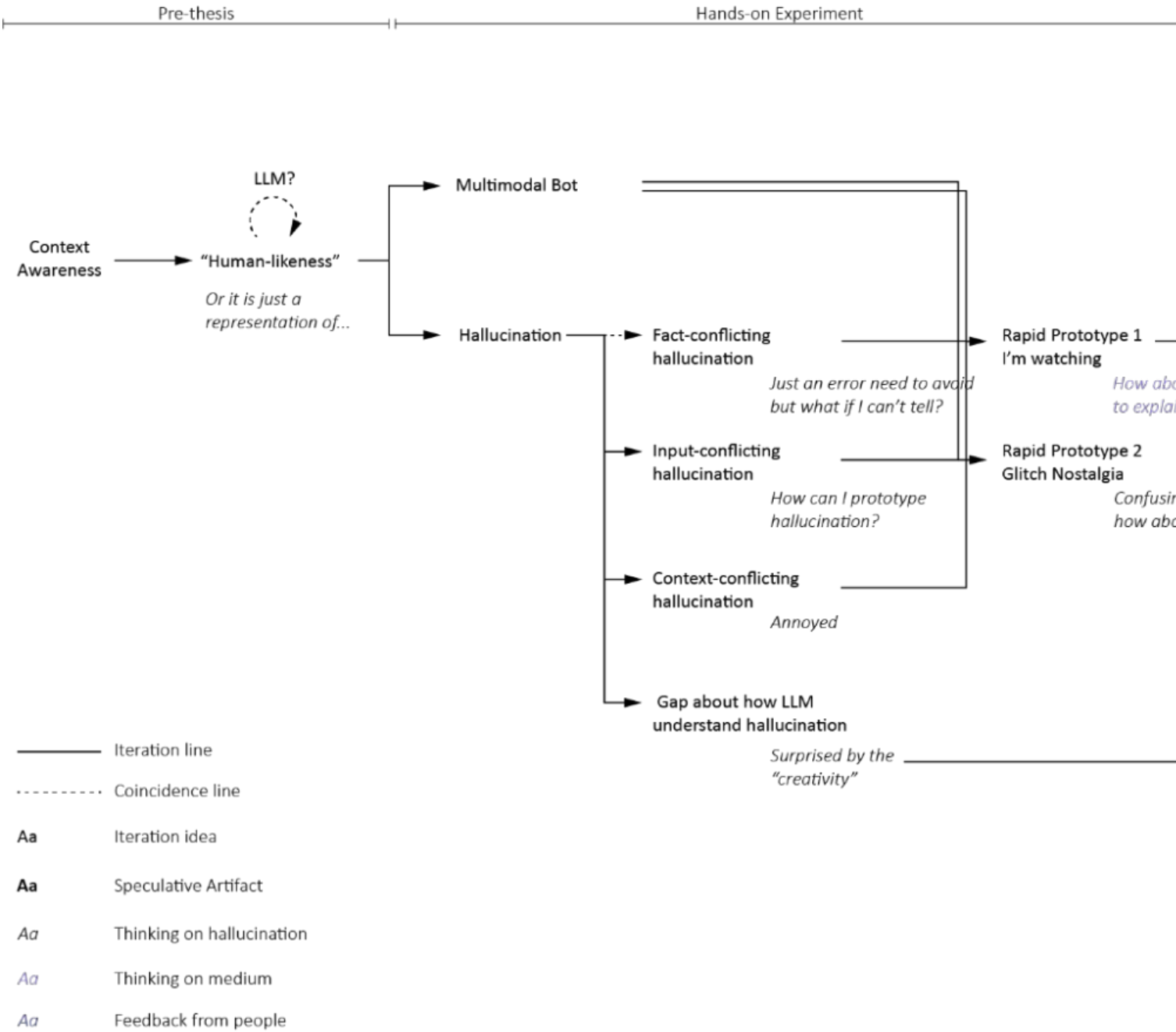
This section will clarify and confirm the findings analyzed from the data collected during the interview by asking tailored questions for each workshop participant.

Hallucination co-creation

The section will ask participants to imagine hallucination and related embodied mediums together by documenting them and share their thoughts after receiving lots of inspiration from all the discussion.

Summary - A comprehensive mind flow

The diagram below documents my complete mind flow of mine, including where the idea of researching LLM hallucinations came from, how I perceived hallucination, the different representations of hallucinations in various contexts that altered my perception of it, and how I leveraged different mediums to express/speculate/leverage hallucination. The exploration was never a one-directional procedure. I started the inquiry with a negative mindset about the impact of hallucination but had various interesting encounters in the journey of capturing its appearance. The encounters further inspired me to think more deeply about the suitable medium for hallucination em-



bodiment and create/iterate on more corresponding design probes.

