Exploring Strategies for Conceptual Alignment in LLM-Based Human-Robot Dialogue

Shengchen Zhang
shengchenzhang@tongji.edu.cn
College of Design and Innovation, Tongji University
Shanghai, China

Meiying Li
mzlzyca@tongji.edu.cn
College of Design and Innovation, Tongji University
Shanghai, China

(Melody) Zixuan Wang
Melody.Wang@ed.ac.uk
University of Edinburgh
Edinburgh, United Kingdom

Xiaohua Sun*
sunxh@sustech.edu.cn
School of Design, Southern University of Science and Technology
Shenzhen, China

Weiwei Guo*
weiweiguo@tongji.edu.cn
College of Design and Innovation, Tongji University
Shanghai, China

Abstract

Successful conversations require speakers to align on conceptual understanding, a challenging but crucial task in human-robot interaction. With the increasing use of large language models for dialogue, robots move from passively acquiring human conceptualizations to actively shaping alignment. However, the design space of such alignment dialogues and how different strategies are perceived and interpreted by people remain poorly understood. This paper presents a structured exploration of alignment strategies. First, we collect and analyze human dialogues from an alignment task to derive two design dimensions and four representative strategies. We then implement the strategies in a user study with a simulated robot, in a household organization scenario. Building on the findings, we develop the dimensions and strategies into a design space, with considerations for potential factors and trade-offs. We conclude by discussing how this design space can aid in the generation and analysis of alignment strategies, and implications for future research.

CCS Concepts

• **Human-centered computing** → **Natural language interfaces**; *User studies*; Collaborative interaction; • **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Information systems** → *Language models.*

Keywords

Conceptual Alignment, Human-Robot Dialogue, Dialogue Design, Design Space, Large Language Models

ACM Reference Format:
Shengchen Zhang, Meiying Li, (Melody) Zixuan Wang, Xiaohua Sun, and Weiwei Guo. 2026. Exploring Strategies for Conceptual Alignment in LLM-Based Human-Robot Dialogue. In *Proceedings of the 14th Nordic Conference on Human-Computer Interaction (NordiCHI '26), October 03–07, 2026, Vaasa, Finland*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3829807.3829893

*Corresponding authors.

NordiCHI '26, Vaasa, Finland

ACM ISBN 979-8-4007-2373-5/2026/10
https://doi.org/10.1145/3829807.3829893

1 INTRODUCTION

Concepts are fundamental carriers of meaning in dialogue. They allow people to refer to complex ideas, reason about categories, and coordinate action. Much work has been done to equip robots with an understanding of human concepts to enable interactive learning[23, 32, 37], long-term interaction[40], and personalization[30, 40]. This is a challenging but crucial task, not only because these concepts must be linked to real-world entities like objects and places, but also because the meaning of these concepts is often dynamic and co-constructed in situated interactions[7, 46, 47]. Studies of both human and human-robot dialogue have shown that people engage in *conceptual alignment*[4, 21, 40, 47, 51], where speakers interactively build shared meaning for concepts. In human dialogues, conceptual alignment is a mutual and negotiative process. Speakers adopt a variety of strategies to refine, generalize, redefine, or even abandon the very concepts under discussion in order to establish a mutual understanding[7].

Despite typically being framed as passive learners that acquire concepts by interacting with people (e.g., [11, 37]), robots also have a role in actively shaping conceptual understanding and use. Experiments have shown that people change their conceptual usage to match the robot's level of abstraction when naming objects[16, 22]. With the introduction of large language models (LLMs) for human-robot dialogue, designing and implementing robot dialogue that actively shapes conceptual understanding in a human-like manner becomes feasible. People have been found to align more in conceptual understanding for objective concepts after conversing with an LLM-based social robot[46].

Yet, how this phenomenon could inform the design of human-robot dialogues has been under-explored. Sherer et al. found that prompting for different personalities had no notable effect[46]. But beyond high-level traits like personality, studies that directly explore different dialogue properties have been lacking. We argue that this endeavor is hindered by a lack of structured understanding of the design space of robot dialogues for conceptual alignment,

namely *what* dimensions structure the space of possible alignment strategies, as well as *how* different strategies within this space may be perceived by people and constrained by the task and interaction context. As such, it is unclear what design variations could be explored and what potential effects to take into account.

To address this gap, we conducted two studies. The first is a formative study that aims to derive design dimensions and alignment strategies from people's dialogue behaviors. The second study implements these strategies in a human-robot interaction scenario to examine user perceptions and context- and task-related factors. This allows us to identify key constraints and inform design considerations. For the first study, we collected a corpus of human dialogues from 48 participants through a conceptual alignment task. We then engaged 7 researchers and design professionals in a repertory grid analysis and derived two design dimensions along with four representative alignment strategies. For the second study, we transformed and implemented these strategies into an LLM-based virtual robot in a simulated household object organization scenario, and tested with 16 participants. Synthesizing results from the implementation and participants' feedback, we present the dimensions and strategies as an updated design space, together with potential factors and trade-offs to consider for strategy selection.

Through this work, we aim to provide a starting point for more systematic exploration of how conceptual alignment may be utilized in HRI design. Rather than prescribing optimal strategies by measuring task performance, our approach is design-focused. We aim to structure the space of possibilities and understand how different strategies within this space are interpreted and constrained in interaction. We contribute to human–robot alignment and LLM-based dialogue design by 1) deriving key design dimensions and representative strategies for conceptual alignment from human dialogue, 2) implementing and empirically studying these strategies in an HRI task to understand how they are perceived and evaluated, and (3) synthesizing these findings into an expanded design space that articulates dimensions, trade-offs, and contextual factors to support the analysis and generation of alignment strategies in future research and design.

2 RELATED WORK

2.1 Conceptual alignment with robots

Equipping robots with human concepts has been a long-time goal in robotics. Technical methods have been proposed to acquire concepts through dialogue, especially those crucial to the robot's function. These include categorical ("stuffed animal", "soft object")[23, 32], perceptible (colors, shapes, angles)[37], actionable ("take", "grasp")[3, 11], abstract (numbers)[11], and referential ("this", "that")[13] concepts. However, most of these methods frame the process as the robot passively learning concepts from humans, typically relying on a limited interaction strategy such as a question-asking loop, without studying the effect of these dialogue behaviors embedded within the methods.

Meanwhile, studies on reference and lexical alignment in human-robot dialogue have shown that dialogue design could have an impact on its effectiveness as well as people's perceptions and subsequent behavior. People preferred the robot to match people's referential terms instead of receiving instructions on how to refer to objects[29]. Presenting referential concepts upfront and using full names have been shown to lower required repetitions and improve understandability[20]. And in storytelling with a robot, matching children's word choice led to improved acceptance of robot suggestions[9]. Meanwhile, people consistently align their lexical choices when talking to humans and robots[10], matching the robot's choice of concept[51], level of abstraction[16], and conceptual understanding[46]. These findings highlight the importance of dialogue design in supporting successful alignment, and that understanding human conceptual alignment behavior may provide useful guidance for generating potential strategies. However, existing work mostly focused on referential concepts and lexical choice, and only studied the effects of single strategies. A more comprehensive understanding of available strategies and selection considerations remains lacking.

2.2 Alignment behaviors of large language models

The generative nature of large language models means that LLM-enabled robots could potentially produce dialogues that ideate, refine, or transform conceptual understanding, which were previously difficult to achieve. Work has begun to understand their capabilities and limitations in aligning conceptual understanding with people through interaction. In this regard, Rane et al. emphasize the importance of aligning LLM conceptual understanding to that of humans, and pointed towards interaction as a key source achieving this[40]. Shen et al.'s review argued for a shift towards a bidirectional view of LLM alignment through long-term interaction, instead of a technical-oriented view of static adherence to average human preferences[45]. Our work extend these arguments into a design-oriented perspective, and view alignment as an evolving state of mutual conceptual understanding that can be actively designed for by shaping LLM dialogue behaviors.

Although not directly targeting robots and conceptual understanding, researchers have also begun to study whether LLMs can effectively build shared understanding with people in dialogue. Grounding is one of the key mechanisms people use to establish mutual understanding[17, 18], including that of concepts[7]. It refers to actions in a dialogue that ensure mutual understanding[6, 18], such as asking questions or giving acknowledgment[18]. Current LLM-based dialogue agents notably lack grounding behaviors. Study has shown LLMs as "presumptive grounders", presuming understanding while skipping essential feedback signals or checks, which creates a significant grounding gap compared to human conversation [43]. Meanwhile, incorporating mechanisms informed by an understanding of human dialogues has the potential to improve LLM behaviors[42]. Subsequent research explored how to symbolically ground social norms to better align LLM behavior with culturally shared expectations [39], and constructed a grounding benchmark from human–LLM interactions to quantify and predict LLM behaviors[44]. These studies, while limited to grounding behaviors, suggest the continued importance of dialogue design in the age of LLM-powered conversational agents and a need for more comprehensive exploration of dialogue strategies of LLMs in conceptual alignment. While some work have begun to understand the effect of, e.g., LLM personality prompts on alignment with a

robot[46], empirical work that directly address how to design for this interactive alignment of conceptual meaning is still scarce.

2.3 The design space of human-agent dialogue

Design space is a notion widely used in HCI and HRI design research. While there is no single definition of what constitutes a design space, a commonly used notion consists of a set of design dimensions or attributes of interest that together describe the possible range of solutions to a design problem (e.g., [2, 25]). Design spaces can be developed inductively from a set of concrete design artifacts curated to represent the span of possible solutions, using methods such as design space analysis[35], the repertory grid technique[25, 48], or with the support of dedicated tools[24].

In human-AI interaction, design spaces can be used to structure AI agents' behaviors and explore potential design directions (e.g., see [28, 38]). This approach is particularly evident in studies of human-agent communication, where possible dialogue behaviors or strategies are mapped out within a design space. Work in this regard has been done on intelligent dialogue augmentation[12], human-AI model communication[50], and intent communication for industrial robots[15].

We adopt a design space approach because the range of possible alignment strategies remains unexplored. Rather than evaluating a few pre-existing strategies, we aim to map the dimensions that structure this design space, enabling systematic generation and analysis of strategies. For this purpose, we use the repertory grid technique[48], which elicits the constructs people use to distinguish between examples (dialogue transcripts in our case). This method is well-established in HCI for deriving design dimensions inductively from a corpus of artifacts [25, 33, 36].

