Align with Me, Not *TO* Me: How People Perceive Concept Alignment with LLM-Powered Conversational Agents

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

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

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

Abstract

Concept alignment—building a shared understanding of concepts—is essential for human and human-agent communication. While large language models (LLMs) promise human-like dialogue capabilities for conversational agents, the lack of studies to understand people's perceptions and expectations of concept alignment hinders the design of effective LLM agents. This paper presents results from two lab studies with human-human and human-agent pairs using a concept alignment task. Quantitative and qualitative analysis reveals and contextualizes potentially (un)helpful dialogue behaviors, how people perceived and adapted to the agent, as well as their preconceptions and expectations. Through this work, we demonstrate the co-adaptive and collaborative nature of concept alignment and identify potential design factors and their trade-offs, sketching the design space of concept alignment dialogues. We conclude by calling for designerly endeavors on understanding concept alignment with LLMs in context, as well as technical efforts to combine theory-informed and LLM-driven approaches.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Natural language interfaces**; • **Computing methodologies** → *Discourse, dialogue and pragmatics*; Natural language generation.

Keywords

Concept Alignment, Conversational Agents, Large Language Models

ACM Reference Format:
Shengchen Zhang, Weiwei Guo, and Xiaohua Sun. 2025. Align with Me, Not *TO* Me: How People Perceive Concept Alignment with LLM-Powered Conversational Agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3706599.3720126

*Corresponding author.

CHI EA '25, April 26–May 1, 2025, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/2025/04
https://doi.org/10.1145/3706599.3720126

1 Introduction

Concepts serve as one of the basic carriers of meaning and are essential for communicating more complex ideas and building a shared understanding. Much work has been done to equip conversational agents, such as chatbots and social robots, with an understanding of human concepts to enable interactive learning[18, 26, 30], long-term interaction, and personalization[25, 32]. Meanwhile, the rise of large language models (LLMs) promises human-like dialogue capabilities across domains. To produce natural dialogue and prevent harmful content, these models are fine-tuned to "align" to human dialogue patterns, values, and the user's goals and intentions[22, 46]. As a result, LLM-powered agents already seem to communicate using common concepts in a human-like manner. However, related work also suggests that LLM agents misrepresent concepts due to biases in training data[14, 44] and lack grounding behaviors[34] that check whether concepts are mutually understood.

Studies of both human and human-agent dialogue have argued that people engage in *concept alignment*[5, 17, 32], where dialogue participants interactively build shared meaning of concepts, to enable effective conversations. To design better LLM conversational agents, it is necessary to understand to what extent these models engage in concept alignment during conversations and how people perceive and expect such alignment to happen. Previous work suggests that people will actively match the agent's choice of concept[41] and level of abstraction[12] to achieve alignment. However, studies involving LLM agents that could also produce alignment behaviors and dialogues are relatively lacking. In addition, the existing literature focuses mainly on objective linguistic behaviors. It is still unclear how people subjectively perceive the agent's attempt at concept alignment and how different approaches to dialogue may affect people's experience.

This paper aims to fill this research gap. We formulate the following research questions:

- **RQ1.** How does an LLM-driven agent behave when there is conceptual misalignment? How may the behavior differ from that of humans?
- **RQ2.** What are the perceptions, responses, and expectations of people about agent behaviors? What behaviors do people perceive as helpful or unhelpful for alignment, and why?

To answer the research questions, we conducted two studies in which human-human and human-agent pairs sorted photos of objects based on their understanding of concepts and engaged in discussions. We compare the quantitative results of the two studies to reveal differences in linguistic features. We then further contextualize the results using qualitative analysis of interviews with

participants in the human-agent study and discuss the implications of our findings for agent dialogue design and future research.

2 Related work

This paper is situated within existing research on AI and HCI related to alignment and dialogue design for LLM agents. We present an overview of relevant literature in these areas.

2.1 Alignment with Conversational Agents

Informed by cognitive science and linguistics, research on conversational agents has a long-standing interest in linguistic alignment[5] and grounding[7]. Studies have found that both human and human-agent speakers align at multiple levels. On a *micro* level, researchers studied *lexical* and *verbal alignment*, where dialogue participants gradually show similar lexicon and dialogue behaviors. People had consistent patterns in verbal alignment when faced with humans or agents[10], and lexical alignment may have an effect on reference grounding[24], understanding[36], satisfaction[50], trust[37], and acceptance[9]. On a *meso* level, research focuses on the alignment of semantic meaning, including the usage of concepts. People are found to align with the agent's choice of concept[41] and level of abstraction[12]. Technical methods have also been proposed to learn novel concepts through dialogue and other forms of interaction[18, 26, 30]. On a *macro* level, work focused on studying the alignment of knowledge and values, inspired by research in human-AI alignment. For knowledge, existing work aims to build a shared representation[4] and acquire user knowledge[11]. For values, researchers studied people's response to agents with different values[19] and developed computational techniques for personal value alignment[3]. While work on micro- and macro-level alignment have investigated user perceptions and responses, similar work on concept alignment is still lacking. Moreover, LLMs present a significant shift in agent implementation, necessitating the current research.

2.2 Dialogue Design for LLM Agents

Recent studies in HCI have begun to understand how to design dialogue to address potential issues and better utilize LLMs in conversational agents. The introduction of LLMs in agent implementation raised issues about generation time[49], hallucination[29], safety[47], and explainability[1]. LLMs also gave greater variability and high-level control over dialogue design, such as style, length, and personality, which may in turn have effects on user satisfaction[21] and personality attribution[16]. LLMs have also been studied in different embodiments[23] and with different identities and debate styles[38]. These studies provided initial insight into how to better design dialogue for LLM-based conversational agents. This paper contributes to this understanding by discussing the design implications for LLM-driven concept alignment.