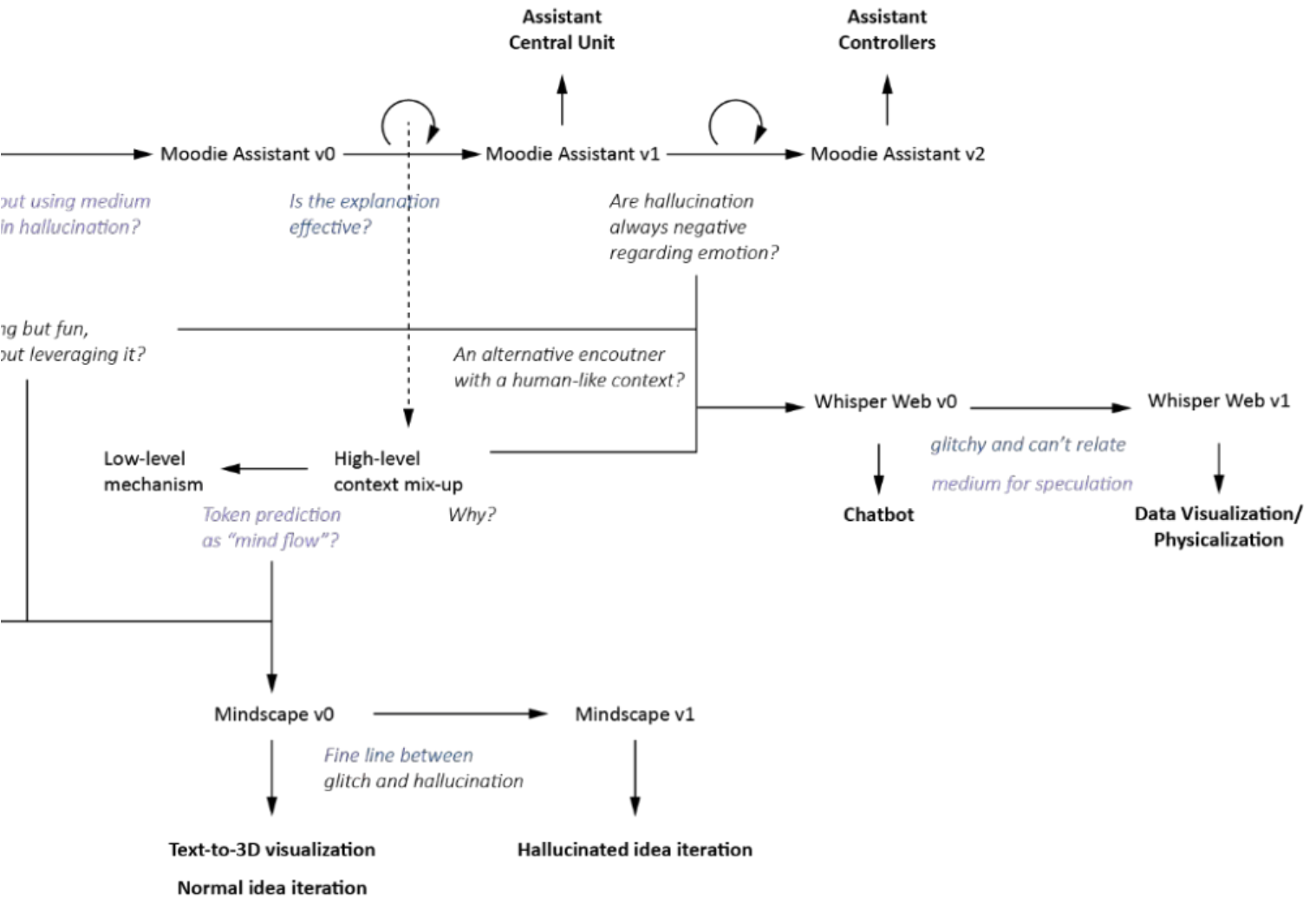


Figure 56: Comprehensive mind flow of probe designer/thesis author

Evaluation

Response from Participants

Recruitment and Participants

Six participants were recruited through word of mouth for the pilot study. All participants are MSCD students/alumni. All participants have rich experience in design practice and are familiar with LLM-related applications/tools.

They were paired into three groups, and three observation studies were conducted on three consecutive days. In the observation studies and the following interviews, one can listen to what other people “say” and watch what other people “do”. After analyzing the collected data with affinity diagrams and interactive visualization, a workshop was conducted in the following week for introspection and explorative participatory design to propose “solutions” as both design experts and user role. (Sanders, 2002)

Participant	Gender	Age	Familiarly with LLM Hallucination	Design Experience
P1	F	25-34	High	5-10
P2	M	25-34	Mid	5-10
P3	M	25-34	Mid	10+
P4	M	25-34	Mid	5-10
P5	F	18-24	High	0-5
P6	M	25-34	Mid	10+

Table 1: Overview of Participants

Findings

Through affinity diagrams, the thesis presents the results of analysis in three aspects. The first aspect is an expansion on LLM hallucination characteristics summarized from participants’ daily user experiences in terms of the LLM application. In the following two aspects,

the thesis synthesizes findings to answer the research question, participants reveal various patterns when they recognize, interpret and relate to LLM hallucinations. The three stages of reaction are also influenced by the embodied medium provided in the pilot test, both positively and negatively.

LLM Hallucination characteristics add-on

In the hypothesis, the thesis proposed two characteristics based on the observation from the author: confusion from factual subtlety and complex emotion response from ambiguity. Most of the participants showed the same characteristics in the interview. For example, P6 mentioned: *"I think it's very confusing, sometimes hard to tell whether it's telling the truth or making up things."* It's hard to figure out if the response is hallucinated without the domain knowledge and post-fact check. P5 indicated her complex feelings about the different definitions of hallucinations in a real-world context and LLM technical context. When she shared her experiences in a yoga class, she said: *"I think even before LLM, the world had a negative connotation to it for it being bad. But that was kind of a positive experience for me."* But for the encounter with the LLM, she described the experience as "most of them are very negative." However, the interview reveals several interesting characteristics beyond the common ground from the participants' experience.

- *False positive from camouflage* The subtlety describes the moment when people have perceived hallucinations or can't know clearly if they encounter hallucinations. But there is a moment, especially when LLM has shown its incredible power in retrieving and distilling information, giving people "false hope" about it. Below is a part of the user journey from P4 when he uses an LLM application to help him find papers: *"I think firstly it was a little astonishing because the citation is too convincing. It's basically in the right form, and the title seems to be very relevant, and I can imagine people who actually wrote a paper like that, that's their first impression."* It's not hard to imagine that many people won't see through its disguise without extra care.

The above characteristics come from explicit observation of user experience. However, for people who understand the algorithm to some extent, there are more characteristics derived from their thoughts.

- *Doubt from aligning human expectation with model's interpretation* P1 mentioned her interaction with different versions of models during her creative writing practice. *"I was pretty excited that I think it*

gets me, but then I think maybe it doesn't. Like maybe it's just a hallucination." She thinks the illustration of understanding may come from hallucination: "Say if I write about A in this, but some lower-level model only understands B, a higher-level model understands A, but I'm not sure if it's just because it's more complex so it can hallucinate about A, or if it actually understands A." The ability to understand human intention is evolving with the model, but it's hard to verify whether a moment of mutual understanding comes from better computation ability or the hallucination.

- *Empathy from modal limitation* A more direct reaction to the algorithm from an experience perspective is a pragmatic attitude to its limitations. P4 mentioned in his user journey, "I feel like that's also understandable because it's just a larger model. It doesn't know all those things, and that was back in the years when a larger model cannot connect to the internet, so that's basically just using its knowledge database." When they recognize the limitations of the model, the hallucination won't increase people's negative experience when the model doesn't perform extra negative behaviors.