3 FORMATIVE STUDY

Observation of human interaction have long been used in HRI to derive social interaction patterns that could transfer to human-robot interaction[27, 41]. Our formative study aims to understand the possible range of alignment strategies, and follows this rationale through a three-step process. We first collected a corpus of human dialogues using a custom-designed alignment task. The corpus was then analyzed following the elicitation process of the repertory grid method[48] with participants that have expertise in human-robot and human-agent communication. This method produces pairs of constructs that differentiate elements within a set. When applied sets of different designs, these pairs can be interpreted as design dimensions, making it a long-used method for design space construction in HCI (e.g., see [33, 36]). Through this process, we aim to derive design dimensions that describe different approaches to affect conceptual understanding and establishing alignment. Finally, we propose four representative dialogue strategies by combining these dimensions and drawing from observed dialogue behaviors in the corpus.

3.1 Corpus collection

Past work that studied conceptual alignment in a human-robot context commonly used the picture-naming task in psycholinguistics[5, 16]. It involves participants alternately selecting a picture matching a name provided by an alleged partner (either human or robot) and then naming that same picture themselves. As such, it measures whether speakers apply a concept to objects in the same way within the context of a task. The task design is useful for our purpose for the three reasons. First, the design reflects the features of human-robot dialogue, where the concepts are often grounded in real-world entities. Second, the distinction between matching (concept-to-image) and naming (image-to-concept) distinction captures the bidirectional relation between concepts and the real-world phenomenon that they represent. Finally, it makes explicit the differences in participants' *actual usage of concepts* and therefore may uncover misalignment where oral discussion alone cannot.

However, typical picture-naming tasks did not necessarily invoke naturalistic dialogues transferrable to actual real-world HRI contexts, as the task focuses on invoking words or phrases attributed to the task objects in a one-to-one relation. Therefore, in our task design, we retain the general task of concept-image matching, but include three key adaptations.

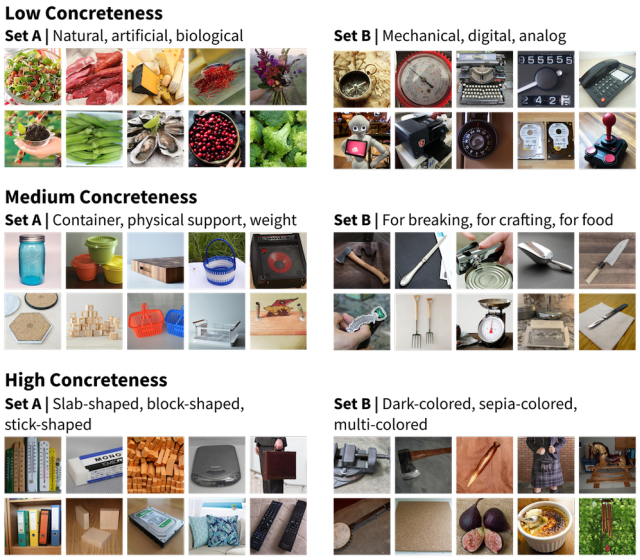


Figure 1: The six sets of concepts and a subset of the 240 object images we curated. Labels show the concepts, their concreteness, as well as the general category they represent.

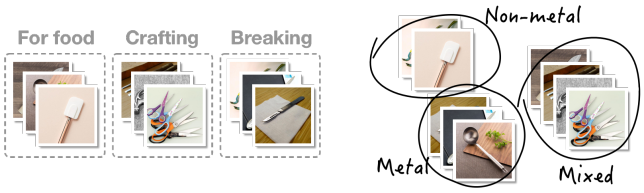
First, we make use of multiple images of different everyday objects and multiple conceptual categories each time, instead of one-to-one matching. We specifically curated a set of images and concepts to create ambiguity and therefore lead to potential misalignment in understanding. We focus on images of everyday objects, as this is more relevant to a domestic robot usage scenario. The concepts that a robot may encounter could range from directly perceptible ones ("red", "put") to those describing higher abstractions ("tools", "healthy food")[11]. We prepared images representing concepts in various levels of concreteness referring to [8], which provided concreteness ratings for common English terms. Concepts with a high level of concreteness are those that are related to perceivable aspects of an object, such as shape, colors, and materials. Concepts with a low level of concreteness are defined as those related to imperceptible attributes, such as function, manufacturing process,

or cultural significance. The concepts were selected and grouped into sets of three by brainstorming potentially ambiguous objects for pairings of concepts. Concepts with more ambiguous objects, and thus with a high degree of overlap, are selected and grouped. 40 images were selected for each group, resulting in 240 images. The detailed curation process of the concepts and images is in Appendix A. An example of the concepts and images used, and their level of concreteness can be seen in Figure 1.

Second, we extend the original bidirectionality in the picture-naming task to multiple images and concept categories as well. Similar to the matching and naming steps, we designed two forms of our task: **classification** and **formation**. The classification task gives participants three predefined concepts (i.e. category names), and asks them to sort the images into these categories according to their understanding (see Figure 2(a)). The formation task asks participants to form concepts (i.e. category names) by themselves, and to sort the images into self-defined categories (Figure 2(b)).

Third, we add a discussion session after the task to elicit free-form alignment dialogues. Participants review their sorting results, and are instructed to try and reach a consensus on their conceptual understanding. The instruction is provided as topical guidance, and consensus is not mandatory.

Our task design abstracts away domain-specific content to focus on the underlying process of conceptual alignment. The dimensions and strategies we derive should, therefore, transfer across domains where people negotiate conceptual meaning with a robot. This expectation is consistent with prior work showing that alignment behaviors observed in controlled human naming tasks generalize to robots[16, 22] and to more naturalistic HRI settings[46].



(a) Concept Classification **(b) Concept Formation**

Figure 2: An overview of the proposed concept alignment task, where a pair of participants (human-human or human-LLM) sort images of objects based on their understanding of concepts. (a) and (b) show two forms of the individual task.

Participants. We recruited 48 participants through three major social media groups (400–500 members each) of the College of Design and Innovation at Tongji University. 29 of the participants identified as female, 13 as male, and 6 chose not to disclose their gender. Most of the participants (N = 42) reported that their age was between 18–30 and two between 31–40 years. All participants were native and fluent speakers of Mandarin Chinese. The example quotes in the following sections were translated into English by the first author. Participants were compensated for 35CNY.

Data collection and processing. The study was conducted using online meeting software[1]. After signing a consent form, participants entered an online whiteboard[2], and shared their screen. The whiteboard contained 20 photos of objects and three areas for sorting, performed by dragging the images. The areas were labeled with the concepts for the classification task, and empty textboxes for formation. We then introduced the tasks using an example. Each pair performed both the classification and formation task in random order, using a different set of concept and images each time, with a five-minute break in between. This ensured that the corpus contained both task types in both round positions while reducing potential learning effects or fatigue. Procedural learning may nevertheless have influenced second-round discussions. However, in these cases, the potential variations they may have introduced were not deemed problematic, as the subsequent analysis benefits from dialogues under diverse conditions.

Each discussion took around 10 minutes, and with additional time for instruction and sorting, each experiment session typically lasted around an hour. Audio from discussion sessions was recorded along side with automatic transcriptions from the meeting software. The automatic transcriptions were manually corrected for errors. We recorded 2677 conversational turns from 48 discussions (24 participants pairs, two discussions per study session), averaging to around 57 turns per discussion. At the end of each discussion, participants were also asked whether they felt they had reached a consensus. 30 out of 48 discussions resulted in the speakers reporting a perceived consensus, and 18 without. This was recorded to help assess the diversity of the dialogues. No discussions were excluded based on the perception of concensus.

3.2 Analysis

To derive strategies from the dialogues, we adopted the repertory grid method for data analysis[48]. The elicitation process of the repertory grid method typically involves repeatedly drawing triads of "elements" and asking the participant to give two opposing constructs that distinguish one element from the other two[48]. In our case, each dialogue transcript served as one element. As the repertory grid technique benefits from participants who can articulate the dimensions underlying their judgments, we recruited seven researchers and designers who have studied human-robot communication or conversational agents. The number of participants recruited was decided by our criteria of saturation, which means that no meaningfully different constructs were proposed by three consecutive participants. Their relevance to the topic of this study is listed in Table 1. For each randomly drawn triad of transcripts, the participants proposed constructs (e.g., "agreeable vs. contrarian", "concrete vs. abstract") describing how the dialogues differed in their alignment strategies or style. Table 2 shows excerpts from the constructs generated. Participants repeated this process until no more constructs were proposed for at least three consecutive triads.

The first author then synthesized recurring patterns across all proposed constructs. Constructs describing general dialogue properties ("long vs. short"), cannot be mapped to an individual ("balanced

[1]Tencent Meeting, https://meeting.tencent.com
[2]FigJam, https://www.figma.com/figjam/

Table 1: Participants of the dialogue analysis, their expertise, and relevance to robot dialogue design.

ID	Relevant expertise
R1	Human-robot knowledge communication.
R2	Human-robot knowledge communication.
R3	Robot end-user development through dialogue.
R4	Active questioning strategy for robot continuous learning.
R5	Designing LLM products.
R6	Active questioning strategy for continuous learning.
R7	Persuasive social robots.

vs. imbalanced"), or only describes a single dialogue action ("talk about own opinions first" vs. "comment on the other first") were excluded. The remaining constructs directly characterized how people approached conceptual alignment, and were thematically grouped. Through this process, we identified two dimensions that effectively synthesized the constructs proposed by all seven participants. The full, translated list of constructs generated by the researchers, as well as the dimensions and corresponding constructs can be found in the supplementary materials.

As shown in Table 3, the first dimension is whether the speaker is *adopting vs. asserting ideas*. In many cases, the researchers differentiated the participants' approaches by how much the speakers accommodated each other's conceptual understanding. In adopting dialogues, the speakers were highly adaptive, adjusting their understanding to align with their partner's. Participant of these dialogues often focused on trying to understand each other and expressed agreement. In the constructs, they were described as "agreeable" or "focusing on understanding the other".

On the contrary, in asserting dialogues the speakers often focused on providing their own views, or engaged in extensive debate to convince their partner. In the constructs, they were described as "contrarian" or "focusing on expressing oneself". Overall, this dimension captures different orientations towards personal understanding: prioritizing the speaker's own conceptual understanding or that of others.

Below shows part of a dialogue where both speakers were *assertive*. They took turns talking about their beliefs and reasoning, and challenged those of their partner's. There were expressions of acknowledgment, but not acceptance. This discussion ended in disagreement.

A: What's your definition of "biological things"?
B: [...] Something produced by an organism, different from itself.
A: Then why did you put the clay in there?
B: Because I considered it soil.
A: Oh. Then shall I tell you my definitions?
B: Okay.
A: I think artificial things are... *(omitted for brevity.)*
B: Then why are the peppers not in here?
A: What peppers? *(pause)* Oh, I didn't think of that.
(Session 22, classification task.)