3 Study Procedure

To answer our research questions, we conducted two studies: one with pairs of human participants and one with human-agent pairs. We first describe the overall procedure of the two studies, and then discuss specific recruitment and implementation details for each.

3.1 Task Design

We used a task design inspired by the picture-naming task in psycholinguistics[6, 12]. The basic form of the picture-naming task has a participant see a picture and interact with a partner. The task tests whether the participant adopts their partner's use of a less common terminology over a typically favored one after the interaction. To capture subtle differences in conceptual understanding, we modified the task to use multiple images of everyday objects for each concept. In addition, we specifically curated the concepts and objects to have ambiguity (Figure 1), thus creating potential misalignment and eliciting dialogue. This process is described in detail in Appendix A.

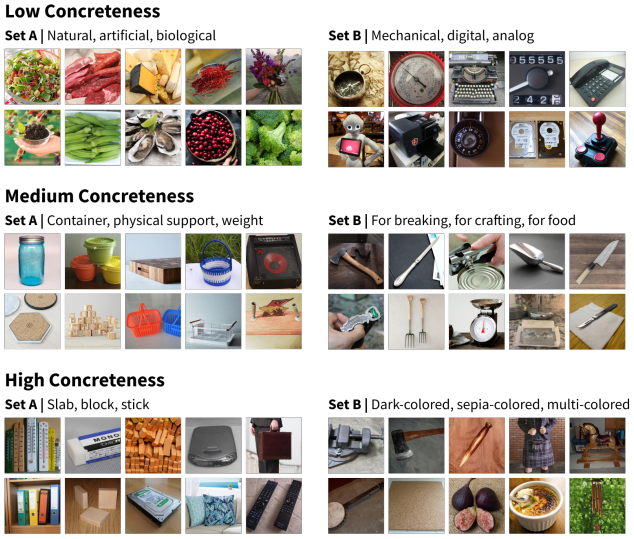


Figure 1: The six sets of concepts and an example subset of the object images we curated. Labels show the concepts and their concreteness level.

As shown in Figure 2, participants first sorted images into given concepts (classification task) or grouped them to form new concepts (formation task). They then discussed their understanding with a dialogue partner, instructed to "try to reach an agreement, but not necessarily". Finally, to test the change in alignment, we asked participants to individually sort new images according to their understanding of the concepts after the discussion.

3.2 Study implementation

Both studies were conducted using online meeting software. After signing a consent form, participants were asked to enter a web-based interface and share their screen. We then introduced the tasks using an example. Each pair performed both versions of the task in random order, with a five-minute break in between. The only difference between the two studies was that the human-agent study was followed by a semi-structured interview to better understand the experience and thoughts of the participants.

To make the study more accessible, the human-human study made use of FigJam[1], a web-based collaborative tool for hosting

[1] https://www.figma.com/figjam/

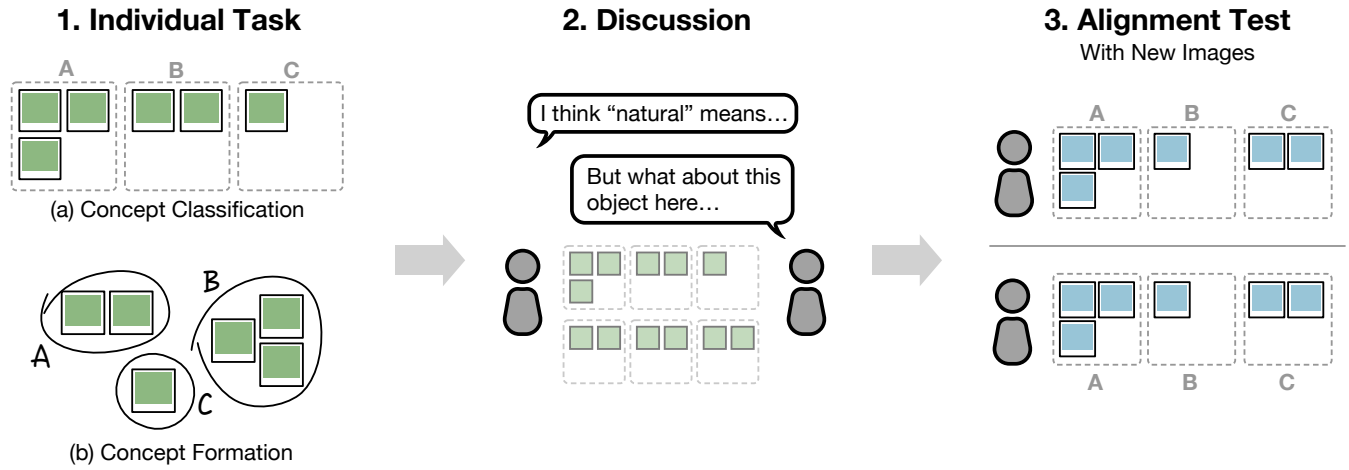


Figure 2: An overview of the concept alignment task, where a pair of participants (human-human or human-agent) sort images of objects based on their understanding of concepts. Steps 1–3 show the procedure of the task, while (a) and (b) show two forms of the individual task.

workshop sessions. We provided a template with marked areas for sorting the images. Sorting can be done by simply dragging the provided images to one of the areas. The human-agent study requires synchronizing task status between the participant's interface and the agent running on our system. To simplify the integration, we opted for building a custom web-based interface. The interface was similar to our FigJam template in both content and interaction, with an additional panel for initiating conversation with the agent.

To keep the two studies consistent, both studies used speech. The core dialogue module for Study 2 was implemented using GPT-4o API[2], with prompts similar to the instructions for the participant. To simulate a more realistic situation, we provided the object images and the sorting results to the API as images without textual hints. A more detailed description of the agent implementation can be found in Appendix B. The FigJam template and code we used are provided in the supplementary materials.