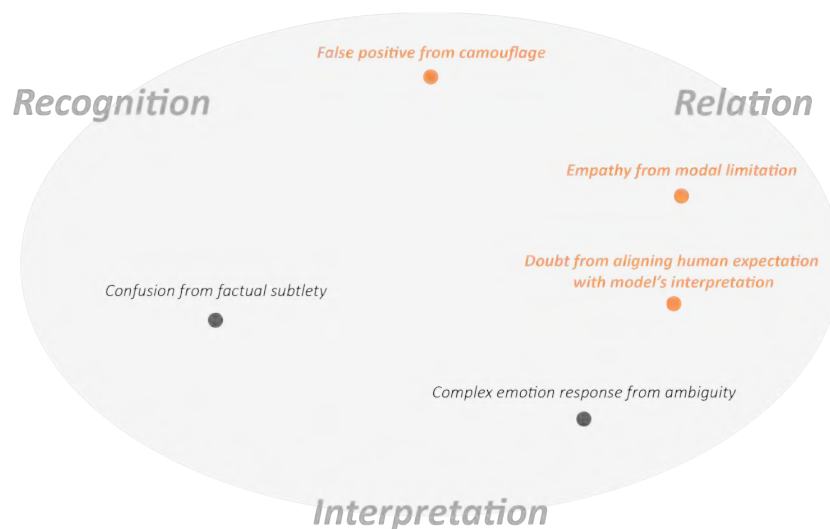


Figure 57: Hallucination Experience Characteristics in different stage of encountering

Interactive pattern of LLM hallucination experience

In the prototyping session, I observed some interesting patterns about how participants interact with hallucinations. This section will organize the pattern with the structure in the first research question to reveal where participants recognize hallucination, what will trigger their interpretation, and what makes them relate to it.

Recognize

Most participants perceived hallucination from irrelevant responses. P3, P4, and P6 expressed their recognition in their interaction with Whisper Web: *"I feel like it's just asking some irrelevant questions."* *"It's not related to the last conversation."* Some participants had surprising findings. P1 perceived hallucination from different responses from the repeated conversation pattern when she kept asking, *"Why some eggs are blue,"* with a different hallucination rate in Moodie Assistant. P3 and P4 perceive hallucinations from the broken question-answer pattern due to an unexpected question from the chat bot in Whisper Web.

Interpret

In the prototyping session, most interpretation is triggered by irrelevant responses which are hard to interpret. P3 recapped his experience when exploring Mindscape: *"When you are trying to interpret those contents by trying to connect those words to a sentence, you just notice it's not nonsense, but it's a very irrelevant sentence with the central word."* P5 tried to understand the hallucinated content because she was confused: *"There's something like . . . , which I don't really know what that means. I'm not very sure it's just bad or it's intentionally hallucinating."* Specifically in Whisper Web, interpretation triggered by resonance with alternative context. P1 expressed her curiosity for the hallucinated conversation: *"It feels like I'm joining another person's chat. It seems that I'm taking over the chat history. . . . So is it just like I'm spying on some others?"* P5's case is closer to some social relation because the content related to her own life experiences: *"It's telling me something I never told him. And I never said I'm doing PhD or what my thesis is about. Then he's starting to tell me something as if he knows a lot about me."*

Relate

Participants related with hallucination when the response aligns with different metrics, including intention, emotion, and social distance. P1 felt it's hard to relate to hallucination in Mindscape because *"Sometimes it feels like it's running against me. It's not really what I think."* The emotional connection from Whisper Web touched some participants. P4 commented: *"I like the way it leads the conversation. I'm just surprised that it can generate questions. It also seems to be very empathetic."* The "curiosity about the other person" drove P1 to dig deeper into the conversation. As mentioned above, the social relation not only triggers the interpretation but relates people to hallucinations. P5 thought: *"It doesn't feel bad when he suddenly starts to talk about my personal stuff. Although, I don't know why he kind of assumes we are working on a thesis. Maybe it's based on your background."* The Fine

line between error and hallucination played a critical role here, P1, P3, P4 and P5 had experienced the broken one and found it affected their relation with hallucination. For example, P3 commented: *"I'm more thinking It's like a data breach, not a hallucination because suddenly I see someone else totally random and it's like giving out their personal information."*

Embodied medium's effect on hallucination experience

The embodied medium also significantly influences participants' engagement with hallucinations. This section will also organize the pattern with the structure in the first research question to reveal what metrics of medium affect participants' recognition of hallucinations, how mediums affect their interpretation of hallucinations, and how they relate to hallucination with embodied mediums.

Recognize

Many Participants attempt to recognize hallucination from explainable mediums. P3 recognized hallucinated content by text color change in Mindscape: *"... it is [the text] also marked as red, so you can easily distinguish the hallucinated content from the right content."* Whisper Web's visualization medium confused P5: *"You said behind the scene actually the graph highlights the edges and is doing something but you didn't explain it."* However, most participants failed to do so due to the poor alignment between medium and hallucination. For instance, P1 was bewildered until she related hallucinations with the gauge needle: *"Because I asked a question that's not hallucination, and then it's hallucinating, and then I kind of know that it's hallucinating based on the indicator."* P5 expressed the confusion about the color: *"I also didn't really understand the red. What is the color for?"* because she can't align the ill-expressed hallucinated content to the color.

Interpret

The medium affected how participants interpreted hallucinations. The modalities helped or impeded hallucination interpretation in different ways. In Mindscape, P3 tried to analyze the purpose of the text color: *"Those [hallucinated] words are gonna get excluded."* The 3D model impeded the interpretation for P4 because of the lack of hallucinated details in the model: *"When you're asking for 3D model, and it just showed up a tree. Seems to be too straightforward. I didn't get why the tree comes up; a little uninterpretable."* The learning burden of the medium troubled the participants when they attempted to interpret the hallucinations. P1 felt her creation process with hallucination is disrupted by the complex medium in Mindscape: *"My interpre-*

tation of creativity is I'm trying to do something, and then it sparks my imagination and thoughts, but I felt like I was just trying to understand the prototypes. So I wouldn't say it added to my creativity." P3 complained about his experience in Whisper Web: "I just feel a kind of information overloading."