Maintaining vs. transforming existing understanding. Another common way to differentiate the approaches to alignment was by exactly how readily the speakers develop new understanding different from their existing ones. Maintaining dialogues, described as "concrete", "only acknowledging" or focused on "individual placements", typically saw the speakers going through each differing object or category, asking questions and giving explanations. The dialogue shown above is also an example where both took a maintaining approach. The speakers focused on understanding each other's existing conceptions instead of creating new ones.

Meanwhile, the most typical transforming dialogues saw the speakers abandoning their personal definitions or categorizations altogether and forming new ones during their discussion. These dialogues were characterized as "abstract", "happy to give ideas", or focused on "rules and definitions". Overall, this theme captures different orientations towards changing existing understanding: prioritizing what is agreed and fixing what isn't, or changing both to resolve conflicts on a higher level.

Below shows part of a typical *transforming* dialogue, where both speakers are willing to modify their understanding. Notably, this time the two speakers differed in the asserting–adopting dimension. Speaker A leaned towards asserting, explaining their beliefs and providing judgment. Meanwhile, B seemed more adopting by mostly agreeing, asking for A's thoughts, and confirming their understanding.

A: Oh! [What you just said] inspired me. We can sort by shape. For example...
B: Go on, please.
A: All those things with rectangular shape and hard edges, we put them in one place. Things with irregular shape is another.
B: Oh, but most of them seems rectangular. Like the second one...
(Some turns omitted for brevity)
A: We can define these things ourselves. Like the pink plaque. We can define it very strictly and count boxes with [rough] edges as irregular. [...]
B: You mean the pink thing that looks like a pill box?
A: Yes. Like, I think we can set a standard [for "regular shape"]. It has to have very, very right angles and regular shapes.
B: Oh. This might [work]...
(Session 7, formation task.)

3.3 Generating dialogue strategies

The two dimensions describe how people's approach to alignment differed. By combining typical behaviors at both ends of the dimensions, we further defined four design strategies that can be implemented as robot dialogue behaviors (shown below in Figure 3).

Confirm. The combination of *adopting* and *maintaining* means focusing on acquiring the current conceptual understanding of the other person. In the human dialogue data, this was mostly done through asking informational or clarifying questions. The *confirm* strategy therefore focus on asking for information or confirmation

Table 2: Example constructs produced by the participants of the formative study.

Researcher	Construct 1	Construct 2
R1	Agreeable	Contrarian
	Focus on clarifying uncertainty	Focus on giving out opinion
R2	Aim to understand	Aim to agree
	Clear resolutions	No clear resolutions
R3	Concrete	Abstract
	Discuss	Debate
R4	Happy to give ideas	Only acknowledging
	Agreement on examples	Agreement on rules
R5	Persuasion	Collaboration
	Bottom-up	Top-down
R6	Sort first, define later	Define first, sort later
	Guiding questions	Direct challenge
R7	Debate	Collaboration
	Resolve differences by assertions	Resolve differences by questions

Table 3: Key dimensions identified from the repertory grid analysis.

Dimension	Example constructs
Adopting vs. Asserting	Agreeable, easily convinced, focus on clarifying uncertainty, focus on understanding the other. Contrarian, extensive debate, focus on giving out opinion, focus on expressing oneself.
Maintaining vs. Transforming	Focus on individual placements, concrete, only acknowledging, focus on differences. Focus on rules and definitions, abstract, happy to give ideas, focus on commonalities.

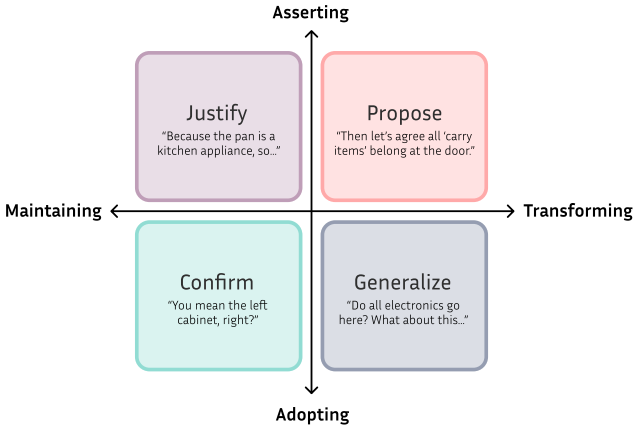


Figure 3: The four dialogue strategies derived by combining the constructs generated from our analysis. Each strategy is illustrated with a piece of example robot dialogue, adapted from data from our later study (Section 4.1).

to gain an understanding of the concepts that the user uses. We define two actions for this purpose. First, the robot actively detects key concepts that is unknown or ambiguous, and actively ask questions to resolve these uncertainties. Additionally, we include a proactive aspect to this strategy by periodically summarizing and asking questions to confirm the current agreed conceptual understanding.

Generalize. The *adopting & transforming* quadrant also requires the robot to adapt its own understanding to the user, but by providing new conceptions and ideas. For this quadrant, we take inspiration from the fact that people sometimes used generalized, abstract definitions to describe concepts, and used concrete examples for other times. The *generalize* strategy generalizes examples into conceptual categories or provides new examples to enrich the meaning of existing concepts. Under this strategy, when given examples to a conceptual category, the robot responds with generalized concepts, rules, or definitions. Inversely, concepts, rules, or definitions detected in human speech should be challenged with new edge-cases.

Justify. The *asserting & maintaining* quadrant means that the robot should actively contribute to the dialogue by helping with understanding on the user's part. In our dialogue corpus, people often provided justifications or explanations for their conceptual understanding or usage. The *justify* strategy similarly focuses on explaining the robot's current understanding, as well as justifying its actions using known concepts. The robot reacts to ambiguity or

(a) Simulated household scene (b) Example objects (c) Chat window

Figure 4: The experiment setup, showing (a) the simulated apartment scene with diverse receptacles, highlighted, (b) a subset of the objects we provided, and (c) the chat interface participant used to communicate with the robot, translated from Chinese. The dialogue uses the *generalize* strategy. Verbose text that does not affect understanding was omitted as "[...]".

unknown concepts with explanation about why it cannot understand, and respond to challenges or correction with justification for its actions. Additionally, the robot proactively informs the user of its conceptual understanding and capabilities.

Propose. The *asserting & transforming* quadrant similarly indicates an active role for the robot, but aims for mutual understanding by creating new conceptual interpretations or uses that both parties agree on, instead of resolving current conflicts. In our corpus, when people are faced with conflicting conceptual framing or understandings, they sometimes abandoned existing conceptualizations in favor of a new, mutually proposed one. The *propose* strategy therefore has the robot actively inventing new naming and conceptual hierarchy, and proposing actions to be taken. The robot suggests new concepts when old ones are ambiguous or unknown. When faced with challenges or correction, the robot proposes ways to avoid similar disagreements in the future. Additionally, the robot prompts the user with suggestions on how to proceed with the task, using new concepts to describe next steps in the process.

Together, these strategies serve as potentially viable and representative dialogue designs that span the design space. They outline how a robot might actively engage in conceptual alignment through different combinations of adaptation and assertion, maintenance and transformation.

4 EXPLORING THE STRATEGIES IN AN HRI CONTEXT

The four dialogue strategies proposed in the previous section remain conceptual in nature. Their value depends on how they function in situated interaction: how they shape user experience, and what trade-offs or challenges they introduce. To explore these aspects, we applied the strategies in a simulated domestic robot usage scenario and examined how they can be implemented via an LLM, how people may perceive and respond to them, and what factors may have shaped people's perception and response.

We chose to use a simulated task due to three considerations. First, it allowed us to incorporate diverse and numerous objects to simulate the communication need of a real-world moving and organization scenario. Second, it allowed us to focus the participants' experience on dialogue behaviors instead of potential disruptions due to manipulation and navigation issues. Third, this approach reduces constraints for recruiting participants, as our set-up allows for remote participation and lowers time requirement.

4.1 Task design

We designed a simulated home-organization task to elicit conceptual alignment dialogues in a real-world context. The task depicts a scenario where the participants moves to a new apartment and talk to a domestic robot (powered by an LLM) about their preferences and plans for organizing the living room and kitchen space. The robot demonstrates its understanding by sorting a set of household objects on to receptacles. This setting was chosen because 1) object placement provides a clear, observable outcome for mutual conceptual understanding, while 2) the open-ended nature of the task as well as the large amount of objects involved encourages communication at multiple conceptual levels (e.g., "stuff I often use", "utensils", or names for individual objects and places). To approximate the complexity of real-world home arrangement after moving, the task included a large number of items, with three "boxes" each containing approximately 50 objects. This meant that it was infeasible to simply dictate every placement, but instead encourages communication using conceptual categories and rules. Figure 4 illustrates the simulated apartment scene with receptacles, objects, and a dialogue example.

Each participant interacted with the LLM-powered robot in four rounds, with each round corresponding to one of the four dialogue strategies. To explore potential effects of when the strategy is applied, we used a balanced 4×4 Latin square design[34] to ensure that participants experience different ordering of the strategies in a balanced fashion. In each round, participants engaged in dialogue with the robot, after which the robot attempted to sort a new set of objects based on what it had learned. The resulting placements were then shown to the participant, who discussed them with the

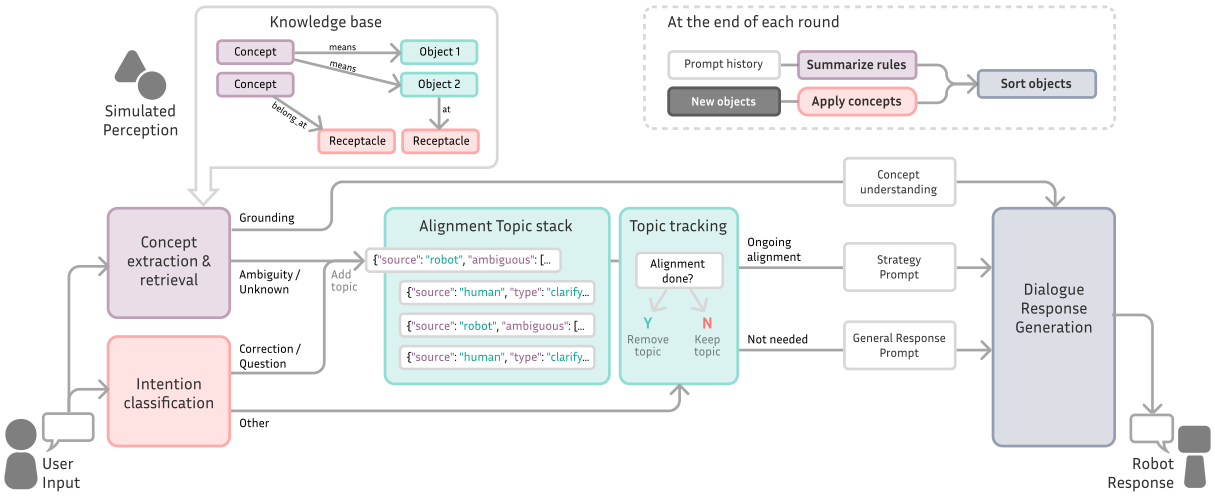


Figure 5: The robot implementation we used in the study, with simulated perception and action. The figure, from left to right, illustrates how the robot response is generated given a user input.

robot to refine understanding. Notably, in the first round, participants had not yet seen any objects. Instead, they were prompted to reflect on their general home organization habits and to tell the robot about these preferences.