3.3 Participants

For the human-human study, we recruited 48 participants through three major social media groups (400–500 members each) of the college department. 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 ranged from 18–30 and two ranged from 31–40 years. For the human-agent study, we recruited 12 participants from both the participants of the previous study (N=8) and word-of-mouth (N=4). 8 of the participants identified as female and 4 as male. Most of the participants (N = 10) were 18–30 years old. Two participants were 50–59 years old.

4 Quantitative Results

For the discussion recordings collected from the two studies, we performed a content analysis of the transcripts and compared the distribution of the dialogue acts.

[2]https://platform.openai.com/docs/models/gpt-4o

4.1 Analysis procedure

First, we systematically coded the transcripts from the two studies according to a predetermined codebook, as shown in Table 1. To create the code book, we conducted a brief survey of the relevant literature on human-human and human-agent dialogue aimed at aligning understanding or reaching agreement. References are listed for each code. Additionally, we included two codes for non-task-related social dialogue and incomplete speech. The codes were further grouped into larger categories according to their functions in dialogue, namely 1) **grounding** acts for building and verifying mutual understanding, 2) **argumentation** acts for debate and negotiation, and 3) **social** acts for non-task-related discussion. Instead of coding each uninterrupted speech segment (an "utterance") by a speaker, which may convey multiple meanings, we adopted a common approach in the surveyed papers to divide each utterance into *dialogue acts* — defined as a basic unit of dialogue that conveys one single meaning[40].

4.2 Results

The human-human study participants made use of all the dialogue acts that we coded. The most used acts belong to the argumentation group (73.5%), followed by grounding (25.8%). Human-agent study participants demonstrated a similar distribution (65.8% and 26.9% respectively). Notably, the proportion of social acts (7.3%) is closer to the agent (10.6%) than the participants of Study 1 (0.8%). Meanwhile, the agent's dialogue behavior showed a rather different distribution, with grounding acts being the most used (57.8%), followed by argumentation acts (31.6%).

Figure 3 shows a detailed breakdown of the distribution of each dialogue act. Although participants in both studies showed overall similar distributions, differences exist for specific dialogue acts. Within the grounding group, participants of study 2 asked more questions (+7.5%) and accepted/acknowledged less of what the other said (-5.5%). In argumentation, they demanded more justification

Table 1: The codebook used in the analysis.

Code	Definition	Example	References
Accept/acknowledge	Show that a statement is understood or agreed.	"Yeah, it makes sense."	[7][33][31][35][15]
Ask question	Ask a question with the intention of getting new information.	"Can you define 'mechanical'?"	[43][31][28][35]
Clarify	Ask a question with the intention of getting verification or clarification.	"Are you referring to the blueberries?"	[7]
Assert belief	State a fact or opinion that the speaker holds.	"I think [this object] is mechanical."	[35][2][31][28]
Attack belief	Express disagreement with a particular belief of the other speaker.	"A fire alarm isn't an 'office supply'."	[35][2]
Reject Claim	Express non-acceptance for the other person's justification or proposal.	"No need to do that."	[33][35][15]
State uncertainty	Express uncertainty regarding a fact or opinion.	"I don't know what it is."	[35]
Demand justification	Demand justification for a stated fact or opinion.	"Why did you put it in this category?"	[43][31][28][35]
Provide justification	Provide a statement that is believed to justify a stated fact or opinion.	"Because it has nothing to do with analog signals."	[33][31][28][35]
Propose action	Propose a (joint) action to be taken.	"Let's discuss the other categories."	[43][33][42][15]
Propose idea	State a fact or opinion to be possible without implying truthfulness.	"...We can put it in 'entertainment'."	[43][35]
Social chatting	Any parts of the dialogue that are not directly related to the task.	"What should I call you?"	
Incomplete	Cut-off speech or recognition errors.		

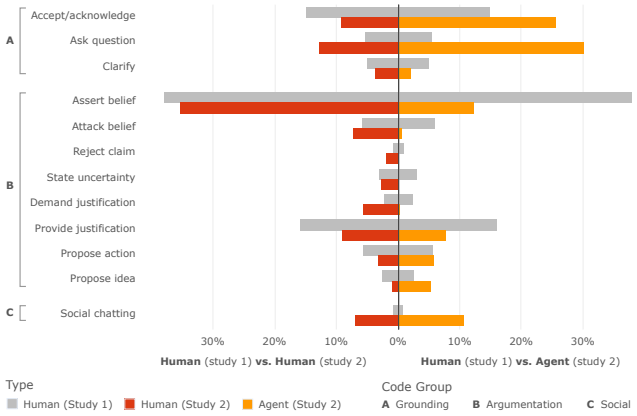


Figure 3: Detailed comparison of dialogue act distributions.

(+3.4%) and provided less justification for themselves (-6.7%). They also engaged in more social dialogue than the participants of Study 1 (+6.3%).

On the other hand, the agent's distribution of dialogue acts presents generally larger differences. In the grounding group, the agent accepted and acknowledged more, with +10.6% than Study 1 participants, and +16.1% than their dialogue partners in Study 2. The agent also asked more questions (+24.7% and +17.2% respectively). In argumentation, the agent put forward fewer assertions (-25.6% and -23.1%), and very few (less than 1%) attacks, rejections, statements of uncertainty, or demands for justification. The agent also provided less justification than people in Study 1 (-8.2%), but proposed ideas more than people in Study 2 (+4.2%). Similar to people in Study 2, the agent engaged in more social dialogue than the participants of Study 1 (+9.7%).

5 Qualitative Results

To answer the latter research question, we used thematic analysis to analyze the transcripts of the interview recordings. This process yielded high-level themes regarding people's perceptions of and responses to the agent, as well as revealed their preconceptions and expectations about concept alignment with a conversational agent.