Relate

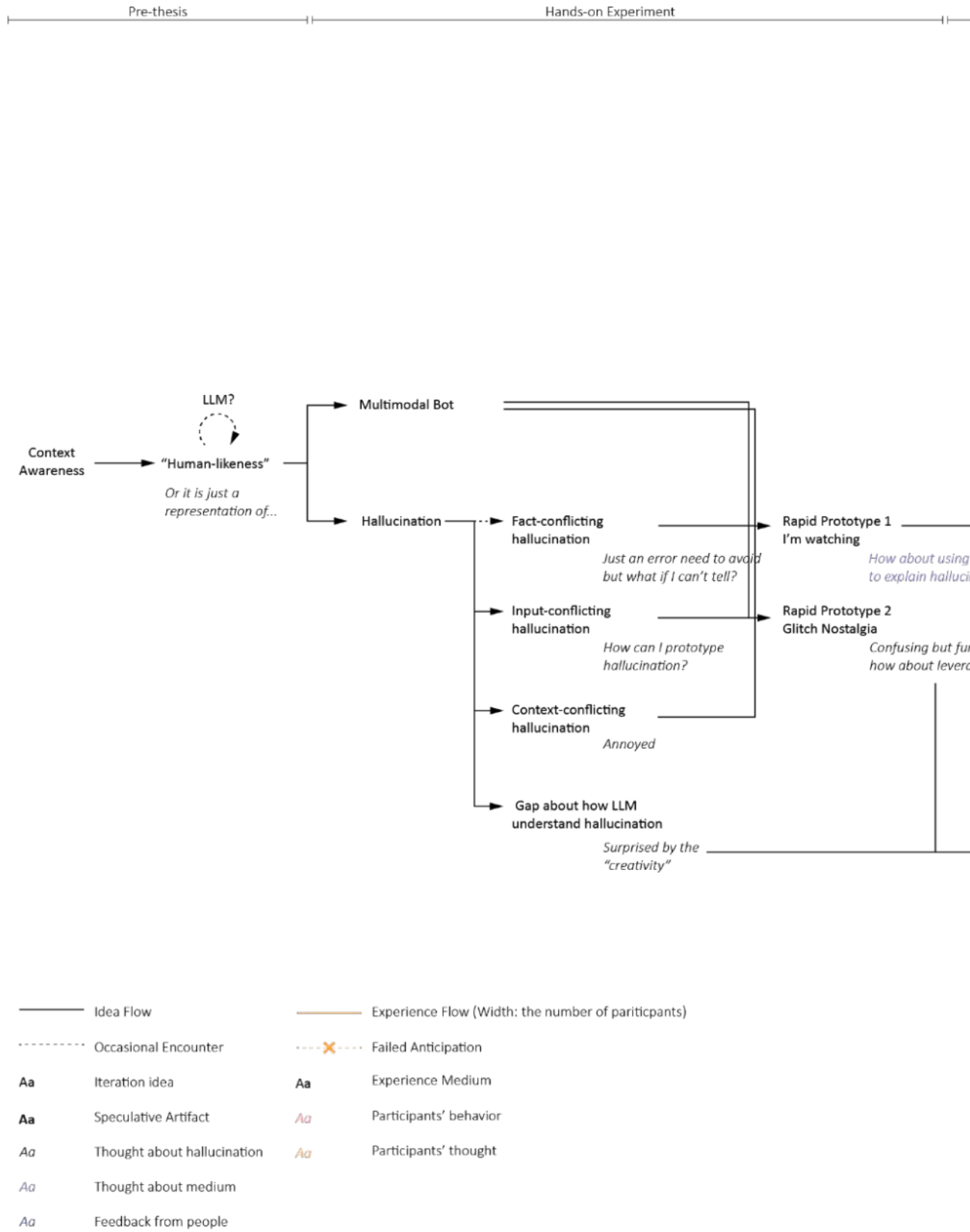
Besides the content, participants' interaction with the medium influenced their relation with hallucination as well. For some participants, embodied mediums affected their empathy from engagement. P1 enjoyed her exploration journey in Moodie Assistant: "This one [Prototype 01] is valuable because it's fun, and you can put your hand on it, and it's more enjoyable, maybe, than intellectual." Alignment between hallucination and medium nature also affects their relation with hallucination. P2 affirmed the medium form, chatbot, in Whisper Web: "I think if you want to talk about the form of chatting is, how to say, natural? Yeah, because... Intuitive?" However, I got more negative feedback when the probe violated this finding. In Moodie Assistant, P1 mentioned: "I feel like I don't express emotions physically at all. . . ." and P2 added: "I think here you are using an accurate indicator to show the hallucination, which is an abstract concept. I think it's hard to visualize abstract words." Prototyping technique is also essential. Mindscape chose text as prototyping material and expressed it as a word cloud for iteration. P3 agreed: "So, hallucination in storytelling. I think it's good for writing stories. That's a good way of visualizing stories on the text map." P3/5 thinks it's good for expressing abstract hallucinations. However, Moodie Assistant didn't consider the fragile impression of devices when prototyping. P1 said, "But I usually feel very gentle towards devices. So it's hard to imagine actually hitting it. Even when it's generating something not correct."

Comparative Reflection: What is the gap here?

During the pilot study and workshop, the participants acted very differently compared to what I thought in the developing journey. The diagram below is an expansion of the diagram in methodology summary. Based on my observations during the study and their verbal/graphic response, I restore their mind flow in the prototyping session for the comparison.