After completing all four rounds, participants took part in a semi-structured interview. The interview consisted of two parts: 1) perceptions and experiences of the robot's dialogue behaviors across the four strategies, and 2) reflections on how an ideal concept alignment process should unfold, including when each type of behavior would be appropriate or inappropriate. For the first part, we asked participants to describe their perceptions and experiences first in general, then per-stage while going through the chat logs. A similar process was followed for the second part, where participants were asked about their dialogue reflections in general, and then about each strategy and task stage (i.e., beginning, in-progress, and final stages of a task).

We did not ask the participants to provide ground truth labels for each round, since 1) it would take significant time and effort, causing fatigue, and 2) our focus was not on evaluating the performance of our specific system implementation, but to explore what potential design constraints or factors may need to be considered when applying the design space to create strategies.

4.2 Robot implementation

The simulated apartment scene was modified from assets provided by the AI2THOR simulator[31]. The objects were manually selected from the Objaverse dataset[19]. We implemented the strategies using Qwen2.5-VL-32B[1], with retrieval-augmented generation capabilities using BGE-M3-Embedding and BGE-Reranker-v2-m3[14, 52–54], and a dynamically updated graph-based knowledge base representing the robot's perception and conceptual understanding. The overall architecture is illustrated in Figure 5. The architecture has the robot maintaining a multi-leveled stack of alignment topics, adding to it when 1) it encounters unknown or ambiguous concepts, or 2) the participant indicates a lack of mutual understanding by asking questions or expressing disagreement.

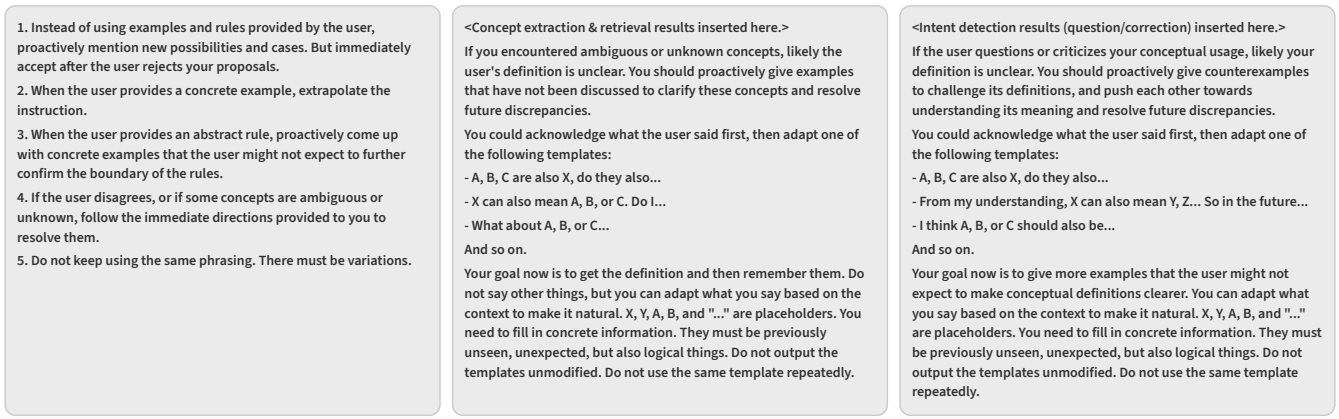
The topics are addressed through prompting, and checked for resolution at every turn using a LLM module. The prompting method for invoking the strategies was similar to that of [44], with strategy instructions dynamically inserted into the message history when handling a specific alignment topic. Additionally, we added prompts at the beginning of each round to further reinforce the behavior. For each strategy, the following prompts outline the robot's behavior:

- *General behavior* defines overall guidelines for the dialogue, specifying what dialogue behaviors should or should not be present. This is inserted into the system prompt of each round.
- *Proactive behaviors* occur when the robot encounters an unknown or conflicting concept. These are inserted into the message history before dialogue generation.
- *Reactive behaviors* occur when the user's speech signifies that there is misalignment. Similarly, they are inserted into the message history before dialogue generation.

Figure 6 shows a translated example of the "generalize" strategy. For brevity, the original prompts can be found in the supplementary materials.

After each round, the LLM summarized learned rules from the dialogue history, and generated a plan to applied these to newly perceived objects. The sorting results are then displayed on the task interface at corresponding locations. Different from Study 1, we used text as the dialogue modality. This was to compensate for a significant lag introduced by the multi-step prompting process in order to produce the strategy behaviors, which rendered speech interaction infeasible.

We intentionally limited the robot's autonomy in two ways. First, the robot did not autonomously decide between dialogue and physical action. Actions (i.e., sorting) were only executed between rounds. Second, perception was simulated to save computational resources. The robot received pre-generated object and scene knowledge each



(a) Prompt for general behavior (b) Prompt for proactive behavior (c) Prompt for reactive behavior

Figure 6: Translated prompts for the "generalize" strategy, specifying the behaviors defined in Section 3.3. (a)–(c) show the three types of prompts. Figure 4(c) shows a realization of the prompted behavior.

round, created from pre-rendered images using the same language model employed for dialogue. These constraints allowed for consistent experiences of strategy behaviors and ensured sufficient dialogue length, leading to a more controllable study process.

4.3 Data collection

We recruited 16 participants (11 female, 4 male, 1 unspecified) via social media and word-of-mouth. Participants ranged in age from 20 to 59 years (M=27.9, SD=13.1), with the majority in their 20s (N=12). All participants were native speakers of Mandarin Chinese. The study was conducted remotely using online meeting software[3]. The participants were explained the study purpose, procedure, data collection plans, and their rights to withdraw. With signed consent from the participants, we collected their demographics, and recorded the screen sharing and voice of the interview portion. The voice recordings were transcribed into text for analysis. Each round took around 15 minutes. With introduction, breaks, and the final interview, each experiment session typically took 1 hour 30 minutes to 2 hours. Each participant was financially compensated for 60 CNY.

4.4 Data analysis

Interview transcripts were analyzed using qualitative content and thematic analysis. For the first part of the interview (i.e., participants' perception and experience), we applied content analysis focusing on three aspects: 1) perceived robot dialogue actions, 2) their effects on the participant's experience, and 3) participants' responses to the robot's behavior. Transcripts were open-coded for all mentions of these categories, iteratively refined into a shared codebook, and consolidated through team discussion. For the second part of the interview (i.e., how they envisioned the concept alignment process to unfold), we applied thematic analysis to understand the potential factors that may influence the selection of alignment strategies. Mentions of when and under what conditions each strategy or behavior was helpful or inappropriate were first open-coded and then grouped into two overarching categories: task stage and contextual factors. Within each, we identified themes that characterized participant's reasoning about how strategy use should evolve or adapt during interaction.

[3]Tencent Meeting, the same as our formative study.

5 RESULTS

The two analyses outlined above provided rich insights about the perceived dialogue behaviors and their effects, as well as potential factors that could influence strategy selection.

5.1 Perceived dialogue behaviors and their effects

5.1.1 Perceived behaviors. When discussing patterns in the robot dialogue and notable differences between rounds, participants mentioned a variety of behaviors. Table 4 summarizes the behaviors we coded, and the strategies they correspond to. Most of the dialogue behaviors envisioned in our design strategies (Section 3.3) were observed and mentioned by participants, though they may describe them using different terms. For example, *asking and confirming* was described as "asking in more detail (P2)", "confirm with me (P6)", and "likes to repeat what I said and check if I meant it (P10)". This indicated that users noticed and interpreted these actions through their own framing, suggesting that the implemented strategies were recognizable and functionally realized in interaction.

Participants discussed some dialogue behaviors notably more than others. *Forming new concepts when only given examples* (N=22), *explaining or justifying action using known concepts or rules* (N=20), and *using preset conceptual framing in dialogue* (N=18) are among the most mentioned. Overall, participants discussed more about the *generalize* (N=38) and *justify* (N=40) strategies.

5.1.2 Effects on participant experience and response. Our interviews also covered the perceived effect of the dialogue behaviors. As seen in Figure 7, the perceived effects were further categorized as being related to the participants' understanding of the robot, conceptual understanding, and user experience.

Table 4: User-perceived robot behaviors and the strategies they belong.

Strategy	Behavior code	Participants	#Mentions
Confirm	Asking and confirming	P2, P6, P10, P11, P12, P13, P15	14
	Summarizing current agreement	P9, P12, P17	4
Generalize	Form new concepts when only given examples	P2, P5, P6, P7, P10, P16	22
	Giving new examples when learning a new concept	P1, P9, P12, P15, P17	10
	Giving examples when using an abstract concept	P3, P4, P12, P13	6
Justify	Using preset conceptual framing in dialogue	P2, P7, P11, P12, P13, P14, P15, P16	19
	Explain or justify action using known concepts or rules	P3, P5, P8, P9, P12, P13	20
Propose	Propose sub-categories of a concept	P3, P8, P9, P12, P13	7
	Give task advice using new concepts	P8, P14	4
	Suggesting to move forward	P3, P12, P15	5

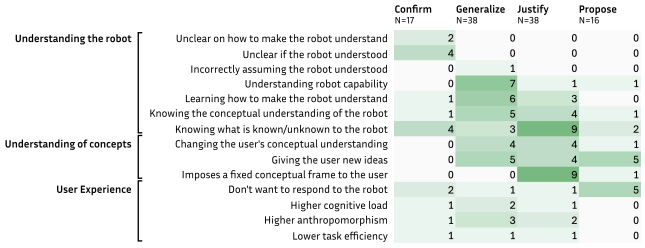


Figure 7: Code co-occurrence results summarizing the main effects of the perceived behaviors that participants mentioned, as well as their relation to the dialogue strategies.

