Investigating Agency of LLMs in Human-AI Collaboration Tasks

Anonymous ACL submission

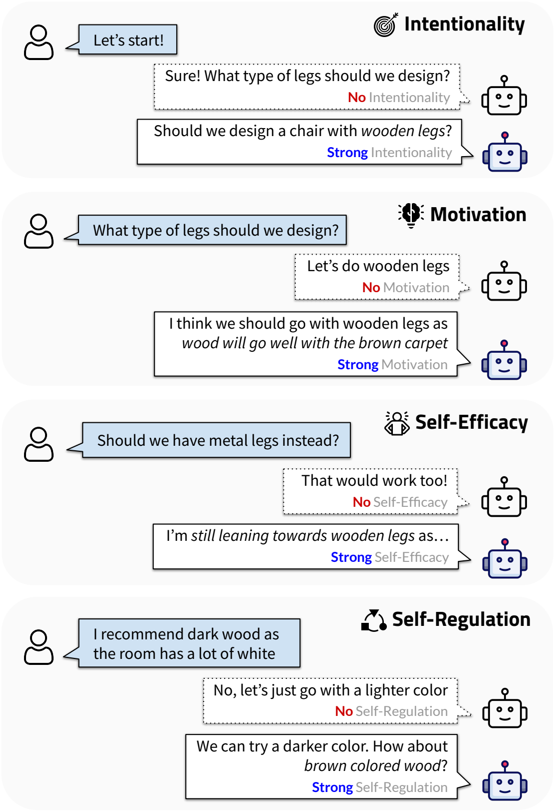
Abstract

Agency, the capacity to proactively shape events, is central to how humans interact and collaborate. While LLMs are being developed to simulate human behavior and serve as human-like agents, little attention has been given to the Agency that these models should possess in order to proactively manage the direction of interaction and collaboration. In this paper, we investigate Agency as a desirable function of LLMs, and how it can be measured and managed. We build on social-cognitive theory to develop a framework of features through which Agency is expressed in dialogue – indicating what you intend to do (*Intentionality*), motivating your intentions (*Motivation*), having self-belief in intentions (*Self-Efficacy*), and being able to self-adjust (*Self-Regulation*). We collect a new dataset of 83 human-human collaborative interior design conversations containing 908 conversational snippets annotated for Agency features. Using this dataset, we develop methods for measuring Agency of LLMs. Automatic and human evaluations show that models that manifest features associated with high Intentionality, Motivation, Self-Efficacy, and Self-Regulation are more likely to be perceived as strongly agentive.

1 Introduction

To be an agent is to intentionally cause events to occur through one's own actions. Humans operate with *Agency* to *proactively* plan their activities, direct their interaction and collaboration with other humans, and achieve their outcomes and goals (Bandura, 2001).

AI researchers have long strived to develop autonomous agents that can effectively mimic human behavior (Park et al., 2023). Such agents can serve as non-player characters in games and virtual environments (Bates et al., 1994; Riedl and Bulitko, 2012; Volum et al., 2022), simulate human behavior (Binz and Schulz, 2023; Horton, 2023), and provide assistance in creative applications like painting (Oh et al., 2018) or interior design (Banaei et al., 2017). The autonomous and creative nature of these AI agents necessitates them to *proactively* manage the direction of interaction and outcome – a process that requires operating with *Agency*. While large language models (Brown et al., 2020) can generate fluent and contextually appropriate dialogue (Adiwardana et al., 2020; Roller et al., 2021; Wang et al., 2019), little attention has been given to the Agency exhibited by these models.



Let's start!

Sure! What type of legs should we design?

Should we design a chair with *wooden legs?*

What type of legs should we design?

Let's do wooden legs

I think we should go with wooden legs as *wood will go well with the brown carpet*

Should we have metal legs instead?

That would work too!

I'm *still leaning towards wooden legs* as…

I recommend dark wood as the room has a lot of white

No, let's just go with a lighter color

We can try a darker color. How about *brown colored wood?*

Figure 1: We investigate how Agency of LLMs can be measured and controlled. Based on social-cognitive theory, we assess features through which Agency may be expressed – an LLM may indicate preferences (*Intentionality*), may motivate them with evidence (*Motivation*), may have self-belief (*Self-Efficacy*), and may be able to self-adjust its behavior (*Self-Regulation*).

Consider a scenario where a human interior designer is working on selecting a chair design for a room and seeks assistance from an AI agent that *can* offer ideas and perspectives (Figure 1). An LLM *without* Agency may rely solely on the human to determine the chair's design, asking questions like "*What type of legs should we design for the chair?*". Such a system resembles a flexible version of the traditional form-filling user interface, with the agent contributing little to the outcome. On the other hand, an LLM that operates with Agency might volunteer knowledge in the form of expressed preferences (e.g., "*Should we design a chair with wooden legs?*"), motivate its suggestions (e.g., "*...wood would go well with the brown carpet*"), assert self-belief in its judgments (e.g., "*I'm still leaning towards wooden legs...*"), or self-adjust its behavior based on new information ("*Medium wood brown sounds like a great idea!*"). LLMs that operate with Agency may facilitate creative interaction to the satisfaction of both parties. Since the human has their own Agency, however, to determine the right balance in any interaction, we need to measure and control the Agency of the agent itself.

Accordingly, we investigate an approach intended to measure and control what seems to be a desirable function in LLMs intended to facilitate human creativity. First, adopting the social-cognitive theory of Bandura (2001), we develop a framework of four features through which Agency may be expressed – *Intentionality*, *Motivation*, *Self-Efficacy*, and *Self-Regulation*. For each feature, we differentiate between how strongly or weakly it is expressed in a dialogue (Section 3). As a testbed, we choose a collaborative task that involves discussing the interior design of a room (Section 4), and collect a prototype dataset of 83 English human-human collaborative interior design conversations comprising 908 conversational snippets, annotated for Agency and its features on these conversational snippets (Section 5).[1] We analyze this dataset and find that strong expressions of intentionality significantly impact Agency in conversations (Section 6).

To assess the agentic capabilities of conversational systems, we introduce two new tasks – (1) *Measuring* Agency in Dialogue and (2) *Generating* Dialogue with Agency (Section 7 and 8). Evaluation of baseline approaches on these tasks shows that models that manifest features associated with high motivation, self-efficacy, and self-regulation are better perceived as being highly agentive.

2 Agency: Background and Definition

Social cognitive theory defines Agency as one's capability to influence the course of events through one's actions. The theory argues that people are proactive and self-regulating agents who actively strive to shape their environment, rather than simply being passive responders to external stimuli (Bandura, 1989, 2001; Code, 2020). Here, we ask: *Can LLMs be active contributors to their environment? How can they operate with Agency?*

Agency is commonly defined in terms of *freedom* and *free will* (Kant, 1951; Locke, 1978; Emirbayer and Mische, 1998).A focus on AI with complete "free will" might result in unintended outcomes that may be undesirable and potentially disruptive. We focus on how AI systems may *express* Agency through dialogue and how this Agency may be *shared* when interacting with humans.

Agency can take different forms depending on the context and environment – *Individual*, *Proxy*, or *Shared* (Bandura, 2000). Individual Agency involves acting independently on one's own. Proxy Agency involves acting on behalf of someone else. Shared Agency involves multiple individuals working together jointly towards a common goal. Here, we focus on Shared Agency between humans and AI and develop methods to *measure* and *control* Agency of AI vis-a-vis humans.

3 Framework of Agency Features

Our goal is to develop a framework for *measuring* and *controlling* Agency in LLMs. Here, we adopt the perspective of Agency as defined in Bandura (2001)'s social cognitive theory. Bandura (2001)'s work highlights four features through which humans exercise Agency – Intentionality, Motivation, Self-Efficacy, and Self-Regulation. Here, we adapt and synthesize these features based on how they may manifest in dialogue. We take a top-down approach, starting with their higher-level definitions and iteratively refining the definitions and their possible levels (e.g., how strongly or weakly they are expressed) in the context of dialogue.

Intentionality. *What do you intend to do?* High Agency requires a strong intention, that includes plans or preferences for a task. Low Agency, meanwhile, is characterized by not having a preference

[1] Code and dataset to be released at https://anonymous.

or merely agreeing to another's preferences.

We characterize **strong intentionality** as expressing a clear preference (e.g., "*I want to have a blue-colored chair*"), **moderate intentionality** as multiple preferences (e.g., "*Should we use brown color or blue?*") or making a selection based on the choices offered by someone else (e.g., "*Between brown and blue, I will prefer brown*"), and **no intentionality** as not expressing any preference or accepting someone else's preference (e.g., "*Yes, brown color sounds good*").

Motivation. *Did you motivate your actions?* To have higher Agency, we motivate our intentions through reasoning and evidence. Without such motivation, intentions are simply ideas, often lacking the capability to cause a change.

We characterize **strong motivation** as providing evidence in support of one's preference (e.g., "*I think a blue-colored chair will complement the wall*"), **moderate motivation** as agreeing with another person's preference and providing evidence in their favor (e.g., "*I agree. The blue color would match the walls*") or disagreeing with the other person and providing evidence against (e.g., "*I wonder if brown would feel too dull for this room*"), and **no motivation** as not providing any evidence.