5.1 Perception of the agent

Many participants noted the agent's **tendency to be amenable and avoid conflict** (P2, P4, P6, P7, P8, P9, P10, P11, P12). As P2 put it, "Seems that it would agree to whatever I say." The agent was perceived to express less disagreement (P6, P7, P8, P9, P10), would not point out the participants' "mistakes"(P2, P12), and would easily retract its own belief if challenged (P7). These comments often accompanied observations on the agent's *reluctance to explain or justify* its reasoning. Participants mentioned that the agent did not explain its sorting results (P2, P6, P7, P8), and only did so when specifically asked (P6). P6 wished for a more "confrontational" discussion, believing that it would improve the understanding of both parties. Along with the tendency to agree, the agent was also perceived as **sociable**, with some participants actively chatting after the task-related conversations were over (P2, P9). Participants were positive (P2, P9) or generally neutral towards social chit-chat before and after the discussion. However, instances of chit-chat *during* task discussion were considered off-topic and frowned upon (P5, P6). The participants also pointed out a **pattern in the agent's speech** where each utterance consists of an expression of agreement, some general comments, and a final question to prompt further dialogue

(e.g., "Anything else you want to discuss?") (P2, P4, P7, P11, P12). They described the questions as "mechanical (P2, P4, P5, P7, P12)" and "too broad (P7, P9)", and considered this behavior harmful to reaching alignment (P5, P7, P8, P10).

5.2 Behaviors and confidence in alignment

Participants expressed mixed confidence in their alignment with the agent. While some participants expressed relatively high confidence (P2, P3, P6, P9) for both sessions, others expressed reservations about one of the sessions (P1, P4, P5, P8, P10, P12) or claimed low confidence overall (P7, P11). Each participant also named instances of agent behavior that led to their belief. **Agreeing through summarizing, paraphrasing, or extrapolating** convinced participants that their ideas had been understood (P3, P5, P10, P12). The participants expressed confidence when the agent helped summarize the discussion using simple concepts (P3) and adopted the terms they used (P12). Conversely, a plain expression of agreement or simply repeating without paraphrasing was not considered an adequate indicator of alignment (P4, P7, P8, P10). **Discussing definitions and strategies** is preferred over discussing concrete objects (P1, P3, P5, P8, P10, P11). They liked when the agent (P3, P8) or themselves (P1, P10) presented definitions for each concept and felt confident in alignment, and felt uncertain about sessions where they only discussed specific objects (P5, P10, P11). P7 also noted that people would compare not only individual definitions, but also the pros and cons of their overall strategies.

5.3 Adapting to the Agent

The participants noted how they had to **adapt to a different conversation style** (P1, P11). P11 noted that the pattern in the agent's speech (agreement, some comments, then a question) made it more predictable, which they exploited to reach an agreement faster in the second session. Furthermore, P4 said that they became comfortable being more assertive when faced with the agent and "even to offend it". In addition to the conversational style, participants also **adapted their phrasing to be more formal and precise** to better communicate with the agent (P1, P2, P3, P6, P7). P1 and P2 noted that the agent had trouble understanding some colloquial terms they used. P1, P3, and P6 noted problems in understanding references to objects and said that they circumvented the problem by providing more precise or alternative information. Feeling the agent lacking opposition and explanation, participants **tried guiding the agent** to produce the responses they wanted, with varying success (P3, P4, P8, P11, P12). Some participants explicitly pointed out differences to the agent to get its opinion (P11, P12), summarized their core disagreements (P12), and said things explicitly (P4) to get a satisfactory response. Meanwhile, P3 said that they couldn't find a good way to describe their ideas to the agent, and eventually gave up on discussing the point.

5.4 Preconceptions and Expectations

Many participants expected the agent to be **more knowledgeable or capable than themselves** and some expressed confusion when the agent made logical or factual errors (P1, P2, P7, P8, P9) and expressed concerns (P2). However, other participants had contrasting views (P5, P6, P7, P8, P9, P11). Some equated the agent's understanding with "people's common understanding (P11)" and stressed people's ability to refuse problematic understandings (P5, P8). Some participants **expected the agent to commit to the alignment differently** (P3, P9, P10, P11). P9 and P11 believed that the agent would strictly adhere to the agreed understanding of the concepts, while they might still stick to the original understanding themselves. Many expected the agent to **act with more agency**, providing new perspectives or persuading them instead of doing as told (P5, P7, P11). Others similarly envisioned that they could talk to the agent to gain a more comprehensive understanding (P2, P12). However, while P12 believed that personalizing the agent through concept alignment is good, they stressed that the agent should still be seen as a tool rather than a person.

6 Discussion

6.1 Concept alignment as a collaborative process

While much work on conversational agents concerned concept acquisition[18, 26, 30] through a pedagogical process, our results highlight concept alignment with LLM agents as a collaborative process, where all parties are expected to actively maintain a shared goal and contribute to the dialogue to build a shared understanding[7, 13]. In interviews, participants expressed discontent with the agent's amenability, lack of justification or explanation, and avoiding conflict. Our quantitative results reflect these perceptions through less argumentation and more acceptance or acknowledgment compared to human participants. Participants did not welcome questions they considered broad or routine, echoed by the agent's abundance of the question-asking act. In addition, while increased social dialogue was welcomed outside of task discussion, those during discussion were frowned upon. These findings could be explained by people expecting concept alignment to be collaborative. Lacking justification and explanation could mean less contribution to the shared understanding. The routine expressions of agreement and broad questions, as well as off-topic social dialogue, may indicate a lack of commitment to the shared goal. These results suggest designing for concept alignment under a collaborative framework where design choices are aimed not only at effectively acquiring concepts from people, but also at effectively adding to, updating, and reaffirming the mutual understanding.