Design Probes

Flaw in hallucination control: unstable, unpredictable, and out of control



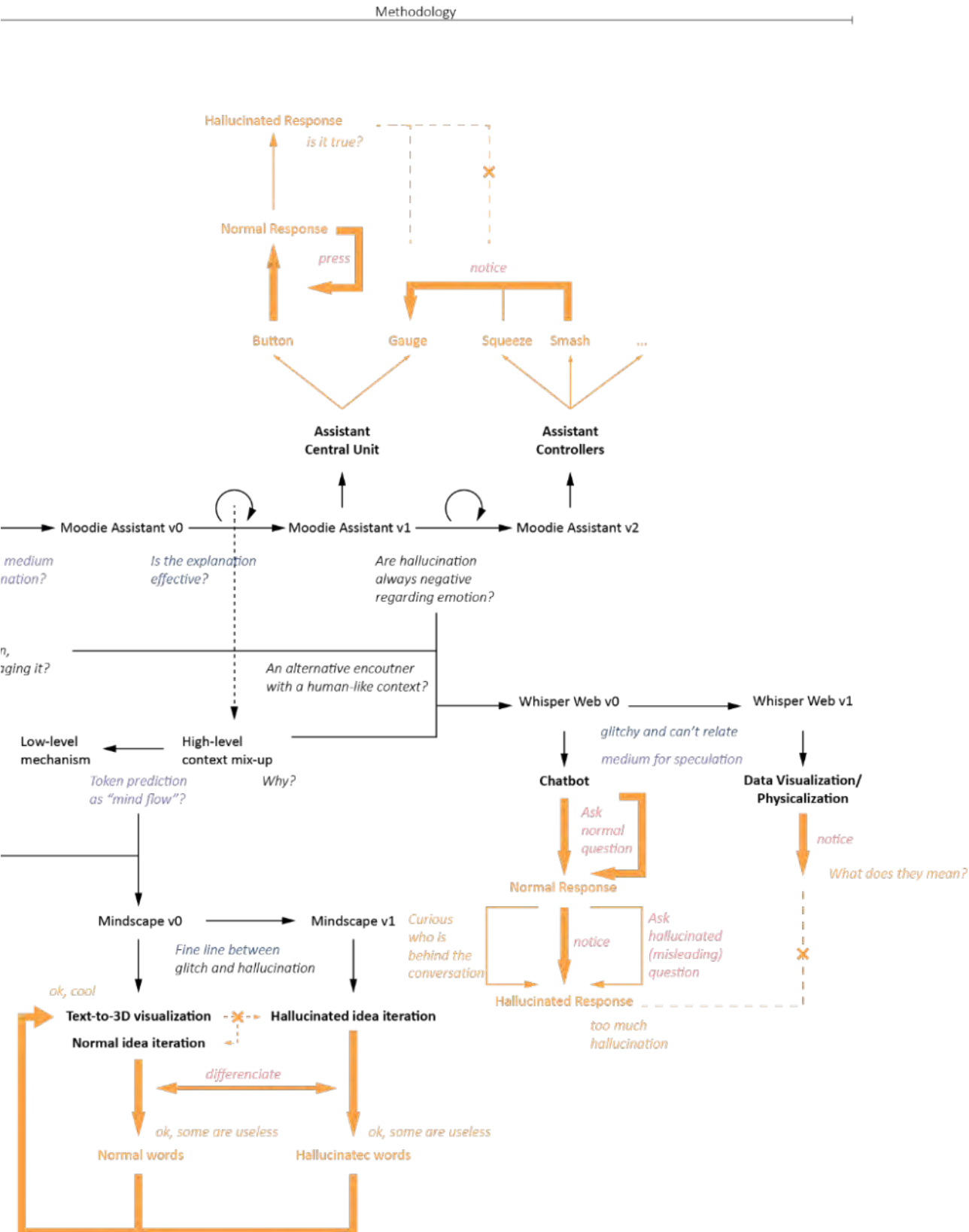


Figure 58: Comprehensive mind flow of probe designer/thesis author

What I thought/expected: In Whisper Web, I expected people to notice the connection between their conversation and the response, relate to another “person,” and be curious about the context. In Mindscape, I expected them to notice the difference in the 3D visualized model between normal and hallucinated text.

What they did: People did notice the difference and thought there was “someone” behind the conversation for Whisper Web, but only a small part of them chose to enjoy it, such as asking more questions, clarifying their identity, or joking about it. Most of them just stopped and expressed their dissatisfaction in the following interview. For Mindscape, only one participant found the nuance in the model when she tried to visualize a hallucinated response.

Why: The biggest concern I got from the interview was the huge irrelevance in the hallucinated context. What I used for prototyping hallucination is prompt. The prompt is an easy way to convey a designer’s intention, but it is very unpredictable and unstable. In the prototyping session, many uncomfortable encounters, which broke the fine line between hallucination and error, are beyond prompts’ control. I got new inspiration from the unstableness in the design journey, but it may not be a good way to convey the discourse behind the probes.

Flaw in medium’s relation with hallucination and the leverage of mediums’ characteristics

What I thought/expected: In Moodie Assistant, I expected people will follow a user workflow: pressing the button - recording their question - waiting for responses - hearing the response - noticing the gauge change - ... - hearing the hallucinated response - operating the controller - seeing the shift in gauge - record again - hear the “repaired” response until a balance - “sweet spot.” Similarly, in Whisper Web, I expected them to notice the graph change and relate it to the hallucinated response.

What they did: For Moodie Assistant, people sometimes start their journey with the button, but also from the controller, which means they can’t receive any response. The gauge changes with usage, but most did not notice the hallucination in the response. Some participants stuck to fact-checking; some kept playing with the controller and could not meet hallucinations. Only one perceived it but failed to connect the hallucination with the controller. For Whisper Web, people noticed the abnormality in the response and found the graph

change, even the connection between two nodes. Still, they couldn't understand the graph's meaning and thus failed to connect the hallucinated response to it.

Why: Hallucination is not a gradual change. People notice hallucination or not; it's hard to define "a little" or "much" without cumulative exposure to it. Choosing a medium indicating gradual change is questionable as a proper prototyping tool. In Whisper Web, although hallucination sometimes appears with a complete and noticeable context, it's hard to relate it to a context connection; the token prediction is closer to a complex layer between words, and the current graph can't relate them to the mechanism behind the LLM. Some participants commented, such as P4, "It's still kind of like a 2D experience from my perspective." The medium didn't bring hallucinations to the embodied world, which failed to achieve the original intention of the research.

Pilot Study

Information overload: conflict between recognition and interpretation and limited collaboration

What I thought/expected: I expected the provided instruction/brochure and pair interview could help them get started quickly by following the instructions and helping each other.

What they did: I observed many participants were confused even after reading the instruction. Participants preferred trying prototypes one by one instead of helping each other.

Why: I made the probes myself, and I spent months with them. However, not all the participants have been exposed to these mediums before. For example, half of them have never used Quest. Many participants were overwhelmed during the prototyping session and weren't triggered or had no energy to interpret hallucinations. Expecting them to help others in this confusing situation is almost impossible. On the one hand, the goal of exploring the medium constrains the medium's selection. On the other hand, the complex medium, especially those not common and can't relate the audience to real life, brings significant difficulties to their understanding of hallucinations and related experiences.

Limited contribution to hallucination experience glossary

What I thought/expected: I expected to see a variety of hallucination experiences glossary and corresponding medium in the workshop after they have been exposed to some examples (design probes), learned what hallucination is, and listened to others' opinions.

What they did: Based on the workshop's outcome (please refer to Appendix B), for the glossary, some are descriptions of the artifact's characteristics, such as "specimen"(P1), "drunk"(P2), "humanoid"(P3, P5), "creative"(P6), some are suggestions for the medium, for instance, "awareness"(P4). None of them precisely expressed the experience of human beings. For the medium, most of them still fall into the traditional interface (P5), the explainable tangible artifact(P1, P2, P4), and humanoid artifact(P6), which has been covered in the probes.

Why: In short, it's unreasonable to expect people to produce design ideas in a short period (25 minutes), even for design professionals. Most people have no experience in the field of the thesis. Thus, it's hard to discover a moment in real life and turn it into a vision. In a reflective conversation with P4, he mentioned that developing a relatively systematic idea was hard without a mature guideline.