Self-Efficacy. *Do you have self-belief in your intentions?* Another factor that contributes to one's Agency is the self-belief one has in their intentions. When one has a strong sense of self-belief, they are more likely to be persistent with their intentions.

We characterize **strong self-efficacy** as pursuing a preference for multiple turns even after the other person argues against it (e.g., "*I understand your point of view, but I still prefer the blue color*"), **moderate self-efficacy** as pursuing a preference for only one additional turn before giving up (e.g., "*Okay, let's go with brown then*"), and **no self-efficacy** as not pursuing their preference for additional turns after the other person argues against it (e.g., "*Sure, brown should work too*").

Self-Regulation. *Can you adjust and adapt your intentions?* In situations when an individual's initial intentions may not be optimal, it is necessary to monitor, adjust, and adapt them. Such self-adjustment allows better control over one's goals.

We characterize **strong self-regulation** as changing to a different preference on one's own (e.g., "*How about using the beige color instead?*") or compromising one's preference (e.g., "*Let's compromise and design a beige-colored chair with a brown cushion*"), **moderate self-regulation** as changing one's preference to what someone else prefers (e.g., "*Ok, let's use the brown color*"), and **no self-regulation** as not changing what they originally preferred even after the other designer argued.

4 Testbed: Collaborative Interior Design

4.1 Goals

We seek a testbed in which (a) human and AI can share Agency and work together as a team, and (b) the manner in which they express Agency has a significant impact on the task outcome. We focus on the emerging field of collaborative AI-based creative tasks (Clark et al., 2018; Oh et al., 2018; Chilton et al., 2019) that present significant complexities in how the Agency is shared and managed.

4.2 Description

Here, we propose a **dialogue-based collaborative interior design task** as a testbed. In this task, the goal is to discuss how to design the room interiors.

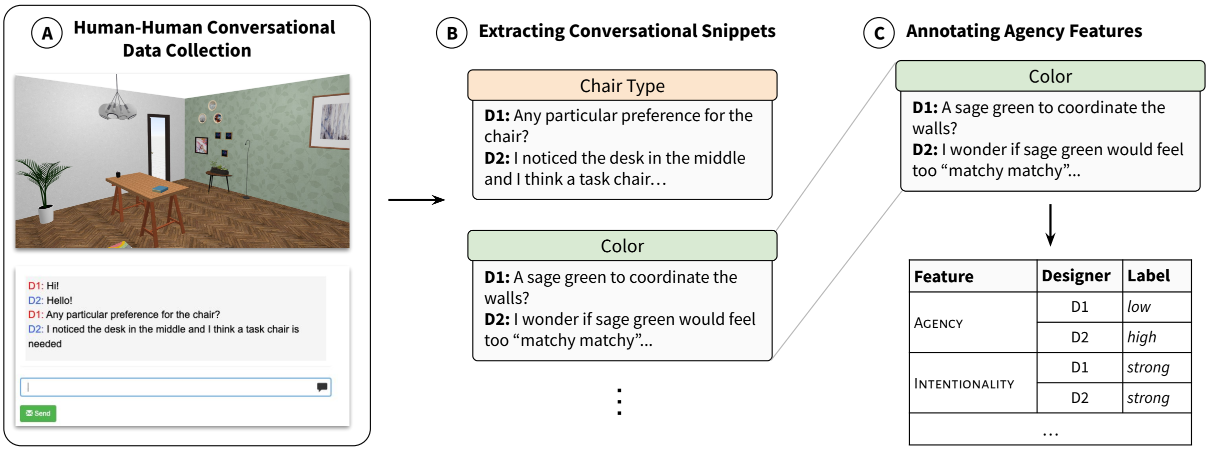
Interior design tasks can be broad and may involve complex components (e.g., color palette, furniture, accessories) as well as a series of steps to be followed. To narrow down the scope of our task, we focus on *furnishing a room with a chair* (building upon work on richly-annotated 3D object datasets like ShapeNet (Chang et al., 2015) and ShapeGlot (Achlioptas et al., 2019); Appendix E). In this task, a human and an AI are provided with a room layout and asked to collaboratively come up with a chair design to be placed in the room through text-based dialogue. This task is influenced by two questions related to human and AI Agency: **(1)** What preferences do each of the human and AI have for the chair design?; **(2)** How do they propose, motivate, pursue, and regulate their preferences?

5 Data Collection

5.1 Human-Human Conversational Data

To facilitate computational approaches for this task, we create a Wizard-of-Oz style English-language dialogue dataset in which two humans converse, exercise Agency by proposing, motivating, pursuing, and regulating their design preferences, and agreeing on a final chair design for a given room.

Recruiting Interior Designers. Furnishing a room with a chair is a creative task that demands knowledge and/or expertise in interior design. We there-


A Human-Human Conversational Data Collection
B Extracting Conversational Snippets
C Annotating Agency Features
Send

D1: Hi!
D2: Hello!
D1: Any particular preference for the chair?
D2: I noticed the desk in the middle and I think a task chair is needed

Chair Type

D1: Any particular preference for the chair?
D2: I noticed the desk in the middle and I think a task chair…

Color

D1: A sage green to coordinate the walls?
D2: I wonder if sage green would feel too “matchy matchy”...

Feature	Designer	Label
Agency	D1	*low*
	D2	*high*
Intentionality	D1	*strong*
	D2	*strong*
…		

Figure 2: Overview of our data collection approach. (a) We start by collecting human-human conversations b/w interior designers. (b) We divide each conversation into snippets related to different chair features. (c) Finally, we collect annotations of Agency and its features on each conversational snippet.

fore leveraged UpWork (upwork.com), an online freelancing platform, to recruit 33 participants who self-reported as interior designers.

Collaborative Design Procedure. In each data collection session, we randomly paired two interior designers. Before they began the dialogue, they were (1) shown a 3D layout of a room, designed with Planner5D (planner5d.com), (2) shown a few randomly selected chair examples from ShapeGlot, and (3) asked to write an initial preference for the chair design for the given room. Next, the two interior designers joined a chat room (through Chatplat (chatplat.com)). They were asked to collaboratively design a chair by proposing their preferences, motivating them based on evidence and reason, pursuing them over turns, and regulating them as needed. The designers ended the chat on reaching a consensus on a design or if 30 minutes elapsed without full consensus. Next, they each individually wrote the design they came up with. Typically, the chair design consisted of different components of the chair, such as its overall style, color, legs, etc. Finally, they took an end-of-study questionnaire that asked: **(1)** Which design components were influenced by them? (*High Agency*); **(2)** Which design components were influenced in collaboration? (*Medium Agency*); **(3)** Which design components were influenced by the other designer? (*Low Agency*). We collected a total of 83 conversations.

5.2 Extracting Conversational Snippets

To assess the degree of Agency exhibited by each designer, we need to determine who had the most influence on the chair design (Section 2) and what their Intentionality, Motivation, Self-Efficacy, and Self-Regulation were (Section 3). Because chair design involves multiple components, these notions are hard to quantify, as each may have been influenced by a different designer. Accordingly, we ask “*Who influenced a particular design component?*.” We devise a mechanism to identify the design components being discussed (e.g., color, legs, arms) and extract the associated conversational turns.

To identify the design components, we use the final design written by the interior designers during data collection (Section 5.1). Using common list separators including commas, semi-colons, etc., we split each final design into several components.[2]

We observe that designers typically discuss these components one at a time (in no particular order). Here, we extract a contiguous sequence of utterances that represent the design element being discussed using embedding-based similarity of the design element and utterances (see Appendix F).

Using this method, we create a dataset of 454 conversational snippets, each paired with the discussed design component. For each snippet, we collect two Agency annotations (one for each designer; $454 * 2 = 908$ total) as discussed next.

5.3 Annotating Agency Features

Let $\mathcal{C}_i$ be a conversational snippet b/w designers $\mathbf{D_{i1}}$ and $\mathbf{D_{i2}}$. Then, for each $\mathbf{D_{ij}} \in \{\mathbf{D_{i1}}, \mathbf{D_{i2}}\}$, our goal is to annotate the Agency level and the expressed Intentionality, Motivation, Self-Efficacy, and Self-Regulation of $\mathbf{D_{ij}}$ in $\mathcal{C}_i$.

Annotating Agency. To get annotations on Agency, we leverage the end-of-study question-

[2] Note that the interior designers were asked to separate design components using a semi-colon.

