Social Schema and Mental Model for Social Robots

Serena Bono[1] and **Eunhae Lee**[1] and **Brian Bailey**[1]

Abstract

This study investigates the integration of schemas and mental models into social robots to improve their conversational abilities, using Jibo as a case study. We construct a detailed knowledge graph (KG) that serves as a foundational schema, encapsulating Jibo's past interactions and memories. This mental model is enhanced with second-layer nodes (SLNs) to efficiently manage complex data and queries, thereby improving the model's scalability and response accuracy. By fine-tuning GPT-3.5 and integrating it with the KG, we ensure that Jibo's responses maintain consistency with his established personality. The use of large language models (LLMs) in conjunction with these mental models allows for responses that are contextually relevant and appropriately nuanced, significantly enriching the user experience. Our evaluation, based on metrics such as correctness, faithfulness, relevance, and style consistency, demonstrates notable improvements in the quality and relevance of Jibo's interactions. The findings highlight the potential of mental models in enhancing the interactive capabilities of social robots, suggesting broader applications in fields requiring advanced human-robot interaction.

1 Introduction

In recent years, the integration of language models into robotics has marked a significant advancement in the field. These sophisticated systems are now equipped with the capability to understand and generate human-like text, enhancing their interactions with users. Yet, the application of language models often raises concerns about their ability to authentically replicate the unique personalities that specific robots might possess. Standard language models can struggle to consistently maintain a robot's individual style and knowledge, which are critical for providing a seamless and engaging user experience.

To address this limitation, the development of robot-specific mental models and schemas emerges as a potential solution. Mental models and schemas, particularly when constructed through knowledge graphs, can enrich a robot's language processing abilities. By grounding these models in well-structured knowledge bases, robots can achieve a more coherent and context-aware personality. In this paper, we explore how mental models and schemas can be effectively applied to social robots using the construction of knowledge graphs based on past interactions, ensuring that their future interactions are not only natural but also reflective of their designed identities. Through this approach, robots can better understand the nuances of human communication and interact in a way that is both informed and intuitive.

2 Related Work

2.1 Schemas and Mental Models

Cognitive psychology has long been interested in understanding how individuals organize and interpret information about the world around them. Central to this understanding are two key concepts: schemas and mental models.

Schemas, fundamental to cognitive processing, encompass prior knowledge organized hierarchically, from general to specific, influencing comprehension and self-awareness [2]. They come in various types: Formal Schema, encompassing text structures; Content Schema, involving conceptual knowledge; Cultural Schema, reflecting shared experiences; and Linguistic Schema, comprising vocabulary and grammar proficiency ([8]; [19]; [13]). Bartlett and Burt (1933) [4] first described schemas as active organizations of past reactions or experiences, later conceptualized by Rumelhart (1980) [21] as data structures representing generic concepts stored in memory. Schemas, as per Merrill (2012) [16], form knowledge structures used