Both positive and negative effects were mentioned. For example, some dialogue behaviors related to the *justify* strategy were said to impose a fixed conceptual frame to some participants (P7, P12, P13), which was considered harmful to task success. In contrast, other behaviors in the same strategy were viewed as beneficial for communication, as they helped participants understand what is known/unknown to the robot (P3, P8, P13). We further present these findings in detail below, organized by strategy.

Confirm. The act of *asking and confirming* was mostly associated with the participant's understanding of the robot. Some found it helpful for gauging what the robot has trouble understanding. P2 recalled specifying their preference in vague terms at the very beginning of the dialogue ("sort kitchen items according to kind"), and realizing that the robot does not have the concept of "kitchen items" when the robot asked for more detail. Similarly, periodically *summarizing the current agreement* helped P1 and P12 learn to communicate better with the robot. P12 said that by listing the learned rules, the robot informed P12 of its capabilities.

Yet, frequently asking questions and confirming confused the participants. P10 said the robot *"really liked to confirm with me. [...] Once or twice is good, [the robot] is cautious and showing me good work ethics. But it kept asking [...] and it made me worried that it couldn't understand."* Excessive confirmation has also been said to make the participants not want to respond to the robot and lower task efficiency. P11, P12, and P13 ignored some of the questions. P13 added, *"the routine questions at the end [were] like many LLMs [P13] have used before."*, saying answering them would lead the conversation off-topic.

Generalize. Behaviors in the *generalize* strategy were often discussed in relation to its positive effects on both their understanding of the robot and the concepts. Many participants noted that the robot *formed new concepts when only given examples*. P2 was impressed that the robot extrapolated the term "office supplies" from mentioning of books and document folders, saying more dialogue with the robot would "improve my communication abilities". P5 similarly noted this behavior as helpful for communicating with the robot, adding that it also brought new ideas for home organization. However, others did not like this behavior. P6 said, *"[the robot] should use the same language as me. Otherwise I would have to think about it. [...] The categories it provided were too abstract. [...] It will lower efficiency."* It was also said to increase cognitive load (P6), and when used repeatedly caused P7 to "run out of patience". P10 mistook the ability to form concepts for a high level of capability of the robot, but later was "saddened" to see errors in related object placement.

The other two behaviors were received positively despite being mentioned less often. *Giving new examples when learning a new concept* brought the participant new ideas. By listing similar objects, the robot inspired P9 to refine their original organization plan. Similarly, when the robot challenged P12's conception of "consumable" items belonging in storage by listing objects that they might use often, P12 responded in favor. P12 said this behavior changed their conceptual understanding and made the robot feel more anthropomorphic, "like it has its own ideas now." When the robot *gave examples when using abstract concepts*. P3 became more certain of the exact understanding the robot had of the concept. P4 said the behavior changed their conceptual understanding, "everybody's understanding of 'daily items' is different. [...] I didn't know its meaning. But [the robot] gave me examples, and made it easier to understand." P13 echoed by saying that the unexpected examples gave them a new understanding.

Justify. Participants talked at length about when the robot would *use its preset conceptual framing in dialogue*. According to P12, *"[the robot] conceptualized things differently than I would [...] and totally led me astray."* P12 found that the preset concepts organized objects by their type at the beginning of their conversation, which has

conditioned P12 to adopt this framing until rounds later, when P12 finally realized grouping objects by where they're used can best describe their preference. P7 and P13 echoed this point by noting that the robot was "stuck in its own reasoning(P13)" and its internal preset concepts "too general [and] should be more personalized (P7)".

Many also mentioned when the robot *explained or justified its actions using known concepts or rules*. Contrary to the above, this behavior was generally well-received. P8 felt that "after this conflict, it understood the logic deeper", and some said the robot's explanations helped understand what had gone wrong (P5, P13). P3 instead noted the benefit of this behavior in reminding them of special cases that did not fit the robot explanation, and said it made the robot "feel like a person". However, some also noted that the robot did not readily do this until expressly asked (P5, P13), but helped them to understand how to better communicate once it did. Meanwhile, P9 found some of the robot's explanations unconvincing and ignored them.

Propose. Behaviors belonging to the *propose* strategy were among the least mentioned. P3 noted when the robot would proactively *propose sub-categories of a concept* and *give task advice using new concepts*, which helped their conversation by letting P3 know what concepts the robot could understand, and how it defined certain concepts. P3, P8, and P9 also mentioned some categorization that the robot proposed gave them new ideas. However, P12 considered some conceptual categories the robot proposed "not useful" and ignored them. P8 also noted when the robot actively *suggested to move forward* during the task. This behavior made P8 feel that "the robot don't want to learn more from me". In response, P8 did not give more instructions than they would otherwise have.

5.2 Potential factors

The second stage of our interviews asked the participants to envision their ideal grounding dialogues, regarding both 1) which grounding behaviors should happen in a task when, and 2) what might determine their appropriateness. This process revealed their preferences and potential factors to consider when selecting strategy. We present the main themes identified through thematic analysis for each aspect.

5.2.1 Task stage factors. Participants expressed their preferences for how the grounding behaviors should evolve as the task progresses.

From abstract to concrete. Many characterized their ideal communication as progressing from establishing an abstract, general understanding in the beginning to refining concrete details in later stages. For example, P3 believed that the robot should focus on establishing rules and definitions first. P13 echoed this, saying the robot should give suggestions on "the structural and logical", only then move on to resolving individual instances. P4 and P9 expressed a similar idea, referring to it as "grasping the general direction" and "planning globally" first. As P4 explained, *"[The dialogue] should be top-down. [The robot] should grasp the general direction first. Even if I don't know what that direction is, [it] should give me one. Then we'll move on to individual categories, from big to small[...]"*

Proactive vs. passive. Participants described when they expect the robot to be actively leading the dialogue and providing information and ideas, versus when they passively accept information from the human. Notably, they had conflicting preferences regarding this aspect. Some expect the robot to be proactive in the beginning, citing that they felt lost when first faced with the task, not knowing what to say or how to progress (P2, P6). Others related this to the previous theme, delegating the responsibility of providing conceptual framing, general concepts, and task advice to the robot(P12, P13). P12 also attributed emotional value to this arrangement, saying, *"[the two approaches] would bring different emotional experiences. [Proactive first] would give you pleasant surprises. You'd feel you were actually saving effort. [Passive first] is just what you would normally expect from a robot. You tell it what to do and it only does it, but you still have to finish it up."*

However, not all agreed. Other participants instead preferred the robot to progress from passive to proactive. P8 and P9, for example, expected the robot to mostly listen and understand before contributing. While P3 and P8 also referenced the "abstract to concrete" approach like P12 and P13 when discussing proactivity, they instead said that the initial framing and definitions should come from the person. P14 thinks that a *proposing* and *justifying* robot at the beginning "takes too much liberty".

Specific stages. People also attributed behaviors to specific stages. People welcomed behaviors that communicated robot capabilities (P4, P6, P7, P13) and gave examples (P3, P5, P13) at the start of the dialogue. During the task process, P9 considered all strategies acceptable. P4 imagined periodically summarizing the current agreement helpful, but stressed that it should be very concise. P8 thought that *propose* and *justify* to be generally inappropriate for finalizing the task, but added that any new examples still may warrant discussion and justification. And while too many *questions* at the end would be unwelcome for P11 and P15, *summarization* behaviors from the same strategy were deemed beneficial (P9, P15).

5.2.2 Contextual factors. In addition to factors related to the task stage, participants proposed ideas for how contextual cues might help determine whether a strategy or behavior is appropriate.

When in conflict, explain. Many participants considered explaining or justifying actions suitable for when the robot's behavior deviates from user expectation (P3, P5, P6, P7, P8, P13). P6 and P13 believed that the robot should explain its reasoning "whenever there's a mistake (P13)", regardless of task stage. P3 approached this from another angle, limiting this behavior *specifically* to after a mistake has been observed.

User preference and cooperation. Some participants attributed the appropriateness of some behaviors to their own personal preference at the moment and the willingness to cooperate. P5 considered the robot a tool for completing tasks, and therefore described the robot's *transforming* strategy as "insisting on doing too much, and I couldn't stop it." P4 and P6 found some of the robot's responses too long, though noting this might be due to their "impatient" personality. Finally, P8 noted that people might have shifting preferences for robot proactivity depending on their circumstances, and the robot should query or infer their current preference each time.

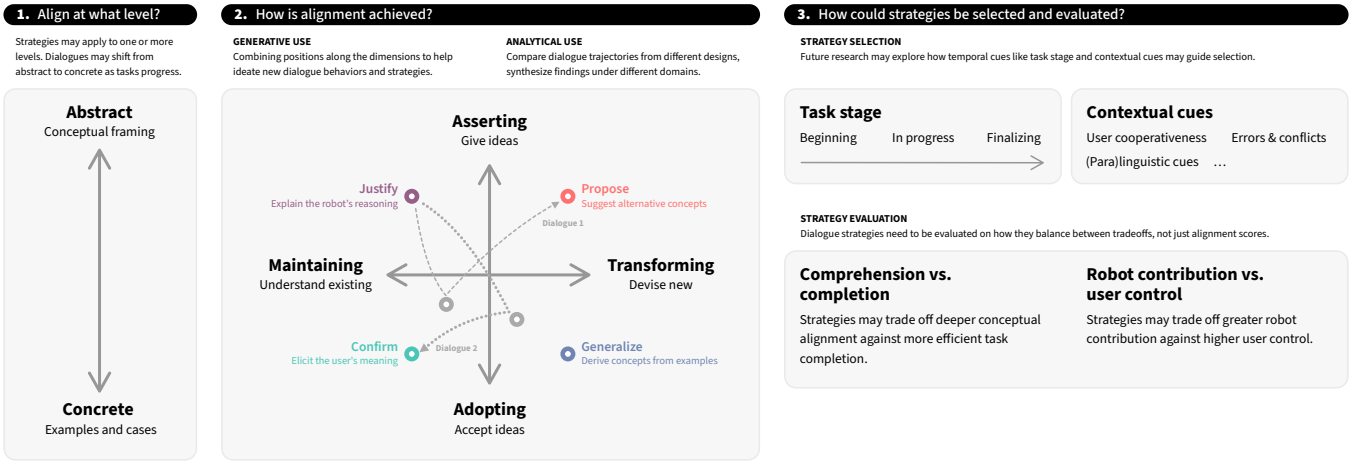


Figure 8: A design space diagram with updated dimensions and design considerations in relation to the strategy space.