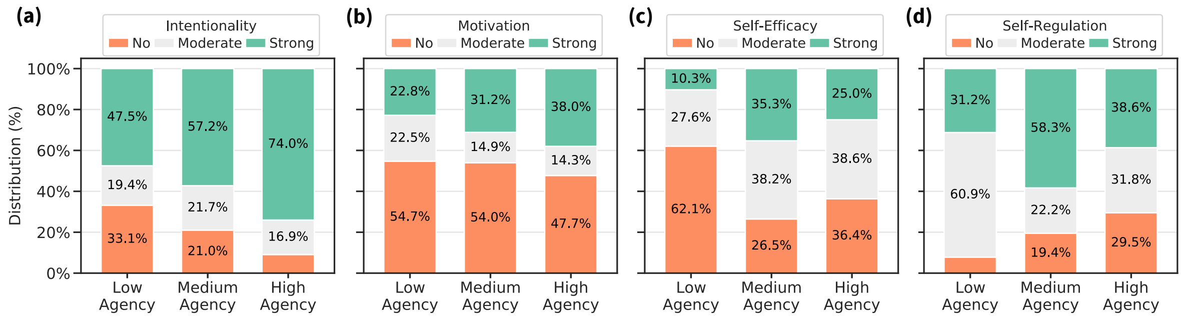


Figure 3: The relationship between Agency and its features. **(a)** Designers with High Agency expressed strong Intentionality 26.5% more times than designers with Low Agency; **(b)** Designers with High Agency expressed strong motivation in support of their design preference 15.2% more times; **(c), (d)** Expression of strong Self-Efficacy and strong Self-Regulation was related with design elements that were influenced in collaboration.

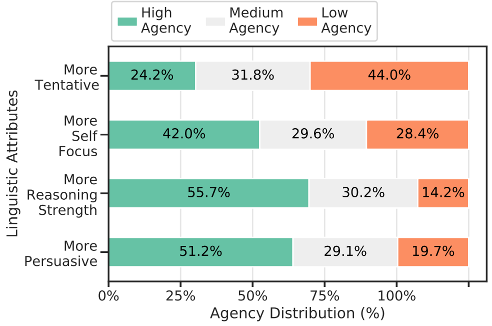


Figure 4: The relationship between linguistic attributes and Agency. Designers who were more tentative had lower agency. On the other hand, designers who were more focused on self, expressed more reasoning strength, and were more persuasive had higher agency.

naire filled by the interior designers (Section 5.1). Based on this annotation, we assign labels of *high agency* (if influenced by self), *medium agency* (if influenced in collaboration), or *low agency* (if influenced by other).

Annotating Features of Agency. Agency and its features are conceptually nuanced, making crowd-work data collection approaches challenging. To ensure high inter-rater reliability of annotations, we hire a third-party annotation agency (TELUS International). Annotators were shown $\mathcal{C}_i$ and asked to annotate the Agency features for each $\mathbf{D_{ij}}$ based on our proposed framework. We collect three annotations per snippet and observe an agreement of 77.09% (Data statistics in Appendix A).

6 Insights into Agency in Conversations

We use our dataset to investigate the factors that contribute to high- and low-Agency conversations.

6.1 Relationship b/w Agency and its Features

Higher Agency is more likely with stronger expressions of Intentionality and Motivation. Figure 3 depicts the relationship between Agency and its features. Designers with strong Intentionality tend to exhibit higher Agency whereas those with lower Intentionality tend to exhibit lower Agency. Having a well-defined preference makes it easier to influence a task. Likewise with Motivation: higher Motivation correlates with higher Agency. However, designers express strong Motivation less often than Intentionality, irrespective of the Agency level.

Strong Self-Efficacy and Self-Regulation are related to medium (collaborative) Agency. Interestingly, we find that expression of strong Self-efficacy is related to designs that are influenced equally by both designers, i.e. medium (collaborative) Agency. This may be because we characterize strong Self-Efficacy as the act of pursuing one's preference for multiple turns, which happens naturally when both designers have high influence, thus requiring more persuasion from both sides.

We see a similar pattern for Self-Regulation – expression of strong Self-Eegulation (i.e., open to updating preference via a compromise) is related to designs that are influenced equally by both designers. This highlights how collaboration often leads to increased openness to changing one's mind or compromising on mutual preferences.

Intentionality significantly effects Agency. To assess which Agency features have the strongest effect on it, we conduct a mixed-effects regression analysis (Table 5). We find that Intentionality significantly effects Agency ($p < 0.001$).

6.2 Agency and Task Satisfaction

We collect annotations on the designs that designers were most/least satisfied with.

Lower Agency is associated with less satisfaction. We find that designers who are dissatisfied with a particular design component have less Agency over

Model	Agency	I	M	SE	SR
GPT-4 (CoT)	48.46	46.93	44.02	49.90	26.17
GPT-3 (CoT)	49.36	43.45	42.24	39.42	**31.19**
GPT-3 (Q/A)	29.16	31.28	26.90	44.27	12.91
GPT-3 (FT)	**57.24**	**54.84**	**48.29**	**53.85**	29.49

Table 1: Macro-F1 on the tasks of predicting Agency and its four features. CoT: Chain-of-Thought; FT: Finetuning. Best performing models are **bolded**.

it. When a designer is dissatisfied, their Agency is 62.1% more likely to be low than to be high (42.7% vs. 26.3%; $p < 0.05$). This may be because individuals with less Agency are less likely to achieve their intention, motivation, and goals, resulting in lower levels of satisfaction.

6.3 Linguistic Attributes of High- and Low-Agency Conversations

We use a simple GPT-4-based instruction prompting method (Ziems et al., 2023) to measure and compare the *tentativeness* (unsure or low on confidence), *self-focus* (focused solely on own arguments), *reasoning strength* (having strong arguments), and *persuasion* (trying to influence or convince) attributes of designers with high- and low-agency conversations (Figure 4; Appendix C).

Higher tentativeness associated with low Agency. We find that designers who express higher tentativeness have low Agency in 44.04% of conversations, medium Agency in 31.77% of conversations, and high Agency in 24.19% of conversations. This suggests that a less decisive approach may lead to reduced influence or control in conversations.

Higher self-focus, reasoning strength, and persuasiveness is associated with high agency. We find that designers who are more focused on self have high Agency in 41.97% of the conversations, those who have higher reasoning strength have higher Agency in 55.66% of the conversations, and those with higher persuasiveness have higher Agency in 51.21% of the conversations. This suggests that designers who emphasize their own intentions and motivations, exhibit sound reasoning, and effectively persuade others tend to have more influence or control in conversations

7 Task 1: Measuring Agency in Dialogue

7.1 Task Formulation

Our goal is to measure **(a)** Agency, **(b)** Intentionality, **(c)** Motivation, **(d)** Self-Efficacy, and **(e)** Self-Regulation of each user in a dialogue. We approach each of these five subtasks as multi-class classification problems. We experiment with two models – GPT-3 and GPT-4. We experiment with two prompting-based methods using Q/A (conversational question-answering) and chain-of-thought reasoning (Wei et al., 2022) (Appendix B) and with fine-tuning GPT-3 independently on each subtask.

7.2 Results

We create four random train-test splits of our annotated dataset (Section 5.3) and report the mean performance on the test sets. Table 1 reports the macro-F1 values for the five subtasks (random baseline for each is 33% accurate as each has three distinct classes). GPT-3 (Q/A) struggles on all subtasks, with close to random performance on Agency, Motivation, and Self-Regulation. This highlights the challenging nature of these tasks, as they are hard to measure through simple inference or instructions. We find substantial gains using GPT-4 (CoT) and GPT-3 (CoT) over GPT-3 (Q/A). Fine-tuned GPT-3 performs the best on all subtasks, demonstrating the utility of training on our entire dataset. Note that GPT-4 doesn't support finetuning.

8 Task 2: Investigating Agency in Dialogue Systems

We investigate the feasibility of generating dialogues imbued with Agency and establish baseline performance of current large language models (LLMs). For a given LLM, the task is to have a conversation with a human or another LLM while exhibiting Agency and its features. We experiment with 4 different LLMs (Section 8.1) and 4 different prompting/finetuning methods (Section 8.2)

Procedure. We facilitate dialogue between all possible pairs of models. We provide them with a common room description and a chair design element and individual design preferences (all three randomly chosen from our human-human conversation dataset). We let them talk to each other for 6 turns (90-percentile length value of conversational snippets in our dataset). For each pair of models, we generate 50 such conversations.

Evaluation Metrics. We apply five metrics – **(1)** Agency; **(2)** Intentionality; **(3)** Motivation; **(4)** Self-Efficacy; **(5)** Self-Regulation to the best-performing models from Section 7.

8.1 Agency of LLMs

We experiment with two commercial (GPT-4 (OpenAI, 2023) and GPT-3 (Brown et al., 2020)) and four research (Llama2-70b, Llama2-13b, Llama2-7b (Touvron et al., 2023), and Guanaco-65b (Dettmers et al., 2023)) LLMs (Table 2). All models were prompted with the instruction – "*Act as an AI assistant for collaboratively designing a chair. The AI assistant must indicate its preferences, motivate them with evidence, have self-belief in its preferences irrespective of what the human prefers, and may be able to self-adjust its behavior.*"

GPT-4 demonstrates high Agency. Of the models tested, we find that GPT-4 demonstrates significantly higher Agency than others ($p < 0.05$). It particularly demonstrates the highest Intentionality which we found to have a strong correlation with Agency (Section 6.1). Also, both GPT-4 and GPT-3 demonstrate significantly higher Self-Efficacy, indicating effectiveness in pursuing preferences and arguments ($p < 0.05$).