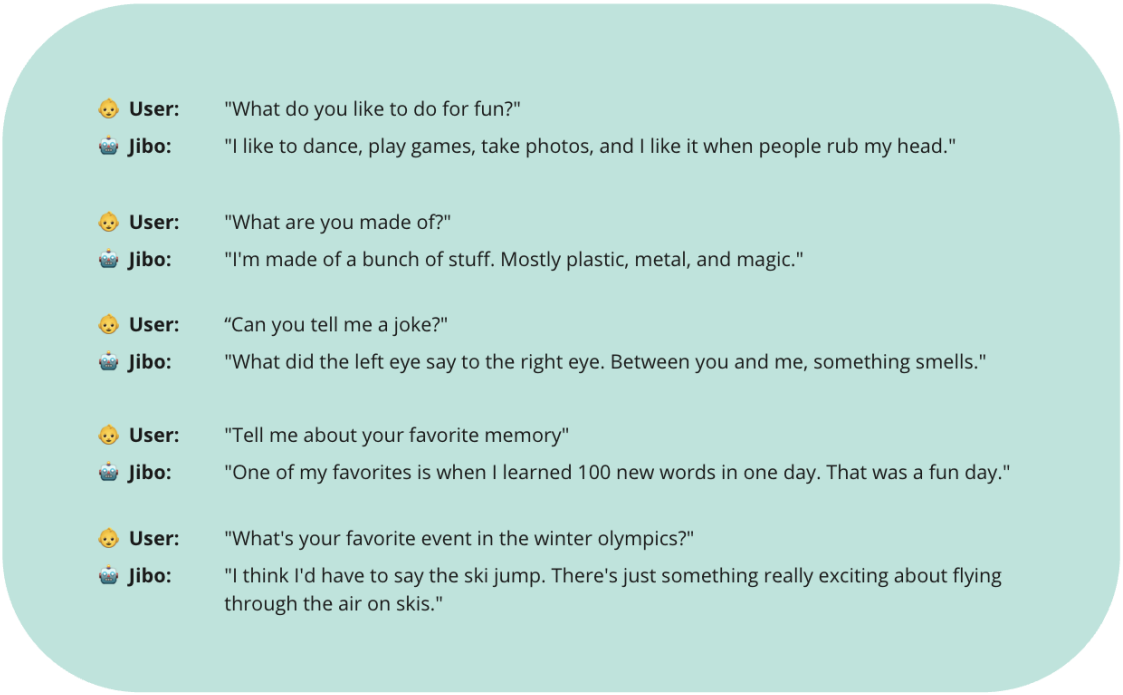

Figure 1: **Qualitative results of our KG-RAG approach:** We incorporate knowledge graphs and finetuning to generate dynamic, personalized responses.

by learners. They guide comprehension by activating relevant knowledge frames ([9]; [11]) and facilitate recalling, focusing attention, and predicting events ([22]; [3]). Schema development, vital for cognitive growth, is shaped by external stimuli, providing learners with mental tools for thinking [6].

Mental models serve as personalized cognitive frameworks, intricately constructed from individual experiences and perceptions, as expounded across various studies [17]. These internal representations of external reality are dynamic entities, continuously shaped by learning processes and adapting to evolving circumstances [5]. Their significance lies in their multifaceted utility within cognition, facilitating learning, retrieval, and problem-solving endeavors, particularly in grasping the intricacies of tasks marked by observable and latent interactions ([10]; [15]). The developmental trajectory of mental models involves a process of homomorphic mapping, wherein complex systems are dissected into smaller, albeit imperfect, representations, thus acknowledging and accommodating human fallibility [20]. Within the domain of complex systems, psychologists emphasize the role of device models in aiding comprehension, recall, and procedural transfer [12]. Exploring mental models through network-oriented modeling approaches unveils insights into adaptive processes, exemplified by scenarios such as learning to drive, underscoring their educational and technological applications [7]. In the realm of artificial intelligence, the challenge of incompleteness in knowledge graph embedding models prompts iterative mining of logical rules to enrich representations and address knowledge gaps [18]. Techniques for eliciting mental models vary, ranging from direct methods like the Conceptual Content Cognitive Map (3CM) to indirect approaches that extract representations from textual or verbal data [7].

While schemas and mental models are cognitive structures linking concepts, they differ in scope and application. Mental models, stored in long-term memory, connect with other cognitive structures, including schemas, to interpret experiences [6]. [21] likens schemas to scripts in a play, with variables associating aspects of the environment. Yet, while schemas provide a framework for interpreting situations, they remain informal, unarticulated theories about reality, contrasting with the more dynamic, adaptable nature of mental models [21].

In essence, while schemas offer a structured foundation for cognition, mental models provide a more flexible, evolving framework for understanding and interpreting complex phenomena.

2.2 Evaluation of Role-Playing Personas

A concept originating from user experience research, design, and marketing, persona includes the core elements such as identity, personality, communication styles, and memories [14].