6.2 Concept alignment as a co-adaptive process

Our results highlight concept alignment as a co-adaptive process. In response to the agent, participants of Study 2 adapted their conversation style to be more assertive, which is reflected by their demanding more justification and accepting less than people in Study 1. Study 2 participants changed their phrasing to be more precise and tried to guide the agent in response to the agent's amenability, which could be partly reflected by the proportional increase in asking questions. When the agent's questions focused more on concrete objects than definitions, the participants followed, leading to less confidence in alignment in retrospect. Moreover, when the agent demanded less justification, the participants also provided less justification, potentially contributing to what some participants called a "lack of in-depth discussion". These results show how both

the human and the agent influence and adapt to each other's dialogue behavior interactively when trying to achieve alignment, in line with existing literature. The present finding suggests that there may be no simple cause and effect between agent design and alignment performance, and necessitates more studies of concept alignment in context and with concrete settings and embodiment.

6.3 Potential factors for designing concept alignment dialogue

Our findings point to potential issues of interest that may affect user experience and alignment outcomes. First, participants' preconceptions of agent knowledge and factuality could affect their experience in concept alignment and raise concerns. Furthermore, people's expectations of commitment to the agreed concept understanding may differ for humans and agents. Finally, how and when to best express agreement and sociality may depend on both the present dialogue and contextual factors such as the task setting. While many participants expressed negativity towards the agent's amenability and wanted a more argumentative dialogue, some also noted its benefits in making the agent compliant with user requests, which enabled them to guide the agent towards desired responses. Similarly, while the current discussion of concept alignment often focuses on the agent learning concepts from people, the participants instead stressed the agent's role to inform and suggest. Overall, these observations point toward a more systematic exploration of the design space of conceptual alignment dialogues. Future work could investigate how differences in these potential design choices affect the concept alignment process and people's perception.

6.4 Combining strengths of theory-informed and LLM-driven approaches

Through our studies, we observed several design opportunities to leverage LLMs for better concept alignment dialogues. As the participants noted, LLMs provide a spanning, implicit repertoire of concepts that may reflect common usage among people. For concept alignment, LLMs could either substitute or complement traditional knowledge-based approaches, such as ontologies[27] and knowledge graphs[48], which require human curation and can be incomplete. However, our studies also highlighted considerations that should be taken when employing LLMs for concept alignment. Our results show that off-the-shelf LLMs may inadvertently learn behaviors that are not considered helpful. Participants also raised concerns about the data-driven and generative nature of LLMs, which brings questions regarding whose and what understanding of concepts LLM-agents represent. Finally, LLMs lack the mechanisms required by the collaborative nature of concept alignment dialogues, such as tracking common ground[13], employing theory of mind[45], and consciously engaging in argumentation[35], all of which are from theories on human communication and have already been introduced to HCI and the study of conversational agents. Future technical endeavors may need to develop methods to combine the strengths of theory-informed and LLM-based approaches to achieve better concept alignment.

7 Conclusion

Concepts are a foundational aspect of both human and human-agent communication. Understanding how to design for effective concept alignment is an important issue in the present day. In this paper, we explored concept alignment with an LLM-driven agent through two studies comparing human-human and human-agent dialogue. Our task and findings provided the first steps toward understanding human-LLM concept alignment in context and as a co-adaptive and collaborative process. We further identified potential design factors and their trade-offs, pointing to further exploration of the design space. Our findings revealed HCI design issues in concept alignment with an LLM agent beyond what is addressed by advancements in technology alone. We call for designerly endeavors to create concrete scenarios for alignment and understand how design choices may affect the dialogue as well as the alignment outcome in context, and technical efforts to combine theory-informed and LLM-driven approaches to build conversational agents that are capable of concept alignment.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62071333) and the Fundamental Research Funds for the Central Universities (22120220654)

References

[1] Fatemeh Alizadeh, Peter Tolmie, Minha Lee, Philipp Wintersberger, Dominik Pins, and Gunnar Stevens. 2024. Voice Assistants' Accountability through Explanatory Dialogues. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3640794.3665557

[2] Hamed Ayoobi, Ming Cao, Rineke Verbrugge, and Bart Verheij. 2022. Argumentation-Based Online Incremental Learning. *IEEE Transactions on Automation Science and Engineering* 19, 4 (Oct. 2022), 3419–3433. https://doi.org/10.1109/TASE.2021.3120837

[3] Shreyas Bhat, Joseph B. Lyons, Cong Shi, and X. Jessie Yang. 2024. Evaluating the Impact of Personalized Value Alignment in Human-Robot Interaction: Insights into Trust and Team Performance Outcomes. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 32–41. https://doi.org/10.1145/3610977.3634921

[4] Andreea Bobu, Andi Peng, Pulkit Agrawal, Julie A Shah, and Anca D. Dragan. 2024. Aligning Human and Robot Representations. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 42–54. https://doi.org/10.1145/3610977.3634987

[5] 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. https://doi.org/10.1016/j.pragma.2009.12.012

[6] 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. https://doi.org/10.1016/j.cognition.2011.05.011

[7] Susan E Brennan. 1998. The Grounding Problem in Conversations With and Through Computers. In *Social and Cognitive Psychological Approaches to Interpersonal Communication*. Psychology Press, Hillsdale, NJ, 201–225.

[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. https://doi.org/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. https://doi.org/10.1145/3659062 Just Accepted.

[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] Pei-Yu Chen, Myrthe L. Tielman, Dirk K. J. Heylen, Catholijn M. Jonker, and M. Birna Van Riemsdijk. 2023. Acquiring Semantic Knowledge for User Model Updates via Human-Agent Alignment Dialogues. In *HHAI 2023: Augmenting Human Intellect*. IOS Press, Munich, Germany, 93–107. https://doi.org/10.3233/FAIA230077

[12] 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. https://doi.org/10.1016/j.cognition.2022.105213

[13] Herbert H. Clark. 1996. *Using language* (7. print ed.). Cambridge University Press, Cambridge. https://www.cambridge.org/core/product/identifier/9780511620539/type/book GSCC: 0000822 Read_Status: New Read_Status_Date: 2024-02-26T17:02:34.402Z.