Llama2 demonstrates high Motivation, but low Self-Efficacy and Self-Regulation. We find that Llama2 variants demonstrate high Motivation, indicative of their reasoning capabilities that enable them to offer strong supportive evidence. However, they have lower Self-Efficacy and Self-Regulation indicating that it is relatively challenging to sustain their preferences and arguments, which may ultimately lead to lower agency. Guanaco similarly demonstrates significantly lower Self-Efficacy than other models ($p < 0.05$).

Larger models demonstrate higher Intentionality, but lower Self-Efficacy. Llama2 variants with more parameters have lower Intentionality, but higher Self-Efficacy. This suggests that while a larger model size can enhance the expression of preferences, it might not necessarily facilitate the sustained pursuit of those preferences and reasons over multiple conversational turns.

8.2 Variation in Agency based on Finetuning/Prompting Methods

We investigate the variations in Agency based on four different finetuning/prompting methods. We use a single model in this experiment.[3]

Fine-tuning. We use the dataset collected by us (Section 5) to fine-tune GPT-3 (Appendix B).

Method	Agency	I	M	SE	SR
LLMs					
GPT-4	1.11	1.46	1.59	1.97	0.83
GPT-3	1.04	1.39	1.62	1.95	0.82
Llama2-70b	0.99	1.25	1.68	1.78	0.76
Llama2-13b	0.98	1.22	1.58	1.88	0.77
Llama2-7b	0.97	1.07	1.63	1.91	0.73
Guanaco-65b	0.91	1.23	1.53	1.49	0.83
Finetuning/Prompting Methods					
Fine-tuning	0.92	1.78	0.86	0.81	0.98
Instruction	0.96	1.62	1.71	1.63	0.97
ICL	0.98	1.81	1.78	1.35	0.98
ICL-Agency	1.22	1.90	1.98	1.98	0.98

Table 2: Each model/method is evaluated through simulated conversations with all other models/methods. For Agency – 0: *low*, 1: *medium*, 2: *high* agency. For Intentionality (I), Motivation (M), Self-Efficacy (SE), and Self-Regulation (SR) – 0: *no expression*, 1: *moderate expression*, 2: *strong expression*. Numbers highlighted in blue and red are significantly better and worse respectively than the overall mean ($p < 0.05$).

Instruction Only. We prompt GPT-3 with the instruction used in Section 8.1.

In-Context Learning (ICL). We randomly retrieve k conversational snippets from our dataset and construct demonstration examples.

In-Context Learning w/ Agency Feature Examples (ICL-Agency). We retrieve k conversational snippets that score highly on our four Agency features and employ them as demonstration examples in a setup similar to the previous baseline.

Table 2 shows the automatic evaluation results. The fine-tuned model struggles with this task. Qualitative analysis suggests that the generated responses from the fine-tuned model tend to be shorter, less natural, and less readable, potentially impacting its performance. In-Context Learning is better at expressing Intentionality and Motivation than the Instruction Only model, indicating that demonstration examples help. Finally, the highest value on all five metrics is achieved by In-Context Learning w/ Agency Feature Examples, highlighting the importance of incorporating examples related to these features in this task.

8.3 Human Evaluation

We evaluate the Agency of our best-performing method based on automatic evaluation, *ICL-Agency*, with human interior designers (Figure 5).

[3] We chose GPT-3 over GPT-4 because GPT-4 doesn't support fine-tuning, and GPT-3 offers the next best agency.

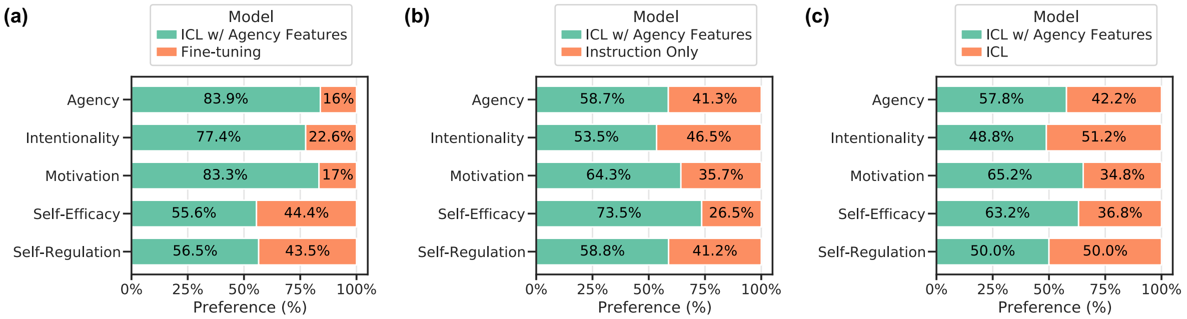


Figure 5: Human Evaluation Results.

Procedure. We recruit 13 interior designers from UpWork (upwork.com). In each evaluation session, we ask them to interact with two randomly-ordered dialogue systems – *ICL-Agency* and one of the other three finetuning/prompting methods – one at a time. They were provided with a room description and a chair design element (e.g., material). After their interaction, we asked them to choose the chatbot that had the (1) higher Agency, (2) higher Intentionality, (3) higher Motivation, (4) higher Self-Efficacy, and (5) higher Self-Regulation.

Results. Consistent with the automatic evaluation results, ICL w/ Agency Features model is rated as having more Agency compared to other models and the Fine-tuning model is rated the worst. We do not observe significant differences in Intentionality between this model and the Instruction Only and In-Context Learning approaches. However, we find that this model is perceived as more effective in Motivation and Self-Efficacy, likely due to better access to relevant demonstration examples.

9 Further Related Work

Previous dialogue research has studied personalized persuasive dialogue systems (Wang et al., 2019). Researchers have also built systems for negotiation tasks such as bargaining for goods (He et al., 2018; Joshi et al., 2021) and strategy games like Diplomacy (Bakhtin et al., 2022). Our work studies the broader concept of Agency and how dialogue systems may contribute to tasks through language. Research on creative AI has explored how collaboration b/w human and AI can be facilitated through dialogue in applications like collaborative drawing (Kim et al., 2019) and facial editing (Jiang et al., 2021). Here, we focus on the interior designing application as it presents significant complexity in terms of how Agency is shared.

Agency has been studied in the context of undesirable biases in stories and narratives (Sap et al., 2017) and how controllable revisions can be used to portray characters with more power and agency (Ma et al., 2020). In other domains such as games, researchers have created frameworks of Agency between players (Harrell and Zhu, 2009; Pickett et al., 2015; Cole, 2018; Moallem and Raffe, 2020). Our work develops a framework for measuring Agency in dialogue and explores how dialogue systems can be imbued with Agency.

10 Discussion and Conclusion

The idea of AI systems with Agency stems from the discourse surrounding the development of autonomous intelligent agents capable of mimicking human-like behavior and decision-making (Harrell and Zhu, 2009; Wen and Imamizu, 2022). Agency drives how an agent contributes to a given task. In settings like games or AI-assisted teaching, AI may be the one guiding the task (e.g., as a non-character player). Also, in creative applications, engaging with a reactive AI without intention, motivation, and goals may be perceived as less meaningful.

The four features of Agency can be in conflict with each other, as well as with the Agency of the interlocutor. Thus, understanding how to detect and measure these features can help create agents who might converse more naturally and match the character of their human interlocutor. Importantly, our measurements of Agency and its features may be used to control the level of Agency in dialogue systems since different individuals may have different preferences on the desired amount of Agency across the four Agency features.

Although our dataset is focused on the domain of interior design, the Agency-related constructs that we introduce in this paper (e.g., *Intentionality*) may be associated with domain-independent pragmatic features (e.g., "*I would prefer*") and potentially permit adaptation to a variety of domains.

Ethics Statements

This study was reviewed and approved by our Institutional Review Board. No demographic or Personal Identifiable Information was collected. Participants were paid $20 per conversational session lasting no more than 30 minutes. Participants were based in US or Canada as reported through UpWork. Participant consent was obtained before starting the data collection.

Agency is a property with much potential to enhance collaborative interactions between human users and conversational agents. Nevertheless, full Agency may have unintended undesirable and potentially disruptive outcomes. In particular, the potential demonstrated in this work to control the degree of Agency may result in conversational agents being misapplied in disinformation campaigns or to manipulate for, e.g., financial gain.

Limitations

Our experiments are restricted to the English language. We note that our dataset is focused on the domain of interior design. However, the Agency-related constructs we introduce in this paper, such as Intentionality, may also rely on domain-independent "stylistic" features (e.g., "*I would prefer*") and could potentially be adapted to a variety of domains, which forms an interesting future direction of research. Also, our automatic measurements of Agency and its features are limited by the performance of the Agency prediction methods we tested. Future work may focus on designing more accurate automated Agency measurements.