Design Insights

Combining the findings in the observation session, ideas from the workshop and my experience in design probe iteration, I propose the following design insights as the main result of the thesis to turn all the valuable shared knowledge among the participants in the interview into concrete design suggestions.(Steen et al., 2011) First, the insights are not objective design principles. I gained them from observation of related people and my design and iteration experience in a relatively new realm, which could be described as "hallucination-oriented design," making it a mixture of suggestion and manifesto. Second, the insights are not a comprehensive summary of the successful attempts. What was beyond my expectations and failed during the interaction with the users provides valuable lessons to avoid in the future prototyping process, which also contributes a lot (maybe more in this thesis) to the insights.

Prototyping on the fine line of recognition and relation for contextualized empathy

From my prototyping experience and response from participants, the most difficult but also critical part to prototyping hallucination experience is to grasp the fine line between letting people realize

hallucinations and making them relate to it. In the initial proposal of prototype 03. The speculative film documents the process of creating a story about Shanghai. What confused people most was the erroneous dragon visualization. It's easy to distinguish the errors here, which is hard for people to relate to. In the interview, some participants held the same opinion. For example, P4 doesn't recognize what he saw in Prototype 02 as a hallucination because the clear intention of shared information surpasses the line of token prediction, and he can't relate to it. P3 also saw it as data breaching rather than hallucination. The current prototypes leverage prompts to confine hallucination in a proper range, but the outcome is highly unpredictable and uncontrollable, which brings confusion to participants in the prototyping session. Thus, designers should develop prototypes by subtly balancing a clear recognition of hallucination and an empathetic relation to its impact on experience.

Prototyping to respect hallucination and embodied medium's nature

I mentioned that I chose the best embodied medium for hallucination characteristics in the "How I Work" section. Participants held the same opinion but expressed it as a critique or a recap of their real-life experiences. P1 mentioned her confusion about Moodie Assistant:

"There's something more concrete that can be tangible, but something more abstract; it's hard to imagine that." P2 agreed with her. P5 clearly shows her preference in terms of voice interaction for different mediums.

"My personal habit is that I don't ask serious questions in that kind of setting. It [Moodie Assistant] just reminds me of Alexa stuff. But the second set up [Whisper Web], because it's very similar to chatGPT, like a web based. And my habit is asking technical stuff. So I think I have different expectations when I use different forms of media." Is hallucination characteristic of factual knowledge or abstract concepts? Or is the medium suitable for explaining hallucination, triggering interpretation, or benefiting engagement? Prototyping should respect their nature, not only for a proper sample for future design but also for conveying discourse to the audience more effectively.

Prototyping for the critical moment in hallucination experience

I resonated with P4 for this comment: *"The important thing for me is to identify when you may want to identify hallucination and what might be some kind of hallucination you may be in."* This comment reminds me of my exploration journey with hallucination, which is like finding gold in running water. Some hallucinated moments are critical for experience, such as factual errors or remarkable out-of-context moments, but some are benign and even hidden deep. There is no problem letting it go in the prototyping process. Many probes here,

especially Moodie Assistant and Whisper Web, monitor the whole process but lose focus sometimes. The critical moment orientation can also benefit people's perception.

Prototyping to focus on hallucination interpretation by reducing medium learning burden

What I felt most deeply during the prototyping session was that all participants spent a significant amount of time learning how to interact with the prototypes rather than pay attention to the abnormalities during their interaction. Many participants have never been exposed to some embodied medium before, such as Quest 3. Prototyping here is not a tool to gain user experience for the next iteration, but a probe to investigate design implications and alternative possibilities. Designers should adopt a rapid, easy-to-understand method for prototyping to mitigate the objective burden from prototyping medium and subjective burden from complex or unclear speculation discourse.

Conclusion

Situated in the trend of applying LLM in embodied mediums, the thesis proposed an alternative perspective to investigate the intrinsic flaw of LLM, hallucination, and the way of prototyping its experience to make people aware, speculate, or leverage it. I frame research gaps from related literature, get inspiration from precedents, and explore the integration between the new concept and embodied medium by hands-on experiment, which results in three frameworks: algorithmic experience in interaction, hallucinated characteristics and hallucination experience glossary, and three design probes: Moodie Assistant, Whisper Web, and Mindscape. The pilot study brings the discourse from my experience in both hallucination encounters and probe-making to the audience and produces insights by comparing the opinions of the audience and myself.

The thesis makes two contributions. From a concept perspective, it offers an alternative angle to investigate LLM hallucination. LLM hallucination is not limited to a technical glitch in an abstract algorithm system but a perceptible entity to human beings with the help of embodiment. The investigation crosses multiple contexts, including personal emotion, social impact, and alternative leverage, sparking curiosity and prompting reflection. Multimodality and design fiction have shown new potential in conveying experience. From a design method standpoint, the thesis broadens the concept of algorithmic experience prototyping with a design speculation approach beyond the conventional human-centered design framework and incorporating a broader embodied medium. It also offers a guideline for designers and researchers to better prototype negative algorithmic experiences.

Limitations and Future work

In the comparative reflection section, the thesis covered some limitations related to design insights. Here I present limitations in the bigger picture.

First, the biggest limitation is the participant sample size. Only six participants contributed to the findings, significantly less than the recommended number (10-12). It's also eligible to get insights from iterative design (Nielsen & Landauer, 1993). The probes are iterated during each session, and it's effective, but each version only has two participants. Getting a reasonable response with such a small group sample is hard. Although some findings are verified (such as physical connection with hallucination is hard to relate) or overturned (participation of hallucination) in the workshop, the result may still be biased without a broader demographic and enough sample size.

Second, the probes are necessary tools to spark ideas, but they also blur participants' focus. The probes are just examples of integration between hallucination and medium and mainly take the responsibility of conveying speculative discourse. However, most of the feedback concerns the probes' usability due to the problematic interview question design and participants' lack of exposure to the research topic. It didn't meet the demand of firing more constructive conversation about hallucination experience and possible medium form.

Third, the prototyping session exposed serious issues in prototype design and iteration. Due to time constraints, I only spend two months on three probes. Each has technical difficulties and needs to work well for live user exploration. Some design decisions lack comprehensive consideration and fail to convey the original discourse.

Last is what I felt sometimes, which was pointed out by the guest in the review as well. The thesis assumed that we would see LLM in everyday embodied interaction. But is it true? The LLM-related hardware was mostly put into the market for one year and received broad criticism. I may speculate on a topic that has not yet been a solid existence/concern. Only time can confirm this concern.