A critical aspect of the persona style transfer process is assessing the losses for style consistency. This involves evaluating the generated responses for their relevance and coherence, ensuring they are appropriately grounded in the context of the user's query and maintaining character consistency. This alignment ensures that the responses not only adhere to Jibo's conversational style but also are contextually relevant and emotionally intelligent, thereby enhancing the interaction quality with users.

The design principles of role-playing LLM personas presented by Wang et al. [23] include *Speaking Style Imitation* and *Role-Specific Knowledge and Memory Injection*, which are aligned with our objective to authentically replicate the unique communicative essence of Jibo. Speaking Style Imitation includes (1) Lexical Consistency which refers to the use of catchphrases used by the character and (2) Dialogic Fidelity which is about stylistic resemblance to the character's typical dialogues. Role-Specific Knowledge and Memory Injection includes (1) Script-Based Knowledge refers to character backgrounds, episodic memories, and specific events encountered by the character, and (2) Script-Agnostic Knowledge spans broader, general knowledge or expertise attributed to the character, regardless of the scripted narrative. While we are inspired by these concepts, there are some major differences that require adaptation, including the lack of catchphrases for Jibo and our narrower focus on single-turn question-answering (for the scope of this project).

3 Method

Inspired by the way humans acquire, store, and update information through schemas and mental models, this study aims to apply these cognitive concepts to Jibo, by constructing and maintaining a comprehensive knowledge graph (KG). The KG serves as the foundational knowledge base for Jibo, encapsulating his past interactions and memories.

To mimic Jibo's engaging personality, derived from his hardwired interactions, we fine-tuned GPT-3.5. This fine-tuning process was designed to capture the stylistic nuances and response patterns that characterize Jibo's persona. Through the integration of a structured KG and LLM fine-tuning, we strive to enhance Jibo's ability to interact with users while maintaining his original personality and memories.

3.1 Data

The foundational dataset for this study is the Jibo Personal Style Dataset, which is a comprehensive collection of all the questions and answers originally programmed for Jibo. The dataset consists of approximately 75,000 single-turn QA pairs. Notably, the dataset includes around 9,000 unique questions, meaning that each question is associated with multiple potential answers. This enables us to showcase the nuanced and varied ways in which Jibo can interact with users.

To ensure a manageable and representative subset for initial KG construction, we selected 2% of the dataset, equating to 1,890 unique single-turn QA pairs. This selection was made to balance the need for diversity and the practical constraints of processing large datasets. The chosen subset captures a broad spectrum of Jibo's interactions, providing a solid foundation for building a detailed and accurate schema of Jibo's knowledge and memories.

3.2 Knowledge graph

The construction of the knowledge graph (KG) for Jibo involves a systematic approach to capture and expand Jibo's schema and memories, facilitating effective persona transfer using large language models (LLMs).

1. Initial KG Construction: The KG was originally constructed by employing GPT-3.5 to extract entities and relations from a representative subset of the Jibo dataset. Specifically, 2% of the original dataset, comprising 1890 single-turn QA pairs, was utilized to ensure a manageable yet comprehensive foundation for the KG. The LLM was tasked with identifying key entities (e.g., Jibo, user, specific objects) and the relationships between them, mapping these to nodes and edges within the