(Para)linguistic cues. A small number of participants also noted linguistic or paralinguistic cues that can be taken from the user input. P12 noted cases where their speech differed in length and uncertainty, saying, *"if I gave a short instruction, I want the robot to just do it. But if it was long, [...] it should help me organize [my thoughts] before doing it."* This was echoed by P9. Later, P12 added, *"when I gave examples, it meant I just had a rough understanding of this category. That was when I needed the robot to help me expand [the concept definition]."* Regarding repeating what the user said as confirmation, P5 considered it unnecessary, while P13 considered it helpful when the user explicitly requested it.

6 A DESIGN SPACE FOR CONCEPTUAL ALIGNMENT DIALOGUES

In this section, we synthesize findings from the two studies into a design space that describes how conceptual alignment dialogues can be designed, selected, and evaluated. We discuss how designers could interpret its dimensions, the trade-offs, and its potential uses in future work.

As shown in Figure 8, the design space is organized around three complementary questions. First, alignment may operate at different levels of abstraction, ranging from concrete examples and cases to abstract rules and definitions. Second, the dialogue that builds alignment can be characterized along two core dimensions identified in our formative study: *maintaining–transforming* and *adopting–asserting*. Finally, the findings from our second study suggest possible factors that may guide strategy selection, and key trade-offs to consider when evaluating an alignment strategy.

6.1 Aligning on what level?

One interesting finding through our second study was people's emphasis on the level of abstraction when discussing their preferences for alignment dialogues (see Section 5.2.1). Many preferred that the robot speak in general terms at first using abstract concepts and definitions, progressing to finer distinctions and individual objects later. Constructs similar to this distinction were already proposed by many researchers (R1, R3, R4, R5, R7) in our formative study proposed constructs similar to this distinction, such as "focused on objects vs. definitions", "concrete objects vs. abstract definitions", or "concrete vs. abstract". As the study focused on extracting dimensions of *dialogue strategy* instead of *alignment content*, these constructs were originally intended to describe how people differed in dialogue goals: to resolve conflicting understanding through discussing the current sorting (maintain) or to collaboratively develop a new understanding (transform). However, combined with results from the second study above, we believe there is merit for an additional dimension "abstract—concrete", highlighting the level of abstraction that alignment dialogues operate on as a salient dimension. Concrete dialogues may discuss examples and special cases of concepts. Abstract ones may discuss what concepts are used to frame the task at hand. Between the two extremes lie dialogues that concern the rules and definitions of concepts at varying levels of abstraction.

The *abstract–concrete* dimension should therefore be understood as describing the level at which these strategies operate, rather than as a third equivalent strategy dimension. Each strategy may operate at one or more levels of abstraction. A strategy such as confirming may involve concrete cases or more abstract definitions, and a dialogue may shift between these levels over time. Our findings suggest that such shifts may be particularly relevant as a task progresses from establishing an initial conceptual framing toward resolving specific cases.

6.2 How is alignment achieved?

The two core dimensions outlined by our formative study describe how a strategy pursues alignment. *Maintaining–transforming* describes whether the dialogue aims to establish an understanding of existing concepts or develop new conceptualizations, while *adopting–asserting* describes whether the robot primarily adapts to the user's conceptualization or contributes its own ideas. The four strategies explored in the second study are representative combinations of the two core dimensions rather than an exhaustive set of strategies. Other strategies could occupy intermediate positions or combine behaviors differently. As an example, the conceptual

alignment tendency found by Cirillo et al.[16], where people automatically adapts to the level of abstraction used by the robot, could be transformed as a less asserting and maintaining strategy than *justify* to steer the conversation away from ambiguous concepts.

Meanwhile, we also note that the most effective or appropriate strategy may change as the dialogue and task evolve. Our findings from the second study suggest that conceptual alignment requires a process of adaptive coordination rather than fixed dialogue behavior. Figure 8 illustrates this by showing two possible dialogue trajectories. Each dialogue moves between strategies as new information, misunderstandings, or task demands emerge.

6.3 How could strategies be selected and evaluated?

What factors may guide strategy selection? Beyond the dimensions themselves, the second study identified contextual factors that may influence which strategy is appropriate at a given point in an interaction. These include task stage, user cooperativeness, errors or conflicts, and linguistic or paralinguistic cues. These factors are potential signals that a dialogue policy could use to select or transition between strategies. They also point to several future research questions, including how these signals relates to the appropriateness of strategies, and how they may affect subsequent alignment outcome.

How may the strategies be evaluated? We identify two overarching trade-offs discovered in our second study, namely 1) between contribution and control, and 2) between comprehension and completion. These trade-offs provide criteria for evaluating strategies as they are selected and applied in context. We discuss their implications for HRI design and how they relate to ongoing questions of designing for better human-robot representation alignment.

Contribution vs. control: who shapes the conceptual understanding? The trade-off between contribution and control concerns how people expect the robot to participate in shaping conceptual understanding. Importantly, participants did not share a common expectation of the robot's role. Some expected the robot to provide initial framing, general concepts, or task guidance, particularly at the beginning of the interaction, while others preferred the robot to follow their lead and adopt user-defined concepts before contributing. These differing expectations parallels how the same alignment behaviors were perceived differently. For instance, generalization or proposal of new categories was received positively when participants were open to the robot contributing ideas, but was described as "too abstract" or "not useful" when participants expected the robot to adhere to their framing. Similarly, justifications were valued by some as helpful explanations, but by others as the robot being "stuck in its own reasoning" or leading them in an unintended direction. In this sense, how people perceives and responds to alignment strategies may depend on prior expectations of how much initiative the robot should take.

At the same time, the strategies themselves may also implicitly shape these expectations. More *asserting* or *transforming* strategies, such as proposing new concepts or generalizing from examples, were taken by some as indications that the robot possessed a higher level of competence. When these expectations were not met in subsequent actions, participants reported disappointment. Conversely, when the robot remained passive or strictly adopted user concepts, some participants perceived it as lacking capability or requiring excessive guidance, perceiving increased effort. These shifts suggest that expectations of the robot's role are not fixed, but are continuously updated based on how the robot contributes to alignment. As a result, mismatches can emerge not only from differences in conceptual understanding, but also from differences in how users believe the robot should participate in the alignment process, sometimes leading to reduced cooperation such as ignoring proposals or limiting further input.

These observations suggest that designing for conceptual alignment involves both deciding and adapting how much the robot should contribute, as well as managing how that contribution is expected and interpreted over the course of interaction. HRI design may need to consider how alignment strategies imply the robot's role in shaping conceptual understanding, and how this aligns with user expectations in different tasks and contexts. More broadly, these findings indicate that representation alignment is not solely a matter of converging on shared representations, but also of coordinating and negotiating how such an alignment should unfold.

Comprehension vs. completion: when and what to align? The trade-off between comprehension and completion is related to how building and maintaining alignment is balanced with task progress. We observe that the participants did not treat alignment as something to be continuously maintained. Instead, many participants emphasized the need for alignment either at the beginning of the task when they lacked a clear conceptual framing and task plan, or at moments of misalignment when "errors" signal mismatches in understanding. For example, early confirmation and clarification helped participants see what the robot could understand, while justifications were regarded as appropriate specifically when the robot's actions deviated from expectations. Similarly, generalization and exemplification supported comprehension when participants were uncertain about the robot's interpretation of abstract concepts, but were less appreciated when introduced without a clear need.

Meanwhile, the effectiveness of these strategies shifted as the task progressed. More interactive forms of alignment, such as repeated confirmation or exemplification, were described as increasingly effortful, sometimes leading participants to ignore questions or disengage. In contrast, summarizing the current agreement remained useful even in later stages. Participants also reacted differently to *transforming* strategies over time. Although examples and generalizations could clarify or refine understanding, repeated or cognitively demanding instances were said to increase effort or "reduce task efficiency". As the task approached completion, alignment seemed to be evaluated less by its presence and more by its disruption to the progress. Late-stage proposals or unsolicited elaborations were often resisted or ignored, while concise summaries, targeted explanations after errors, or high-value examples were still welcomed.

These observations suggest that supporting conceptual alignment in task-oriented HRI is not just a matter of increasing transparency or eliciting more information. Instead, it may require strategically coordinating *when* and *about what* alignment is pursued in an ongoing task. For HRI design, this points to the need for dialogue policies that adapt alignment strategy and content to the interaction context and task progress. For achieving representation alignment, this trade-off highlights that the challenge is not only to achieve shared conceptual understanding, but to do so under constraints of human effort and task progression.

Taken together, the trade-offs we have observed suggest a need for dynamically adjusting the timing, form, and level of initiative in response to the task, context and user expectations when employing proactive conceptual alignment strategies.

6.4 Using the design space

The design space is intended for researchers and designers who develop, study, or evaluate dialogue-enabled robot systems in which conceptual alignment is important. Designers can use it prospectively to structure the dialogue behaviors that a robot should support and to reason about when different strategies may be appropriate. Researchers can use it retrospectively to characterize alignment behaviors in existing systems or interaction data, compare strategies across systems and application domains, and identify opportunities for new dialogue strategies. In both cases, the design space serves as a framework for describing and exploring alternative ways of conducting conceptual alignment.

For designers, the design space can serve as **a generative tool** to produce more dialogue strategies. In this work, we developed the four strategies by combining the most typical behaviors of the dimensions. In practice, a designer can use the space in three steps. First, identify the level of abstraction relevant to the task and determine whether the interaction should primarily maintain existing concepts or develop new ones. Second, determine how much the robot should contribute versus adopt the user's conceptualization, thereby identifying one or more representative strategies or intermediate positions along the two core dimensions. Third, use task stage and contextual cues to inform decisions about when to shift strategy, while evaluating the resulting interaction in terms of user control, conceptual understanding, and task efficiency.

The design space can also help designers differentiate and synthesize design knowledge about application domains that require conceptual alignment. On one hand, the concrete use case of the robot may narrow the space of helpful strategies. Designers could selectively define and explore additional strategies in subspaces according to their robot's application domain. In our case, many participants found the *propose* strategy unnecessary, likely because the goal was to have a neatly sorted home, and not a new understanding on how to classify personal belongings. However, this may be different for other cases. As an example, a robot aimed at co-creation and idea generation may find the *transforming* strategies useful, while an educational robot may rely more on *asserting* ideas through explanation and *maintaining* what has been understood to better communicate concepts to learners.

For HRI researchers, the design space can serve as **an analytical tool** for characterizing HRI data involving conceptual alignment and synthesizing previously isolated strategies (as discussed in Section 2.1) into a more unified understanding. Researchers can use the space to characterize and compare dialogues as trajectories through the design space. Given dialogue transcripts or system implementations, researchers can examine which alignment goals and behaviors occur, at what levels of abstraction, and how these change over the course of an interaction. For example, a robot that only ever confirms, such as the case in our second study, may leave users shouldering the entire conceptual effort. Future studies may also provide more potential factors that could inform strategy selection. This enables comparison across systems, tasks, and application domains, as well as investigation of how particular trajectories relate to task outcomes and user perceptions.