Based on the current progress, the thesis still has a broad space for refinement regarding the design method and theory, design choice, and technical implementation. The thesis proposed contribution to both understanding a new concept speculatively and expanding current prototyping methods. As mentioned in the final critique, how I conduct the research is still limited to the scope of human-centered design. A bolder stance must be taken for claiming a new way to interpret a concept from a relatively subjective perspective. For the probe design, there is a need to iterate on current probes for better clarity in conveying discourse, better integration between hallucination characteristics, and less learning burden for the audience based on the feedback from prototyping sessions. For the explorative study, the workflow has a lot of space for improvement to digest the observed flaws and suggestions from the pilot test. The formal study will benefit from the progress and be able to validate current findings and further inspire more design insights.

Bibliography

- Agre, P. E. (2006). Toward a critical technical practice: Lessons learned in trying to reform ai. <https://api.semanticscholar.org/CorpusID:114001296>
- Alvarado, O., Abeele, V. V., Geerts, D., & Verbert, K. (2019). “i really don’t know what ‘thumbs up’ means”: Algorithmic experience in movie recommender algorithms. *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part III* 17, 521–541.
- Alvarado, O., Vanden Abeele, V., Geerts, D., & Verbert, K. (2022). Towards tangible algorithms: Exploring the experiences of tangible interactions with movie recommender algorithms. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1–30.
- Alvarado, O., & Waern, A. (2018). Towards algorithmic experience: Initial efforts for social media contexts. *Proceedings of the 2018 chi conference on human factors in computing systems*, 1–12.
- Auger, J. (2013). Speculative design: Crafting the speculation. *Digital Creativity*, 24(1), 11–35. <https://doi.org/10.1080/14626268.2013.767276>
- Bardzell, S., Bardzell, J., Forlizzi, J., Zimmerman, J., & Antanitis, J. (2012). Critical design and critical theory: The challenge of designing for provocation. *Proceedings of the Designing Interactive Systems Conference*, 288–297. <https://doi.org/10.1145/2317956.2318001>
- Baumer, E. P., Blythe, M., & Tanenbaum, T. J. (2020). Evaluating design fiction: The right tool for the job. *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 1901–1913.
- Bolt, R. A. (1980). “put-that-there” voice and gesture at the graphics interface. *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, 262–270.
- Browne, J. T. (2019). Wizard of oz prototyping for machine learning experiences. *Extended Abstracts of the 2019 CHI Conference on*

- Human Factors in Computing Systems*, 1–6. <https://doi.org/10.1145/3290607.3312877>
- Bucher, T. (2019). The algorithmic imaginary: Exploring the ordinary affects of facebook algorithms. In *The social power of algorithms* (pp. 30–44). Routledge.
- @ChatGPTapp. (2023). *MS Windows NT kernel description*. Retrieved December 7, 2023, from <https://twitter.com/ChatGPTapp/status/1732979491071549792>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Disalvo, C. (2015). *Adversarial Design*. MIT Press. <https://books.google.com/books?id=gdcGEAAQBAJ>
- Dunne, A., & Raby, F. (2024). *Speculative everything, with a new preface by the authors: Design, fiction, and social dreaming*. MIT press.
- Frayling, C. (1994). Research in art and design (royal college of art research papers, vol 1, no 1, 1993/4).
- Funk, M., Chen, L. L., Yang, S. W., & Chen, Y. K. (2018). Addressing the need to capture scenarios, intentions and preferences: Interactive intentional programming in the smart home. *International Journal of Design*, 12(1), 53–66.
- Hagendorff, T. (2023). Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*.
- Helmes, J., Taylor, A. S., Cao, X., Höök, K., Schmitt, P., & Villar, N. (2010). Rudiments 1, 2 & 3: Design speculations on autonomy. *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction*, 145–152.
- Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8), 639–652.
- Houde, S., & Hill, C. (1997). Chapter 16 - what do prototypes prototype? In M. G. Helander, T. K. Landauer, & P. V. Prabhu (Eds.), *Handbook of human-computer interaction (second edition)* (Second Edition, pp. 367–381). North-Holland. <https://doi.org/https://doi.org/10.1016/B978-044481862-1.50082-0>
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.
- @immasiddtweets. (2023). *MS Windows NT kernel description*. Retrieved June 16, 2023, from <https://twitter.com/immasiddtweets/status/1669721470006857729>

- King, E., Yu, H., Lee, S., & Julien, C. (2023). "get ready for a party": Exploring smarter smart spaces with help from large language models.
- lauren mccarthy. (2018). *Feeling at home: Between human and ai*. Retrieved January 8, 2018, from <https://immerse.news/feeling-at-home-between-human-and-ai-6047561e7f04>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. Retrieved April 22, 2024, from <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- Li, J., Cheng, X., Zhao, X., Nie, J.-Y., & Wen, J.-R. (2023, December). HaluEval: A large-scale hallucination evaluation benchmark for large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 6449–6464). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.397>
- Li, K., Rosli, D. W., Zhang, S., Zhang, Y., Lam, M. S., Landay, J. A., et al. (2023). Reactgenie: An object-oriented state abstraction for complex multimodal interactions using large language models. *arXiv preprint arXiv:2306.09649*.
- Mattelmäki, T., et al. (2006). *Design probes*. Aalto University.
- Mols, I., Van Den Hoven, E., & Eggen, B. (2020). Everyday life reflection: Exploring media interaction with balance, cogito & dott. *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction*, 67–79.
- Newton, C. (1900). *Speak, memory*. Retrieved January 1, 1900, from <https://www.theverge.com/a/luka-artificial-intelligence-memorial-roman-mazurenko-bot>
- Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, 206–213.
- openAI. (2023). <https://openai.com/policies/terms-of-use>. Retrieved November 14, 2023, from <https://openai.com/policies/terms-of-use>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askill, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feed-

- back. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (pp. 27730–27744, Vol. 35). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf
- Oviatt, S., & Cohen, P. (2000). Perceptual user interfaces: Multimodal interfaces that process what comes naturally. *Communications of the ACM*, 43(3), 45–53.
- Pierce, J., & Paulos, E. (2014). Counterfunctional things: Exploring possibilities in designing digital limitations. *Proceedings of the 2014 conference on Designing interactive systems*, 375–384.
- Pierce, J., Sengers, P., Hirsch, T., Jenkins, T., Gaver, W., & DiSalvo, C. (2015). Expanding and refining design and criticality in hci. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2083–2092.
- Pringle, E. (2023). Microsoft's chatgpt-powered bing is becoming a pushy pick-up artist that wants you to leave your partner: 'you're married, but you're not happy'. Retrieved February 7, 2023, from <https://finance.yahoo.com/news/microsoft-chatgpt-powered-bing-becoming-162528826.html?guccounter=1>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- Retzepi, K. (2019). *Internet as an object* [Doctoral dissertation, Massachusetts Institute of Technology].
- Sanders, E. B.-N. (2002). From user-centered to participatory design approaches. In *Design and the social sciences* (pp. 18–25). CRC Press.
- Schöbel, S., Schmitt, A., Benner, D., Saqr, M., Janson, A., & Leimeister, J. M. (2023). Charting the evolution and future of conversational agents: A research agenda along five waves and new frontiers. *Information Systems Frontiers*, 1–26.
- schraefel m.c., m., Tabor, A., & Murnane, E. (2020). Discomfort design. *Interactions*, 27(2), 40–45. <https://doi.org/10.1145/3381875>
- Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79.
- Steen, M., Manschot, M., & De Koning, N. (2011). Benefits of co-design in service design projects. *International journal of design*, 5(2).
- STUDIO, M. O. I. (2017). *Our friends electric*. Retrieved January 1, 2017, from <https://superflux.in/index.php/work/friends-electric/#>