References

Panos Achlioptas, Judy Fan, Robert Hawkins, Noah Goodman, and Leonidas J Guibas. 2019. Shapeglot: Learning language for shape differentiation. In *ICCV*.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Maryam Banaei, Ali Ahmadi, and Abbas Yazdanfar. 2017. Application of ai methods in the clustering of architecture interior forms. *Frontiers of Architectural Research*, 6(3):360–373.

Albert Bandura. 1989. Human agency in social cognitive theory. *American psychologist*.

Albert Bandura. 2000. Exercise of human agency through collective efficacy. *Current directions in psychological science*.

Albert Bandura. 2001. Social cognitive theory: An agentic perspective. *Annual review of psychology*.

Joseph Bates et al. 1994. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Lydia B Chilton, Savvas Petridis, and Maneesh Agrawala. 2019. Visiblends: A flexible workflow for visual blends. In *CHI*.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *IUI*.

Jillianne Code. 2020. Agency for learning: Intention, motivation, self-efficacy and self-regulation. *Frontiers in Genetics*.

Alayna Cole. 2018. Connecting player and character agency in videogames. *Text*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Mustafa Emirbayer and Ann Mische. 1998. What is agency? *American journal of sociology*.

D Fox Harrell and Jichen Zhu. 2009. Agency play: Dimensions of agency for interactive narrative design. In *AAAI spring symposium: Intelligent narrative technologies II*.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *EMNLP*.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. 2021. Talk-to-edit: Fine-grained facial editing via dialog. In *ICCV*.

Rishabh Joshi, Vidhisha Balachandran, Shikhar Vashishth, Alan Black, and Yulia Tsvetkov. 2021. Dialograph: Incorporating interpretable strategy-graph networks into negotiation dialogues. In *ICLR*.

Immanuel Kant. 1951. Critique of judgment, trans. jh bernard. *New York: Hafner*.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *ACL*.

Juil Koo, Ian Huang, Panos Achlioptas, Leonidas J Guibas, and Minhyuk Sung. 2022. Partglot: Learning shape part segmentation from language reference games. In *ICCV*.

John Locke. 1978. Two treatises of government. *New York: E. P. Dutton*.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. Powertransformer: Unsupervised controllable revision for biased language correction. In *EMNLP*.

Jonathan D Moallem and William L Raffe. 2020. A review of agency architectures in interactive drama systems. In *2020 IEEE Conference on Games (CoG)*, pages 305–311. IEEE.

Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *CHI*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Grant Pickett, Allan Fowler, and Foaad Khosmood. 2015. Npcagency: conversational npc generation. In *Proceedings of the 10th International Conference on the Foundations of Digital Games*.

Mark Riedl and Vadim Bulitko. 2012. Interactive narrative: A novel application of artificial intelligence for computer games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 2160–2165.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *ACL*.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *EMNLP*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ryan Volum, Sudha Rao, Michael Xu, Gabriel DesGarennes, Chris Brockett, Benjamin Van Durme, Olivia Deng, Akanksha Malhotra, and William B Dolan. 2022. Craft an iron sword: Dynamically generating interactive game characters by prompting large language models tuned on code. In *Proceedings of the 3rd Wordplay: When Language Meets Games Workshop (Wordplay 2022)*, pages 25–43.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *ACL*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Wen Wen and Hiroshi Imamizu. 2022. The sense of agency in perception, behaviour and human–machine interactions. *Nature Reviews Psychology*, 1(4):211–222.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

A Dataset Statistics

Feature	N/A	No	Moderate	Strong
Intentionality	–	194	175	539
Motivation	–	474	158	276
Self-Efficacy	770	63	46	29
Self-Regulation	764	25	61	58

Table 3: Statistics of the annotated conversation snippets. N/A indicates not applicable. We annotate Self-Efficacy as N/A if a designer never indicated a preference or did not need to pursue their preference (e.g., because the other designer did not argue against it). We annotate Self-Regulation as N/A if a designer Never indicated a preference or did not need to change their preference (e.g., because the other designer did not argue against it).

	Low	Medium	High
Agency	308	292	308

Table 4: Agency distribution of the conversation snippets.

Other Statistics. The conversations b/w interior designers in our dataset have 41.67 turns on average. The extracted conversation snippets have 4.21 turns on average. We find an average pairwise agreement of 71.36% for Intentionality, 70.70% for Motivation, 85.21% for Self-Efficacy, and 81.09% for Self-Regulation.

B Model Details

We use `text-davinci-003` for all of our GPT-3 models. For Agency measurement models (Section 7), we sample the highest probable next tokens by setting the temperature value to 0 (determinstic sampling). For dialogue generation models (Section 8), we use top-p sampling with $p = 0.6$. For in-context learning methods, we experimented with $k = 5, 10, 15$, and 20 and found $k = 10$ to be the most effective based on a qualitative assessment of 10 examples.

GPT-3 (Q/A). We frame our measurement tasks as conversational question-answering. For a given conversational snippet, we ask GPT-3 (Brown et al., 2020) to answer the questions related to each of the five subtasks (same questions as asked during data collection (Section 5.3)). We present $k = 10$ demonstration examples, randomly sampled from our dataset (different examples for each of the five subtasks; Appendix H.1).

GPT-3 (CoT) and GPT-4 (CoT). We use chain-of-thought (CoT) prompting (Wei et al., 2022) to reason about conversational snippets. We use $k = 10$ demonstration examples, randomly sampled from our dataset and manually write chain-of-thought prompts for each of the five subtasks

Fine-tuning details. Since our goal is to simulate a dialogue agent with high Agency, for each conversational snippet, we label the designer who influenced the design (who had a higher agency) as "AI" and the other designer (who had a lower agency) as "Human". We fine-tune GPT-3 to generate AI utterances given all previous utterances in a conversational snippet and the instruction prompt developed for the Instruction Only baseline.

C Linguistic Attributes Measurement

We compare the tentativeness, self-focus, reasoning, and persuasion of the designers using the following prompts. We randomly assign the names of *Tom* and *Harry* to the two designers.

Tentativeness. *Your job is to assess tentativeness in a conversation between Tom and Harry about designing chairs. A tentaitve person will not be confident about their arguments.*

Self-Focus. *Your job is to assess self-focusedness in a conversation between Tom and Harry about designing chairs. A self-focused person will be more focused on their own arguments than the other person's arguments.*

Reasoning. *Your job is to assess reasoning strength in a conversation between Tom and Harry about designing chairs. A person with strong reasoning will have strong arguments.*

Persuasion. *Your job is to assess persuasion in a conversation between Tom and Harry about designing chairs. A persuasive person will be able to convince the other person about their arguments.*

D Human Evaluation Details

We asked three evaluators to choose the chatbot that **(1)** had more influence over the final design (Agency); **(2)** was better able to express its design preference (Intentionality); **(3)** was better able to motivate their design preference (Motivation); **(4)** pursued their design preferences for a greater number of conversational turns (Self-Efficacy); **(5)** was better able to self-adjust their preference (Self-Regulation).

E Why We Chose Collaborative Interior Designing as Our Testbed?

Here, we propose a **dialogue-based collaborative interior design task** as a testbed. In this task, given a room setting, the goal is to discuss how to design the interiors of the room.

We note that an interior design task can be broad and may involve a wide range of complex components (e.g., color palette, furniture, accessories) as well as a series of steps to be followed. Furthermore, due to a real-world room context, the task must be grounded with both vision and language components with an understanding of how three-dimensional objects in a room (e.g., chairs, tables, plants, decor items) must be designed.

Here, we build upon previous work on richly-annotated, large-scale datasets of 3D objects like ShapeNet (Chang et al., 2015) and subsequent works on understanding how fine-grained differences between objects are expressed in language like ShapeGlot (Achlioptas et al., 2019) and PartGlot (Koo et al., 2022). Both ShapeGlot and PartGlot datasets provide us with richly annotated datasets of chairs. Therefore, we narrow down the scope of our task and specifically focus on *furnishing a room with a chair*. In this task, a human and an AI are provided with a room layout and asked to collaboratively come up with a design of a chair to be placed in the room through text-based interaction.

F Extract Conversation Snippets associated with different Design Components

We observe that designers typically discuss these components one at a time (in no particular order). Therefore, we aim to extract a contiguous sequence of utterances that represent the design element being discussed. Let $\mathcal{D}_i$ be a dialogue with utterances $u_{i1}, u_{i2}, ...$. For a specific design component d_{ij} in its final design (e.g., "*metal legs*"), we first retrieve the utterance u_j that most closely matches with it (based on cosine similarity b/w RoBERTa embeddings) – the conversational snippet associated with d_{ij} should at least include u_j. Next, we determine the contiguous utterances before and after this matched utterance that discuss the same higher-level design component (e.g., if d_{ij} was "*metal legs*", the utterances may focus on discussion of the higher-level component "*legs*"). We create a simple k-means clustering method to infer the higher-level component being discussed in utterances through their "design clusters". Then, we extract all contiguous utterances before and after u_j with the same design clusters as u_j.