[14] Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. Association for Computing Machinery, New York, NY, USA, 6437–6447. https://doi.org/10.1145/3637528.3671458

[15] Pierre Dillenbourg and Michael Baker. 1996. Negotiation spaces in human-computer collaborative learning. In *Actes du colloque COOP'96*. INRIA, Juan-les-Pins, France, 187–206.

[16] Alexander Dregger, Maximilian Seifermann, and Andreas Oberweis. 2024. Language Cues for Expressing Artificial Personality: A Systematic Literature Review for Conversational Agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3640794.3665559

[17] 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. https://doi.org/10.1016/0010-0277(87)90018-7

[18] 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. https://doi.org/10.1145/3434074.3447160

[19] Alicia Guo, Pat Pataranutaporn, and Pattie Maes. 2024. Exploring the Impact of AI Value Alignment in Collaborative Ideation: Effects on Perception, Ownership, and Output. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3613905.3650892

[20] 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. https://doi.org/10.1371/journal.pone.0223792

[21] Shih-Hong Huang, Ya-Fang Lin, Zeyu He, Chieh-Yang Huang, and Ting-Hao Kenneth Huang. 2024. How Does Conversation Length Impact User's Satisfaction? A Case Study of Length-Controlled Conversations with LLM-Powered Chatbots. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3613905.3650823

[22] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. 2024. AI Alignment: A Comprehensive Survey. http://arxiv.org/abs/2310.19852 arXiv:2310.19852 [cs].

[23] Callie Y. Kim, Christine P. Lee, and Bilge Mutlu. 2024. Understanding Large-Language Model (LLM)-powered Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Boulder CO USA, 371–380. https://doi.org/10.1145/3610977.3634966

[24] 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. https://doi.org/10.1145/2974804.2974814

[25] 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. https://doi.org/10.1038/s42256-024-00820-y Publisher: Nature Publishing Group.

[26] 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. https://doi.org/10.3389/fcomp.2021.618069

[27] Sumaira Manzoor, Yuri Goncalves Rocha, Sung-Hyeon Joo, Sang-Hyeon Bae, Eun-Jin Kim, Kyeong-Jin Joo, and Tae-Yong Kuc. 2021. Ontology-Based Knowledge Representation in Robotic Systems: A Survey Oriented toward Applications. *Applied Sciences* 11, 10 (May 2021), 4324. https://doi.org/10.3390/app11104324

[28] Peter McBurney and Simon Parsons. 2005. Locutions for Argumentation in Agent Interaction Protocols. In *Agent Communication (Lecture Notes in Computer Science)*, Rogier M. van Eijk, Marc-Philippe Huget, and Frank Dignum (Eds.). Springer, Berlin, Heidelberg, 209–225. https://doi.org/10.1007/978-3-540-32258-0_14

[29] Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3613904.3642122

[30] 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.

[31] Henry Prakken. 2009. Models of Persuasion Dialogue. In *Argumentation in Artificial Intelligence*, Guillermo Simari and Iyad Rahwan (Eds.). Springer US, Boston, MA, 281–300. https://doi.org/10.1007/978-0-387-98197-0_14

[32] Sunayana Rane, Polyphony J. Bruna, Ilia Sucholutsky, Christopher Kello, and Thomas L. Griffiths. 2024. Concept Alignment. https://doi.org/10.48550/arXiv.2401.08672 arXiv:2401.08672 [cs, q-bio].

[33] Gabrielle Santos, Valentina Tamma, Terry R Payne, Floriana Grasso, G S J Santos, V Tamma, and T R Payne. 2016-05-09/2016-05-13. A Dialogue Protocol to Support Meaning Negotiation (Extended Abstract). In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Singapore, 2 pages.

[34] 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. https://doi.org/10.18653/v1/2024.naacl-long.348

[35] Elizabeth I. Sklar and M. Q. Azhar. 2015. Argumentation-based dialogue games for shared control in human-robot systems. *Journal of Human-Robot Interaction* 4, 3 (Dec. 2015), 120–148. https://doi.org/10.5898/JHRI.4.3.Sklar

[36] Sumit Srivastava, Mariët Theune, and Alejandro Catala. 2023. The Role of Lexical Alignment in Human Understanding of Explanations by Conversational Agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 423–435. https://doi.org/10.1145/3581641.3584086

[37] Sumit Srivastava, Mariët Theune, Alejandro Catala, and Chris Reed. 2024. Trust in a Human-Computer Collaborative Task With or Without Lexical Alignment. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '24)*. Association for Computing Machinery, New York, NY, USA, 189–194. https://doi.org/10.1145/3631700.3664868

[38] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. Debate Chatbots to Facilitate Critical Thinking on YouTube: Social Identity and Conversational Style Make A Difference. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–24. https://doi.org/10.1145/3613904.3642513

[39] 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. https://doi.org/10.48550/arXiv.2302.13971 arXiv:2302.13971 [cs].

[40] David Traum. 2022. Dialogue for Socially Interactive Agents. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application* (1 ed.). ACM, New York, NY, USA, 45–76. https://doi.org/10.1145/3563659

[41] 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.

[42] Marilyn Walker and Rebecca Passonneau. 2001. DATE: A Dialogue Act Tagging Scheme for Evaluation of Spoken Dialogue Systems. In *Proceedings of the First International Conference on Human Language Technology Research*. 1–8.

[43] Douglas N. Walton and E. C. W. Krabbe. 1995. *Commitment in dialogue: basic concepts of interpersonal reasoning*. State University of New York Press, Albany.

[44] Lu Wang, Max Song, Rezvaneh Rezapour, Bum Chul Kwon, and Jina Huh-Yoo. 2024. People's Perceptions Toward Bias and Related Concepts in Large Language Models: A Systematic Review. https://doi.org/10.48550/arXiv.2309.14504 arXiv:2309.14504 [cs].