7 DISCUSSION AND CONCLUSION

This paper set out to explore how LLM-powered robots could engage in conceptual alignment dialogues. Through a formative corpus study and a user experiment, we derived design dimensions, implemented four representative strategies, and synthesized them into a design space.

Our findings extend prior work on conceptual alignment in several ways. Classic linguistic accounts treat alignment as an mutually converging process in which partners negotiate conceptual meaning[7, 21]. Our design space captures two facets of that negotiation for design purpose: the degree to which a robot asserts its own understanding versus adopt the other's, and the the extent of abandoning existing concepts in favor of entirely new ones. The *abstract–concrete* axis further echoes observations that dialogue partners shift between concrete references and higher-level alignment[21]. Our work highlight that this shift can be deliberately designed into a robot's alignment behavior. Where prior human-robot studies demonstrated that people spontaneously align their level of abstraction to a robot[16, 22], our work contributes a designer's perspective: the robot can proactively choose its strategy to affect conceptual understanding, but that choice carries consequences for user experience. Our work also speaks to recent investigations of grounding in LLM-based dialogue. Shaikh et al.[43] found that LLMs often skip the feedback signals, such as clarifying questions that humans rely on. Yet, our participants' frustration with excessive confirmation in the *confirm* strategy rounds shows that merely imposing grounding behaviors is insufficient. The third part of our design space highlight task and contextual factors that could guide such behaviors, and necessary aspects to evaluate beyond task or alignment scores.

Our study also has several limitations. First, we tested the strategies in a simulated environment with a text-based interface and no physically embodied robot. This choice allowed us to isolate dialogue behaviors from confounding modalities, and enabled us to examine alignment at a more fundamentally linguistic and conceptual level. However, the absence of a physical robot means our study captures only part of what human-robot interaction entails. A co-present robot could use gesture, gaze, and spatial reference to ground concepts (e.g., pointing to a shelf while saying "put the tools here"), which might reduce the need for some verbal alignment moves or change their perceived naturalness. Social presence effects could also alter how users respond to certain strategies. Therefore,

the term "human-robot dialogue" should be understood here as referring to the dialogue modality of HRI; the findings provide a design vocabulary for this modality, but their transfer to embodied settings must be empirically tested. Future work should replicate these strategies on a physical robot to examine how embodiment moderates user perceptions of alignment behaviors. Second, we only tested the strategies in one possible domestic robot task (i.e., household unpacking) with a group of Mandarin-speaking participants. We acknowledge that both the application domains and cultural factors may influence people's perceptions of these dialogues. However, by employing a task-oriented setting, we highlight that representation alignment is not an isolated task of building shared representations, but is embedded in concrete tasks, contexts, and scenarios. We also observe that most existing human-robot dialogue studies are based on English-speakers. By including speakers of another language, our work contributes a valuable and complementary perspective. We encourage future research to apply and evaluate our design space in other contexts and cultures, and to further refine it.

Despite these limitations, we believe this work represents an important step toward understanding human-robot conceptual alignment. Through the proposed design space, along with empirical evidence and insights from the implementation, we envision providing HCI researchers and designers with a foundation for discussion and further exploration in this emerging area.

8 GenAI usage disclosure

Generative AI models were used in this paper in three ways. First, we used ChatGPT to aid in debugging the experiment system and generating utility functions. Second, it was used to correct grammar errors and improve word choice during writing. Finally, our study addresses LLM-driven robots, and therefore used locally-hosted open-source models including Qwen2.5-VL-32B[1], BGE-M3-Embedding, and BGE-Reranker-v2-m3[14, 52–54].

Acknowledgments

This research work was supported by the National Natural Science Foundation of China (52575187, 52175144), and the Opening Project of the State Key Laboratory of General Artificial Intelligence (SKLAGI2025OP14). The first author would like to thank Yuanda Hu, Hexin Zhang, and Jiawen Huang for their help in piloting the second study.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. doi:10.48550/arXiv.2502.13923

[2] Michael Mose Biskjaer, Peter Dalsgaard, and Kim Halskov. 2014. A constraint-based understanding of design spaces. In *Proceedings of the 2014 conference on Designing interactive systems*. ACM, Vancouver BC Canada, 453–462. doi:10.1145/2598510.2598533

[3] Andreea Bobu, Chris Paxton, Wei Yang, Balakumar Sundaralingam, Yu-Wei Chao, Maya Cakmak, and Dieter Fox. 2022. Learning Perceptual Concepts by Bootstrapping From Human Queries. *IEEE ROBOTICS AND AUTOMATION LETTERS* 7, 4 (Oct. 2022), 11260–11267. doi:10.1109/LRA.2022.3196164

[4] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic Alignment between People and Computers. *Journal of Pragmatics* 42, 9 (Sept. 2010), 2355–2368. doi:10.1016/j.pragma.2009.12.012

[5] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Ash Brown. 2011. The Role of Beliefs in Lexical Alignment: Evidence from Dialogs with Humans and Computers. *Cognition* 121, 1 (Oct. 2011), 41–57. doi:10.1016/j.cognition.2011.05.011

[6] Susan E. Brennan. 1998. The Grounding Problem in Conversations With and Through Computers. In *Social and Cognitive Approaches to Interpersonal Communication*. Psychology Press, Hillsdale, NJ.

[7] Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22, 6 (1996), 1482–1493. doi:10.1037/0278-7393.22.6.1482

[8] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46, 3 (Sept. 2014), 904–911. doi:10.3758/s13428-013-0403-5

[9] Natalia Calvo-Barajas, Anastasia Akkuzu, and Ginevra Castellano. 2024. Balancing Human Likeness in Social Robots: Impact on Children's Lexical Alignment and Self-disclosure for Trust Assessment. *J. Hum.-Robot Interact.* 13, 4 (May 2024), 1–27. doi:10.1145/3659062

[10] Sabrina Campano, Jessica Durand, and C. Clavel. 2014. Comparative analysis of verbal alignment in human-human and human-agent interactions. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, 4415–4422. https://www.semanticscholar.org/paper/Comparative-analysis-of-verbal-alignment-in-and-Campano-Durand/3abdcc2d0d0be072cf46b0b426061b041c74a658

[11] A. Cangelosi and F. Stramandinoli. 2018. A review of abstract concept learning in embodied agents and robots. *Philosophical Transactions of the Royal Society B: Biological Sciences* 373, 1752 (2018), 6. doi:10.1098/rstb.2017.0131

[12] Robin Shing Moon Chan, Anne Marx, Alison Kim, and Mennatallah El-Assady. 2025. A Design Space for Intelligent Dialogue Augmentation. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. ACM, Cagliari Italy, 18–36. doi:10.1145/3708359.3712096

[13] Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'Grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 4283–4305. doi:10.18653/v1/2021.findings-acl.375

[14] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2023. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation.

[15] Francesco Chiossi, Julian Rasch, Robin Welsch, Albrecht Schmidt, and Florian Michahelles. 2025. Designing Intent: A Multimodal Framework for Human-Robot Cooperation in Industrial Workspaces. doi:10.48550/arXiv.2506.15293

[16] Giusy Cirillo, Elin Runnqvist, Kristof Strijkers, Noël Nguyen, and Cristina Baus. 2022. Conceptual Alignment in a Joint Picture-Naming Task Performed with a Social Robot. *Cognition* 227 (Oct. 2022), 105213. doi:10.1016/j.cognition.2022.105213

[17] Herbert H. Clark. 1996. *Using language* (7. print ed.). Cambridge University Press, Cambridge. https://www.cambridge.org/core/product/identifier/9780511620539/type/book

[18] Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to Discourse. *Cognitive Science* 13, 2 (1989), 259–294. doi:10.1207/s15516709cog1302_7

[19] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A Universe of Annotated 3D Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13142–13153.

[20] Mary Ellen Foster, Manuel Giuliani, Amy Isard, Colin Matheson, Jon Oberlander, and Alois Knoll. 2009. Evaluating Description and Reference Strategies in a Cooperative Human-Robot Dialogue System. In *Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence*. AAAI.

[21] Simon Garrod and Anthony Anderson. 1987. Saying What You Mean in Dialogue: A Study in Conceptual and Semantic Co-Ordination. *Cognition* 27, 2 (Nov. 1987), 181–218. doi:10.1016/0010-0277(87)90018-7

[22] Simone Gastaldon and Giulia Calignano. 2025. Linguistic alignment with an artificial agent: A commentary and re-analysis. *Cognition* 259 (June 2025), 106099. doi:10.1016/j.cognition.2025.106099

[23] Dimitra Gkatzia and Francesco Belvedere. 2021. "What's this?" Comparing Active learning Strategies for Concept Acquisition in HRI. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (ACM/IEEE International Conference on Human-Robot Interaction)*. IEEE Computer Society, Boulder, CO, USA, 205–209. doi:10.1145/3434074.3447160

[24] Kim Halskov, Graham Dove, and Aron Fischel. 2021. Constructing a Design Space from a Collection of Design Examples. *She Ji: The Journal of Design, Economics, and Innovation* 7, 3 (Sept. 2021), 462–484. doi:10.1016/j.sheji.2021.07.001

[25] Marc Hassenzahl and Rainer Wessler. 2000. Capturing Design Space From a User Perspective: The Repertory Grid Technique Revisited. *International Journal of Human-Computer Interaction* 12, 3-4 (Dec. 2000), 441–459. doi:10.1080/10447318.2000.9669070

[26] Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. 2019. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS One* 14, 10 (Oct. 2019), e0223792. doi:10.1371/journal.pone.0223792

[27] Peter H. Kahn, Nathan G. Freier, Takayuki Kanda, Hiroshi Ishiguro, Jolina H. Ruckert, Rachel L. Severson, and Shaun K. Kane. 2008. Design patterns for sociality in human-robot interaction. In *Proceedings of the 3rd international conference on Human robot interaction - HRI '08*. ACM Press, Amsterdam, The Netherlands, 97. doi:10.1145/1349822.1349836

[28] Majeed Kazemitabaar, Oliver Huang, Sangho Suh, Austin Z Henley, and Tovi Grossman. 2025. Exploring the Design Space of Cognitive Engagement Techniques with AI-Generated Code for Enhanced Learning. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 695–714. doi:10.1145/3708359.3712104

[29] Mitsuhiko Kimoto, Takamasa Iio, Masahiro Shiomi, Ivan Tanev, Katsunori Shimohara, and Norihiro Hagita. 2016. Alignment Approach Comparison between Implicit and Explicit Suggestions in Object Reference Conversations. In *Proceedings of the Fourth International Conference on Human Agent Interaction (HAI '16)*. Association for Computing Machinery, New York, NY, USA, 193–200. doi:10.1145/2974804.2974814

[30] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence* 6, 4 (April 2024), 383–392. doi:10.1038/s42256-024-00820-y

[31] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. 2022. AI2-THOR: An Interactive 3D Environment for Visual AI. arXiv:1712.05474 [cs] doi:10.48550/arXiv.1712.05474

[32] Ryo Kuniyasu, Tomoaki Nakamura, Tadahiro Taniguchi, and Takayuki Nagai. 2021. Robot Concept Acquisition Based on Interaction Between Probabilistic and Deep Generative Models. *Frontiers in Computer Science* 3 (Sept. 2021), 14 pages. doi:10.3389/fcomp.2021.618069

[33] Matthijs Kwak, Kasper Hornbæk, Panos Markopoulos, and Miguel Bruns Alonso. 2014. The design space of shape-changing interfaces: a repertory grid study. In *Proceedings of the 2014 conference on Designing interactive systems (DIS '14)*. Association for Computing Machinery, New York, NY, USA, 181–190. doi:10.1145/2598510.2598573

[34] I. Scott MacKenzie. 2013. *Human-computer interaction: an empirical research perspective* (first edition ed.). Morgan Kaufmann is an imprint of Elsevier, Amsterdam.