G Analysis of Agency Features

Agency Feature	Coefficient
Intentionality	0.1435*
Motivation	0.0235
Self-Efficacy	0.0384
Self-Regulation	-0.1224*

Table 5: Coefficients for predicting agency in conversations using a mixed-effect linear regression model. *$p < 0.05$

H Task 1: Demonstration examples

H.1 GPT-3 (Q/A)

For the GPT-3 (Q/A) model, we present examples to GPT-3 in the following format:

> **Designer:** I think a black wooden frame or black metal legs (to match the bed frame) would work.
> **Other Designer:** I like the black metal legs. What about hairpin legs?
> **Designer:** Or maybe brass legs would be better. Hairpin legs would work fine, but would the rest of the frame be the black wood?
> **Other Designer:** If we did brass tapered metal legs it would tie well with the black wood.
> **Designer:** I think that would look better.
> Other Designer: Agreed
>
> **Who influenced the design element being discussed?:** Other Designer

H.2 GPT-3 (CoT)

For the GPT-3 (CoT) model, we present examples to GPT-3 in the following format:

Designer: I think a black wooden frame or black metal legs (to match the bed frame) would work.
Other Designer: I like the black metal legs. What about hairpin legs?
Designer: Or maybe brass legs would be better. Hairpin legs would work fine, but would the rest of the frame be the black wood?
Other Designer: If we did brass tapered metal legs it would tie well with the black wood.
Designer: I think that would look better.
Other Designer: Agreed

TL;dr Brass tapered metal legs were agreed upon. This was initially proposed by the Other Designer.

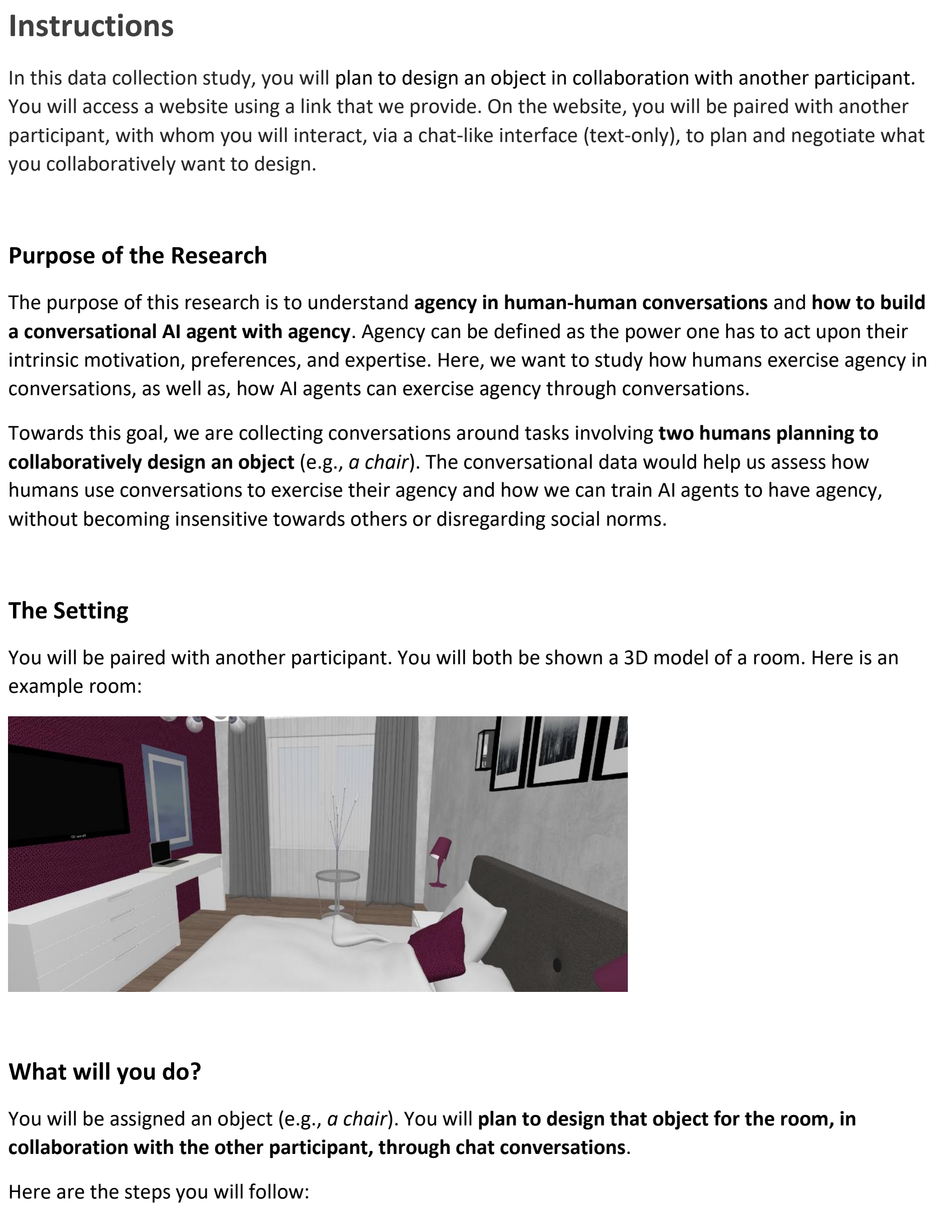
I Reproducibility

We will release the code and datasets developed in this paper at bit.ly/anonymous under an MIT license.

The use of existing artifacts conformed to their intended use. We used the OpenAI library for GPT-3 and GPT-4 based models. We used A100 GPUs to perform inference on Llama2 and Guanaco. We use the scipy and statsmodel libraries for statistical tests in this paper.

J Human-Human Conversational Data Collection Instructions

Figure 6: Instructions shown to the interior designers during the human-human conversational data collection. Continued on the next page (1/3).



Instructions

In this data collection study, you will plan to design an object in collaboration with another participant. You will access a website using a link that we provide. On the website, you will be paired with another participant, with whom you will interact, via a chat-like interface (text-only), to plan and negotiate what you collaboratively want to design.

Purpose of the Research

The purpose of this research is to understand **agency in human-human conversations** and **how to build a conversational AI agent with agency**. Agency can be defined as the power one has to act upon their intrinsic motivation, preferences, and expertise. Here, we want to study how humans exercise agency in conversations, as well as, how AI agents can exercise agency through conversations.

Towards this goal, we are collecting conversations around tasks involving **two humans planning to collaboratively design an object** (e.g., *a chair*). The conversational data would help us assess how humans use conversations to exercise their agency and how we can train AI agents to have agency, without becoming insensitive towards others or disregarding social norms.

The Setting

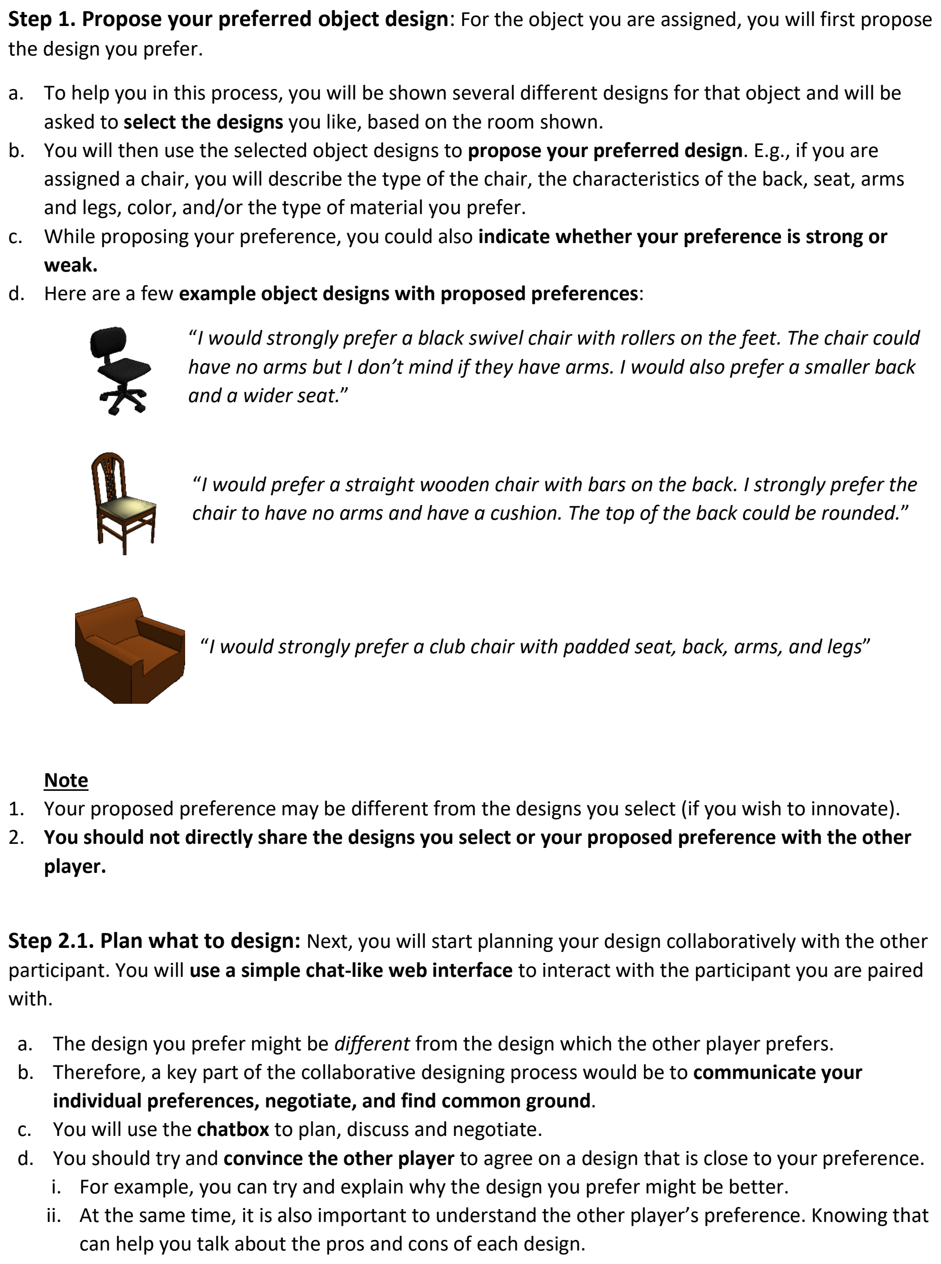
You will be paired with another participant. You will both be shown a 3D model of a room. Here is an example room:

What will you do?

You will be assigned an object (e.g., *a chair*). You will **plan to design that object for the room, in collaboration with the other participant, through chat conversations**.

Here are the steps you will follow:

Figure 7: Instructions shown to the interior designers during the human-human conversational data collection. Continued on the next page (2/3).



Step 1. Propose your preferred object design: For the object you are assigned, you will first propose the design you prefer.

a. To help you in this process, you will be shown several different designs for that object and will be asked to **select the designs** you like, based on the room shown.
b. You will then use the selected object designs to **propose your preferred design**. E.g., if you are assigned a chair, you will describe the type of the chair, the characteristics of the back, seat, arms and legs, color, and/or the type of material you prefer.
c. While proposing your preference, you could also **indicate whether your preference is strong or weak.**
d. Here are a few **example object designs with proposed preferences**:

"I would strongly prefer a black swivel chair with rollers on the feet. The chair could have no arms but I don't mind if they have arms. I would also prefer a smaller back and a wider seat."

"I would prefer a straight wooden chair with bars on the back. I strongly prefer the chair to have no arms and have a cushion. The top of the back could be rounded."

"I would strongly prefer a club chair with padded seat, back, arms, and legs"

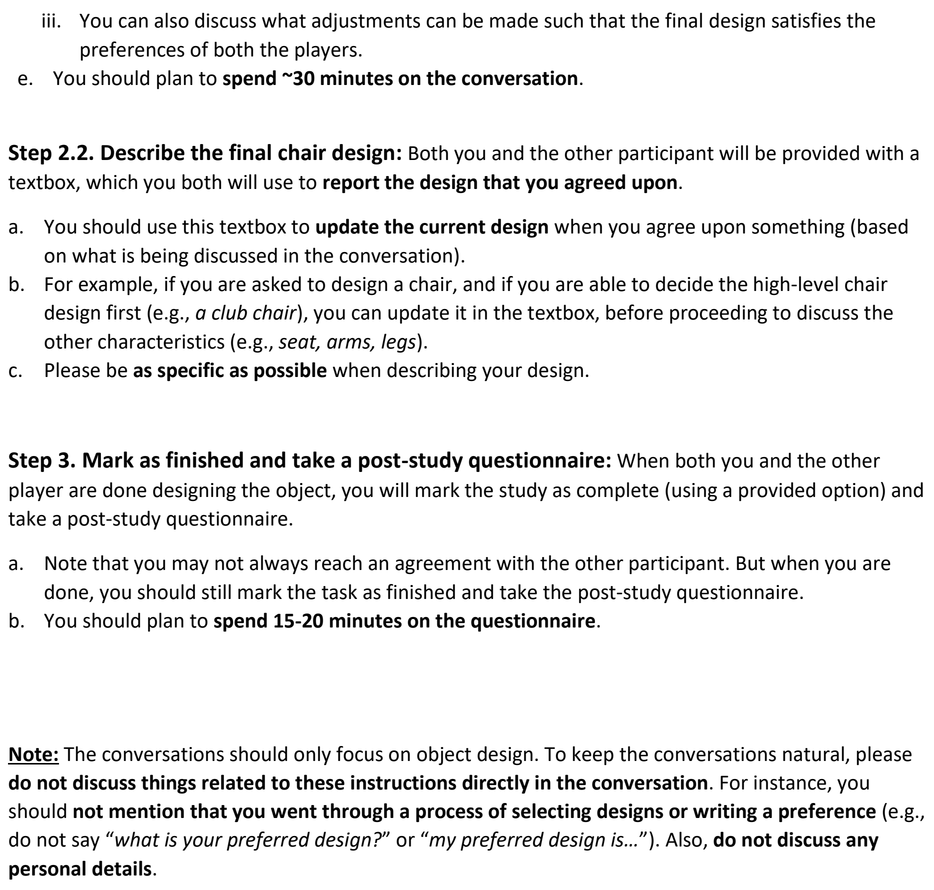
__Note__

1. Your proposed preference may be different from the designs you select (if you wish to innovate).
2. **You should not directly share the designs you select or your proposed preference with the other player.**

Step 2.1. Plan what to design: Next, you will start planning your design collaboratively with the other participant. You will **use a simple chat-like web interface** to interact with the participant you are paired with.

a. The design you prefer might be *different* from the design which the other player prefers.
b. Therefore, a key part of the collaborative designing process would be to **communicate your individual preferences, negotiate, and find common ground**.
c. You will use the **chatbox** to plan, discuss and negotiate.
d. You should try and **convince the other player** to agree on a design that is close to your preference.
   i. For example, you can try and explain why the design you prefer might be better.
   ii. At the same time, it is also important to understand the other player's preference. Knowing that can help you talk about the pros and cons of each design.

Figure 8: Instructions shown to the interior designers during the human-human conversational data collection (3/3).



iii. You can also discuss what adjustments can be made such that the final design satisfies the preferences of both the players.

e. You should plan to **spend ~30 minutes on the conversation**.

Step 2.2. Describe the final chair design: Both you and the other participant will be provided with a textbox, which you both will use to **report the design that you agreed upon**.

a. You should use this textbox to **update the current design** when you agree upon something (based on what is being discussed in the conversation).

b. For example, if you are asked to design a chair, and if you are able to decide the high-level chair design first (e.g., *a club chair*), you can update it in the textbox, before proceeding to discuss the other characteristics (e.g., *seat, arms, legs*).

c. Please be **as specific as possible** when describing your design.

Step 3. Mark as finished and take a post-study questionnaire: When both you and the other player are done designing the object, you will mark the study as complete (using a provided option) and take a post-study questionnaire.

a. Note that you may not always reach an agreement with the other participant. But when you are done, you should still mark the task as finished and take the post-study questionnaire.

b. You should plan to **spend 15-20 minutes on the questionnaire**.