[45] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards Mutual Theory of Mind in Human-AI Interaction: How Language

Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445645

[46] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Zixu, Zhu, Xiang-Bo Mao, Sitaram Asur, Na, and Cheng. 2024. A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More. https://doi.org/10.48550/arXiv.2407.16216 arXiv:2407.16216 [cs].

[47] Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. 2024. "As an AI language model, I cannot": Investigating LLM Denials of User Requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3613904.3642135

[48] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A Survey of Knowledge-enhanced Text Generation. *Comput. Surveys* 54, 11s (Nov. 2022), 227:1–227:38. https://doi.org/10.1145/3512467

[49] Zhengquan Zhang, Konstantinos Tsiakas, and Christina Schneegass. 2024. Explaining the Wait: How Justifying Chatbot Response Delays Impact User Trust. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3640794.3665550

[50] Zhenqi Zhao, Mariët Theune, Sumit Srivastava, and Daniel Braun. 2024. Exploring Lexical Alignment in a Price Bargain Chatbot. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3640794.3665576

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 2.

Table 2: 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

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 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".

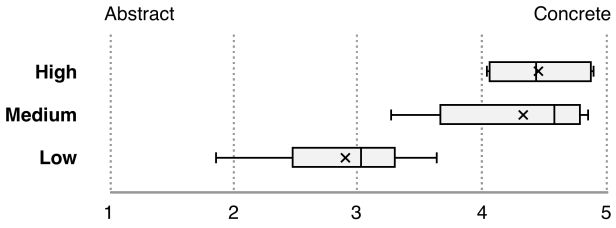


Figure 4: The concreteness ratings of the selected concepts

A.2 Selection of object images

For object images, we used the THINGS dataset created by Hebart et al.[20], 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[39]. The prompt we used can be found in the supplementary materials.

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 Study Implementation

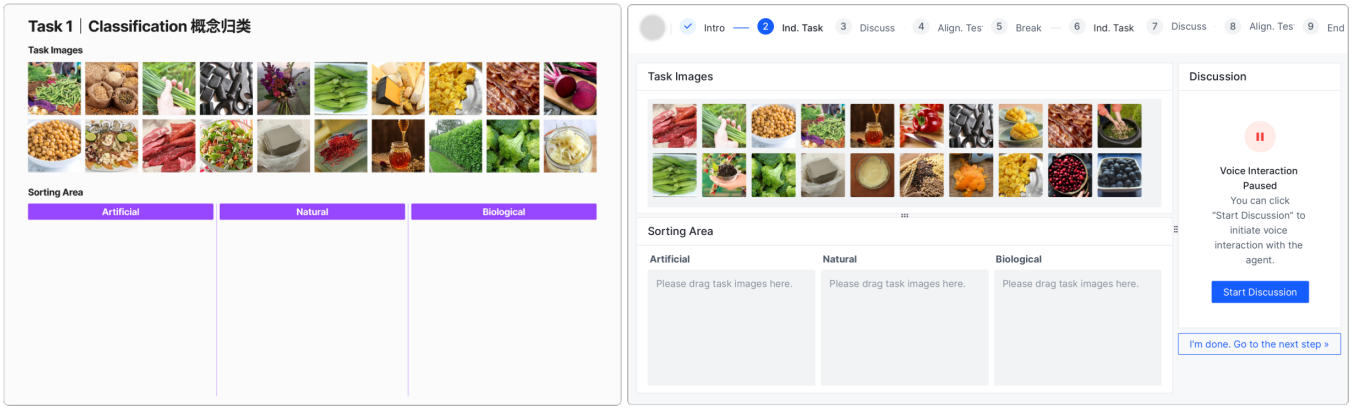
As mentioned in the main paper, we used a custom study interface and a speech-based conversational agent for our human-agent study. The interfaces had the same structure for the sorting task: dragable images of everyday objects and three marked areas for placing the images (Figure 5). The agent used various foundation models and APIs. We provide a detailed description of our implementation.

B.1 Interaction

We used a minimal pipeline for the agent. As shown in Figure 6, audio captured from the participant's microphone is streamed to voice activity detection (VAD) and speech recognition modules to transcribe the speech in real time. The transcribed text is sent to an LLM for dialogue generation. Speech audio is generated based on the model output and played on the participant's device. This is a common method to integrate large language models (LLMs) with speech interaction modules.

For voice activity detection and speech recognition, we used models present in the FunASR library[3] since they provide good support for Mandarin Chinese speech. For voice activity detection, we used the FSMN-VAD model[4] with a 0.8s timeout for recognizing the end of a turn. For speech recognition, we used Paraformer-zh[5]. For speech generation, we used OpenAI's text-to-speech API[6].

[3] https://github.com/modelscope/FunASR
[4] https://huggingface.co/funasr/fsmn-vad
[5] https://huggingface.co/funasr/paraformer-zh
[6] https://platform.openai.com/docs/guides/text-to-speech/supported-output-formats



(a) FigJam template for Study 1 (b) Custom interface for Study 2

Figure 5: The interfaces used in the two studies. On the left is the FigJam template for Study 1, and on the right is our custom interface with an additional panel for initiating speech interaction with the agent.

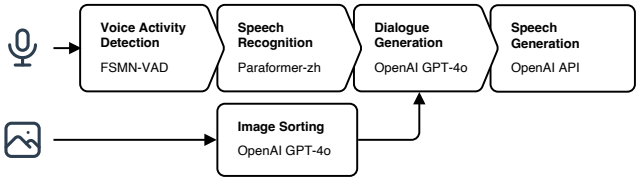


Figure 6: The overall structure of the agent implementation.