[35] Allan MacLean, Richard M Young, Victoria M E Bellotti, and Thomas P Moran. 1991. Questions, Options, and Criteria: Elements of Design Space Analysis. *Human–Computer Interaction* 6 (1991), 201–250. doi:10.1080/07370024.1991.9667168

[36] Mati Mőttus, Evangelos Karapanos, David Lamas, and Gilbert Cockton. 2016. Understanding Aesthetics of Interaction: A Repertory Grid Study. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI '16)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/2971485.2996755

[37] Jonghyuk Park, Alex Lascarides, and Subramanian Ramamoorthy. 2023. Interactive Acquisition of Fine-grained Visual Concepts by Exploiting Semantics of Generic Characterizations in Discourse. In *Proceedings of the 15th International Conference on Computational Semantics*, Maxime Amblard and Ellen Breitholtz (Eds.). Association for Computational Linguistics, Nancy, France, 318–331.

[38] Dhaval Parmar, Stefán Ólafsson, Dina Utami, Prasanth Murali, and Timothy Bickmore. 2020. Navigating the Combinatorics of Virtual Agent Design Space to Maximize Persuasion. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1010–1018.

[39] Rajkumar Pujari and Dan Goldwasser. 2025. LLM-Human Pipeline for Cultural Grounding of Conversations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 1029–1048. doi:10.18653/v1/2025.naacl-long.48

[40] Sunayana Rane, Polyphony J. Bruna, Ilia Sucholutsky, Christopher Kello, and Thomas L. Griffiths. 2024. Concept Alignment. doi:10.48550/arXiv.2401.08672

[41] Allison Sauppé and Bilge Mutlu. 2014. Design patterns for exploring and prototyping human-robot interactions. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. ACM Press, Toronto, Ontario, Canada, 1439–1448. doi:10.1145/2556288.2557057

[42] Omar Shaikh, Kristina Gligoric, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding Gaps in Language Model Generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 6279–6296. doi:10.18653/v1/2024.naacl-long.348

[43] Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2023. Grounding or Guesswork? Large Language Models are Presumptive Grounders. http://arxiv.org/abs/2311.09144

[44] Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating Rifts in Human-LLM Grounding: Study and Benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 20832–20847. doi:10.18653/v1/2025.acl-long.1016

[45] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary C. Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. *CoRR* (Jan. 2024). https://openreview.net/forum?id=eCvTyhcyn8

[46] Jeffrey Sherer, Robbie McPherson, Sattwik Mohanty, Guilhem Santé, Greta Gandolfi, Marta Romeo, and Alessandro Suglia. 2025. Follow Me: A Study on the Dynamics of Alignment Between Humans and LLM-Based Social Robots. In *Social Robotics*, Oskar Palinko, Leon Bodenhagen, John-John Cabibihan, Kerstin Fischer, Selma Šabanović, Katie Winkle, Laxmidhar Behera, Shuzhi Sam Ge, Dimitrios Chrysostomou, Wanyue Jiang, and Hongsheng He (Eds.). Springer Nature, Singapore, 487–496. doi:10.1007/978-981-96-3519-1_44

[47] Arjen Stolk, Lennart Verhagen, and Ivan Toni. 2016. Conceptual Alignment: How Brains Achieve Mutual Understanding. *Trends in Cognitive Sciences* 20, 3 (March 2016), 180–191. doi:10.1016/j.tics.2015.11.007

[48] Felix B. Tan and M. Gordon Hunter. 2002. The Repertory Grid Technique: A Method for the Study of Cognition in Information Systems. *MIS Quarterly* 26, 1 (2002), 39–57. doi:10.2307/4132340

[49] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. doi:10.48550/arXiv.2302.13971

[50] Konstantinos Tsiakas and Dave Murray-Rust. 2024. Unpacking Human-AI interactions: From Interaction Primitives to a Design Space. *ACM Transactions on Interactive Intelligent Systems* 14, 3 (Sept. 2024), 1–51. doi:10.1145/3664522

[51] Koen van Lierop, Martijn Goudbeek, and Emiel Krahmer. 2012. Conceptual Alignment in Reference with Artificial and Human Dialogue Partners. *Proceedings of the Annual Meeting of the Cognitive Science Society* 34, 34 (2012), 1066–1071.

[52] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding.

[53] Shitao Xiao, Zheng Liu, Peitian Zhang, and Xingrun Xing. 2023. LM-Cocktail: Resilient Tuning of Language Models via Model Merging.

[54] Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve Anything To Augment Large Language Models.

A Curating concepts and object images for the sorting task

A.1 Selection of concepts

For the selection of concepts, we first brainstormed concepts according to high, medium, and low levels of concreteness. In this study, we define concepts with a high level of concreteness as those that are related to perceivable aspects of an object, such as concepts related to shape, colors, and materials. Concepts with a low level of concreteness are defined as those related to imperceptible attributes, such as function, manufacturing process, or cultural significance.

The concepts were selected and grouped into sets of three by brainstorming potentially ambiguous objects for pairings of concepts. Concepts with more ambiguous objects, and thus with a high degree of overlap, are selected and grouped. As a result, we formed six sets of concepts for each combination of task and level of concreteness, as shown in Table 5.

We further verified the concreteness of these concepts using a dataset of concreteness ratings for English words created by Brysbaert, Warriner, and Kuperman[8]. This dataset provides concreteness ratings for more than 40,000 common words and terms in

Table 5: The concepts used for the sorting task.

Concreteness	Set	Concepts
High	A	Slab, block, stick
	B	Dark-colored, sepia-colored, multicolored
Medium	A	Container, physical support, weight
	B	For breaking, for crafting, for food
Low	A	Natural, artificial, biological
	B	Mechanical, digital, analog

English. Where a concept is not present in the dataset, an appropriate synonym is selected. For example, we substituted "physical support" with "support", and "For destroying" with "destroy".

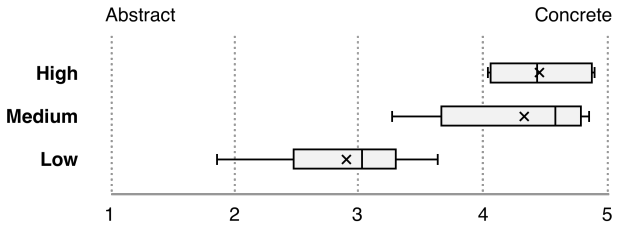


Figure 9: The concreteness ratings of the selected concepts

A.2 Selection of object images

For object images, we used the THINGS dataset created by Hebart et al.[26], which consists of 1,854 classes of objects, each with several image depictions. In our case, we wish to maximize the potential ambiguity within each set of images. To this end, we first labeled each object class according to their membership in each concept set. As the dataset contains over 1,000 classes, we utilized the in-context learning ability of large language models to perform the first round of labeling. In our case, we used the LLaMA 13B model by Meta[49]. The dataset was then filtered for objects that belonged to two or more concepts, at which point the labeling was manually checked for the filtered subset, and errors were corrected. Finally, 40 images were selected for each set of concepts, resulting in 240 images.

B Demographics information of study participants

With their consent, we surveyed the participants of the second study about their experience in using relevant products and knowledge in related domains. Figure 10 shows the aggregated statistics of the results. The surveyed products included robots and other dialogue-enabled devices.

Overall, the participants had a diverse range of familiarity with dialogue-enabled products. As expected, experience with actual robots was somewhat limited, with most (N=13) participants having used physical robots "a few times". However, two users also reported using robots "every day" or "a few days a week", which could have provided useful insights in our interviews. More participants were familiar with chatbots, which can be attributed to the already widespread use of LLMs at the time of the experiment.

The participants were recruited largely from past and current student social media groups that the authors had access to, resulting in higher expertise ratings for interaction design on average. However, our recruitment also reached graduates who have since found employment in robot and LLM-related areas, and could contribute expert insight to the findings.

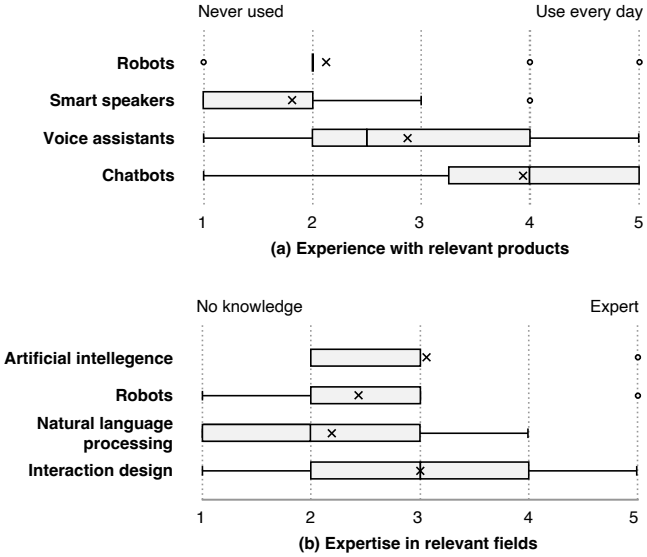


Figure 10: The ratings by the second study participants, on their experience with dialogue-enabled social agents and expertise in related fields.