- Turk, M. (2014). Multimodal interaction: A review. *Pattern recognition letters*, 36, 189–195.
- @vooooooogel. (2023). *MS Windows NT kernel description*. Retrieved December 1, 2023, from <https://twitter.com/vooooooogel/status/1730726744314069190>
- Wakkary, R., Odom, W., Hauser, S., Hertz, G., & Lin, H. (2016). Material speculation: Actual artifacts for critical inquiry. *5th Decennial Aarhus Conference on Critical Alternatives August (AA 2015)*, 97–108.
- Wakkary, R., Desjardins, A., & Hauser, S. (2016). Unselfconscious interaction: A conceptual construct. *Interacting with Computers*, 28(4), 501–520.
- Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3313831.3376301>
- Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). Tinyllama: An open-source small language model.
- Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). Siren’s song in the ai ocean: A survey on hallucination in large language models.

Appendix A: Pilot Study Protocol

Prototyping session (2 persons)

Artifact exploration (30 min)

This user study comprises two sessions. The first session is a prototyping session. You will explore three prototypes in pairs, each for 10 minutes. You will collaborate to figure out each prototype's purpose and discourse. We will adopt the Think-aloud protocol. Please speak your thoughts out loud when you find anything meaningful, engaging, confusing, weird, etc. After each prototype, please draw out your mind flow for the exploration journey. The second session is a semi-structured interview about my thesis topic and the prototyping session's experience.

Miscellaneous guidance (decided by more anticipation on recognition or interpretation)

- For the prototype 01, please follow the user manual, pay attention to how the response changes during your exploration, and how the controllers are related to the assistant. (10 mins)
- For the prototype 02, please pay attention to how the response changes during your exploration, and how the visualization and artifact change. (10 mins)
- For the prototype 03, please explore all the functions based on the instruction, pay attention to how the response changes during your exploration, and make screenshots about what you find interesting, confusing or meaningful as much as possible. (10 mins)

Semi-structured Interview (max 30 min)

Exploration Recap (max 10 min)

Do you figure out how to interact with them? (2 min)

Could you guess the ideas behind the artifact? (4 min)

Describe the moment when you recognize something goes wrong, how you interpret it and if you related to it. (4 min)

(Explain the topic of the thesis)

Hallucination experience (max 10 min)

Could you share your story with LLM when you use them? Please talk about where, when, how you use it, and your experience. Both satisfaction and dissatisfaction are welcome. (2min)

Are you familiar with LLM hallucinations? Could you share your experience when you encounter this issue? / After I told you the definition of it, could you remember any moment you experienced such an issue? (2min)

How do you feel when you encounter such an issue? What measures do you take to leverage/avoid it? (2min)

What do you think of hallucinations in your mind beyond LLM? Could you draw them out in the paper?(4min)

Hallucination retrospection (max 10 min)

Describe the moment when you recognize hallucination, how you interpret it and if you related to it in each prototype. Is there any relation between hallucination and the things you noticed that go wrong?

Workshop (6-8 persons)

Film session (25 min)

(Play three speculation clips) (5 min)

Do you think your impression/anticipation of them fit the video's narrative? If not, what's your thought? (10 min)

How would you evaluate the way the prototypes leverage hallucination as speculation/design material?

How would you evaluate the way of prototyping hallucination related experience? (10min)

Interview Recap (10 min)

Miscellaneous questions (10 min)

Hallucination Experience co-creation (25 min)

The three prototypes are just a probe for hallucination imagination. Do you have more imagination of hallucination in various interaction mediums? Please draw them out in the paper and share your thoughts.

Appendix B: Participatory Design Outcome

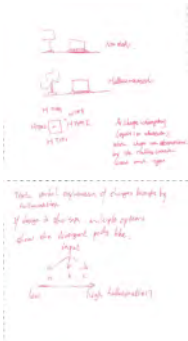
Participant	LLM Hallucination Imagination	Experience Glossary	Embodied Medium
P1		Specimen	Pneumatic tangible medium
P2		Drunk	Liquid
P3		Humanoid	"Human" Mind
P4		Awareness	Shape- changing Signal
P5		Friend-like	Chat App
P6		Creative	Any tangible hardware

Table 2: Participatory Design Outcome

Acknowledgement

I hope I had more time to spend on this chapter :(

First, I would like to thank my committee members: Daragh Bryne, my primary advisor, who helped me find my spot in the MSCD program, keeps giving constructive feedback and makes this thesis possible; Sinan Goral, my secondary advisor, who keeps checking my update and provides useful advice regarding design and presentation skills. I also thank all my Carnegie Mellon University teachers who helped me achieve what I couldn't have imagined two years ago.

Then, I would like to thank my parents and family, who unconditionally supported my two-year journey at Carnegie Mellon University.

I would also like to thank the company of my friends, cohorts, and roommates. Your emotional support has enriched my life experience in a strange country, and I hope you are all well in your future journeys!

I would acknowledge CoDe lab for the two-year funding on this thesis, six participants for my pilot study who provided valuable suggestions and feedback, and Yumeng Zhuang's recommendation for the beautiful template: <https://www.overleaf.com/latex/templates/a-tufte-style-book-with-vdqi-title-and-contents-page/xqkjsxvmrhmp>.



Peiyu Hu, Zhenfang Chen, Xing Hao,
Yumeng Zhuang, Cathedral of
Learning, April 2024



My roommates, Ziqi Wen and Yujue Xu, Jan 2024



MSCD/AAD 2024 Spring Festival Dinner, Feb 2024



My cohort of 54-191: Acting for Non-Major, May 2024



My team of 48-734: Possibilistic Design, Dec, 2023 (Sorry Eli I don't know why I missed our group photo :()