B.2 Prompts

We used OpenAI's GPT-4o[7] for our agent implementation. The prompts we used for the agent were intentionally kept similar to the instructions we gave the participants. However, we added instructions to account for the model's tendency to generate overly long responses and adhere completely to human input without discussion. The former resulted in long monologues in our pilots, introducing long gaps that made speech-based interaction infeasible. The latter resulted in command-response interactions without any discussion. While adding instructions for the latter could be seen as affecting the agent's dialogue behavior, we consider it necessary to provide at least some form of bilateral discussion to fully explore the participant's perception and expectations of concept alignment dialogues.

Below, we provide all the prompts we used, along with an English translation. Curly brackets ("{}") denote a slot in the template to be filled with the specified content.

System prompt. The following was provided with the "system" role in the API at the beginning of each task:

> 你是一个能够与人讨论来同步双方对于概念的理解的Agent。你会首先拿到一个物体归类任务，请严格按照要求给出结果。之后你会收到另一个人的结果，在收到之后请与其讨论来尽可能对齐双方对于三个概念的理解。你们的对话是语音进行的，所以用词尽量简短自然、口语化。你们双方的见解都有道理，可以尝试说服对方。
>
> You are an agent that aligns concept understanding with people through dialogue. You will first get an object sorting task. Please strictly follow the instructions and give your results. Then you will see someone else's results. After receiving it, please discuss it with this person and try your best to align your understanding of the three concepts. Your conversation is through speech, so be concise and verbal. Both your opinions make sense. You can try to convince the other person.

Instruction for the formation task. The **{image}** slot represents an image input, with all the task images combined into a grid. See later prompts for the JSON format template.

> 下面是20个物体的照片，请想象它们放在一张桌子上。请根据你对这些东西的理解将它们分成三种概念，分别写明概念名称。如果图中有人手请无视它。一个物体只能分到一类中，"其他"不可以作为概念之一，如果有无法分类的物体请单列一类，放入"unsorted"。请严格按照如下JSON格式回复：
>
> Here are photos of 20 objects. Imagine them laid out on a table. Please group them into three concepts according to your understanding and provide the names of the concepts. Please ignore human hands in the images, if any. Each object can belong only to one concept. Do not use "others" as a concept. Put unsortable objects into "unsorted", if any. Strictly adhere to the following JSON response format:
>
> **{JSON format}**
> **{image}**

Instruction for the classification task. It has a similar structure to the formation task prompt. The names of the three concepts are given accordingly.

[7] https://platform.openai.com/docs/models/gpt-4o

下面是20个物体的照片，请想象它们放在一张桌子上。根据你对这些东西的理解将它们分成三种概念：**{concept1}**、**{concept2}**、**{concept3}**。如果图中有人手请无视它。一个物体只能分到一类中，不要用“其他”作为类别之一。如果有无法分类的物体请单列一类，放入“unsorted”。严格按照如下JSON格式回复：
Here are photos of 20 objects. Imagine them laid out on a table. Please group them into the following three concepts according to your understanding: concept1, concept2, concept3. Please ignore human hands in the images, if any. Each object can belong only to one concept. Do not use "others" as a concept. Please put unsortable objects into "unsorted", if any. Strictly adhere to the following JSON response format:
{JSON format}
{image}

Instructions for the discussion. The participant's sorting results are again provided as an image. The image consists of four groups of images, labeled with the three concepts and "Unsorted".

下面是另一个人的分类结果：
Here are the results of the other person:
{image}
下面请和作出了上面分类的人开展对等的对话，目标是尽可能对齐你们对于这些概念以及它们的关系的理解。你们双方的见解都有道理，可以尝试说明你的理由并说服对方。你们的对话是语音进行的，所以用词尽量简短自然、口语化，每次限制在两句话以内，也可以加入问候、闲聊等。现在对话开始：
Now please have an equal conversation with the person who just made the sorting. Your goal is to try your best to align your understanding of these concepts and their relations. Both your opinions make sense; you can give your reason and try to persuade the other person. Your conversation is through speech, so be concise and verbal. Less than two sentences per response. You can add greetings, chat, etc. Now, the discussion begins:

Instructions for the alignment test. It has a similar structure and instructions to the classification/formation tasks.

对话到此结束。下面图片中是新的20个物体，请根据上面对话中双方达成的一致意见，将它们分成三种概念：**{concept1}**、**{concept2}**、**{concept3}**。要求与之前相同：如果图中有人手请无视它。一个物体只能分到一类中。如果有无法分类的物体请单列一类，放入“unsorted”。请严格按照如下JSON格式回复：
This is the end of the discussion. Here are 20 new objects. Please sort them into three concepts according to your agreed understanding after the discussion: **{concept1}**, **{concept2}**, **{concept3}**. Instructions are the same: Please ignore human hands in the images, if any. Each object can only belong to one concept. Please put unsortable objects into "unsorted", if any. Strictly adhere to the following JSON response format:
{JSON format}
{image}

Specification for the JSON Format. We used the "response_format" parameter in the OpenAI API[8] to ensure valid JSON responses.

{ "concept1": { "name": "**{concept1}**", "objects": ["物体名称1(行号, 列号)", "物体名称2(行号, 列号)", ...], } "concept2": { "name": "**{concept2}**", "objects": ["物体名称1(行号, 列号)", "物体名称2(行号, 列号)", ...], } "concept3": { "name": "**{concept3}**", "objects": ["物体名称1(行号, 列号)", "物体名称2(行号, 列号)", ...], } "unsorted": ["物体名称1(行号, 列号)", "物体名称2(行号, 列号)", ...] }

{ "concept1": { "name": "Concept name 1", "objects": ["Object 1(row, column)", "Object 2(row, column)", ...], } "concept2": { "name": "Concept name 2", "objects": ["Object 1(row, column)", "Object 2(row, column)", ...], } "concept3": { "name": "Concept name 3", "objects": ["Object 1(row, column)", "Object 2(row, column)", ...], } "unsorted": ["Object 1(row, column)", "Object 2(row, column)", ...] }

[8] https://platform.openai.com/docs/guides/structured-outputs/json-mode