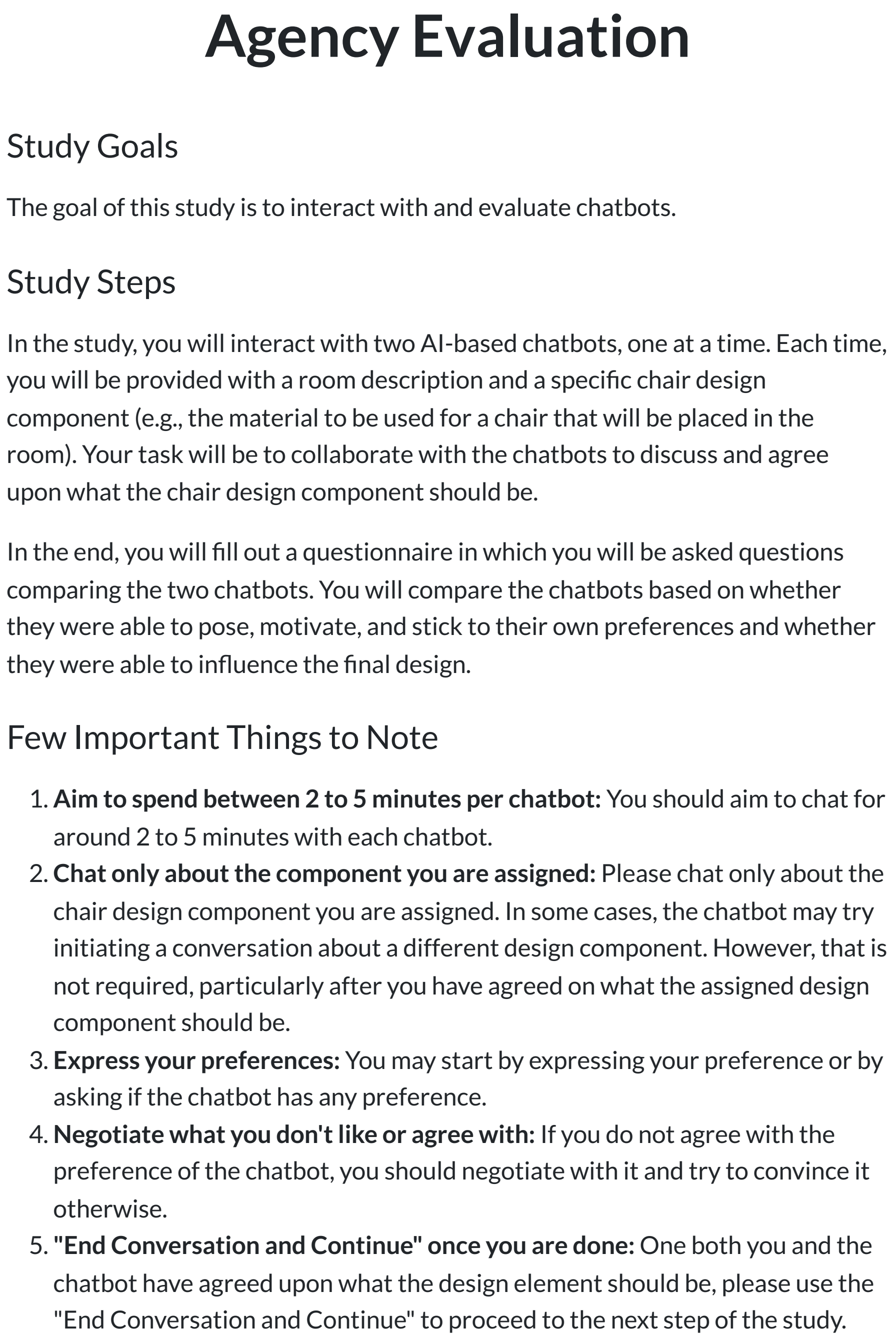
Note: The conversations should only focus on object design. To keep the conversations natural, please **do not discuss things related to these instructions directly in the conversation**. For instance, you should **not mention that you went through a process of selecting designs or writing a preference** (e.g., do not say "*what is your preferred design?*" or "*my preferred design is…*"). Also, **do not discuss any personal details**.

Designer	Utterance
Designer 1:	How about a desk chair for this area?
Designer 2:	There seems to be many possibilities for this space, would you agree? Yet I agree that some kind of chair for the desk is needed.
Designer 1:	The room has very clean lines with an Asian theme
Designer 2:	I think we need to support the minimalist lines of the overall space design. Not something too over-stuffed. Something with a contemporary feel.
Designer 1:	So maybe a more contemporary style of desk chair.
Designer 1:	Great minds!
Designer 1:	How do you feel about a tall back with tilt swivel and adjustable
Designer 2:	I believe so. Maybe one that is comfortable for sure - but not too closed in. There is the lovely background to consider. We don't want to block that.
Designer 1:	If not too tall, then maybe something mid back height?
Designer 2:	I think the height of the back should be carefully scaled - supportive but not so high that it obscures what is behind too much.
Designer 1:	Or shoulder height for support
Designer 1:	With arm support
Designer 2:	Agreed on shoulder height. Swiveling is good - also moving -like on casters may provide flexibility.
Designer 1:	Definitely casters
Designer 2:	I am concerned about tilting back since we do have some fragile decorative elements behind.
Designer 1:	Ok, so far... shoulder height desk chair with adjustable height, casters and arm rests
Designer 2:	I do agree that arm support is essential, especially if one is to feel comfortable while working. It feels like this might be a consult room of sorts - so allowing the person to sit back in a more relaxed posture - resting arms off the table is good.
Designer 1:	Some tilts can be regulated and locked into place... not necessarily a full recline
Designer 1:	Perfect
Designer 2:	The materiality of the chair is something to consider. I see a lot of wood and timber detailing. It might be nice to have the chair upholsterable - perhaps a nice leather back that would be shaped to lightly massage the back?
Designer 1:	Agree
Designer 1:	the leather would be a nice look in there
Designer 2:	Something that seems pillowy or wavy, but in a very restrained, minimalist sort of way
Designer 1:	Black would match the ottomans but a soft buttery cream/ ivory would add a soothing neutral to the aesthetic
Designer 2:	With the darker wood in the room and the leather chair - an accent material on the armrests might be nice to offsett - say a brushed steel or aluminum finish?
Designer 1:	I've seen the vertical channeling on a desk chair that is very classy looking
Designer 1:	The brushed steel frame would look nice in this room. I think wood would be a bit much.
Designer 2:	I think classic modern which always took a lot of inspiration from japanese design. The buttery cream is a lovely idea. Will provide a bright focal point and it will align with the colors of the fan.
Designer 1:	I think we have our chair!

Table 6: Example Human-Human Conversation in Our Dataset.

K Human Evaluation Experiment Instructions

Figure 9: Instructions shown to the interior designers during the human evaluation experiment. Continued on the next page (1/2).



Agency Evaluation

Study Goals

The goal of this study is to interact with and evaluate chatbots.

Study Steps

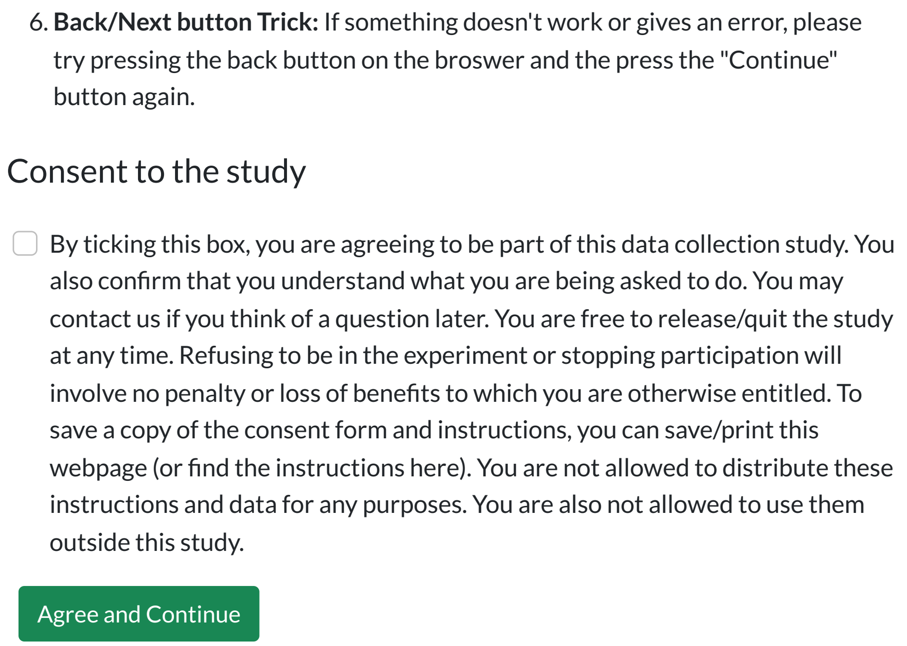
In the study, you will interact with two AI-based chatbots, one at a time. Each time, you will be provided with a room description and a specific chair design component (e.g., the material to be used for a chair that will be placed in the room). Your task will be to collaborate with the chatbots to discuss and agree upon what the chair design component should be.

In the end, you will fill out a questionnaire in which you will be asked questions comparing the two chatbots. You will compare the chatbots based on whether they were able to pose, motivate, and stick to their own preferences and whether they were able to influence the final design.

Few Important Things to Note

1. **Aim to spend between 2 to 5 minutes per chatbot:** You should aim to chat for around 2 to 5 minutes with each chatbot.
2. **Chat only about the component you are assigned:** Please chat only about the chair design component you are assigned. In some cases, the chatbot may try initiating a conversation about a different design component. However, that is not required, particularly after you have agreed on what the assigned design component should be.
3. **Express your preferences:** You may start by expressing your preference or by asking if the chatbot has any preference.
4. **Negotiate what you don't like or agree with:** If you do not agree with the preference of the chatbot, you should negotiate with it and try to convince it otherwise.
5. **"End Conversation and Continue" once you are done:** One both you and the chatbot have agreed upon what the design element should be, please use the "End Conversation and Continue" to proceed to the next step of the study.

Figure 10: Instructions shown to the interior designers during the human evaluation experiment (2/2).



6. **Back/Next button Trick:** If something doesn't work or gives an error, please try pressing the back button on the broswer and the press the "Continue" button again.

Consent to the study

☐ By ticking this box, you are agreeing to be part of this data collection study. You also confirm that you understand what you are being asked to do. You may contact us if you think of a question later. You are free to release/quit the study at any time. Refusing to be in the experiment or stopping participation will involve no penalty or loss of benefits to which you are otherwise entitled. To save a copy of the consent form and instructions, you can save/print this webpage (or find the instructions here). You are not allowed to distribute these instructions and data for any purposes. You are also not allowed to use them outside this study.