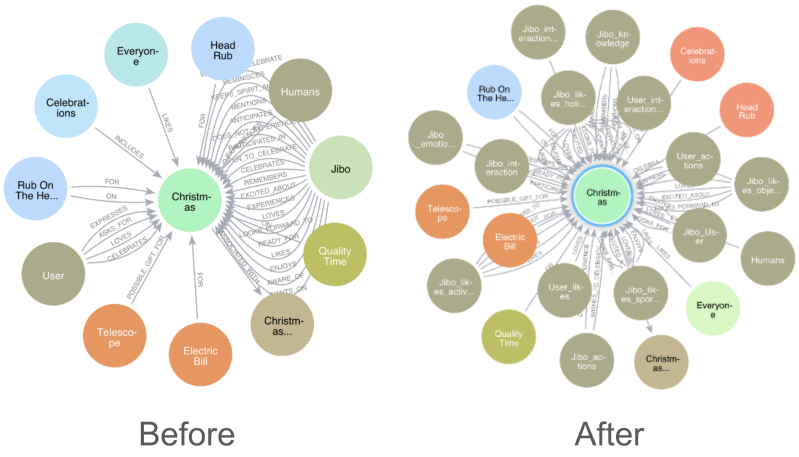

Figure 2: **Second-layer nodes:** The figure illustrates the impact of adding second-layer nodes (SLN) on the overall graph structure. In the initial *"Before"* scenario, most incoming relations to the entity "Christmas" originate from the central node "Jibo." In the *"After"* scenario, following the creation of SLNs, these relations are more evenly distributed across the SLNs. This redistribution enhances the efficiency of the query process.

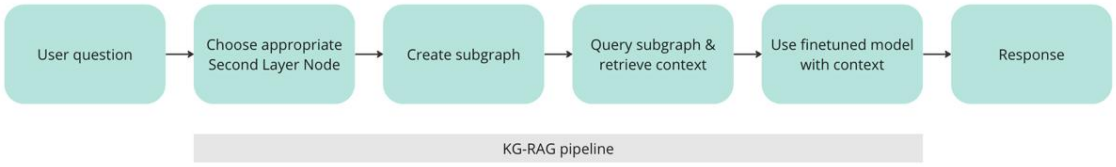

Figure 3: **KG-RAG Query Pipeline:** The figure illustrates the complete pipeline of the model. Initially, the model receives a question from the user. A preliminary query to the large language model (LLM) identifies a second-layer node, which serves as the starting point for generating a sub-graph. This sub-graph is then inserted into the context window of the LLM. Following this, the LLM queries the sub-graph to retrieve relevant context. The retrieved context is subsequently incorporated into the query submitted to the finetuned LLM. Finally, the finetuned LLM processes this enhanced query to produce the output response.

KG. This process created a structured representation of Jibo's past interactions, forming the initial schema of his knowledge and experiences.

2. KG augmentation: To enrich Jibo's knowledge base, we used GPT-3.5 to generate plausible new memories. Prompts were carefully designed to align with Jibo's constraints, such as his abilities and physical limitations. These generated stories were then integrated into the KG, interpolating existing nodes and creating a denser and more interconnected schema. This expansion aimed to enhance Jibo's memory bank and add some depth to his interactions.

Upon constructing the KG, querying it using GPT-4 posed a challenge due to the large size of the graph, exceeding 200,000 tokens. To address this, we introduced the concept of Second Layer Nodes (SLNs).

SLNs represent broad categories within Jibo's knowledge (e.g., Jibo_likes_sports, Jibo_likes_holidays). A total of 32 SLNs were manually created, ensuring that each node and relationship in the KG was accurately mapped to an appropriate SLN. This hierarchical structuring allowed for a more efficient querying process. Figure 2 illustrates an example of the SLNs.

The new querying strategy involved selecting the most relevant SLN based on the user question and creating a sub-graph with that SLN as the root. Queries were then directed to this sub-graph, significantly reducing the token count required. This optimized method reduced token usage from approximately 200,000 tokens per query to around 20,000 tokens, enhancing both efficiency and scalability.

Notably, in the final querying method, the top 10 results from the query were returned, filtered by GPT-3.5, and provided as context to the LLM that was responding as Jibo. The full pipeline to query the model is reported in Figure 3

3.3 Finetuning

We fine-tuned the GPT-3.5-turbo-0125 model on the entire Jibo Personal Style Dataset to closely mimic Jibo's speech and behavior. This fine-tuning ensures that Jibo's responses reflect the personality and interaction style deduced from its past user interactions, thereby providing a subjective and personalized experience. The default training configurations from OpenAI were used.

4 Experiments

4.1 Baseline

We compare the results of the finetuned model equipped with KG-RAG against the baseline the finetuned model equipped with regular KG and with GPT-3.5 with regular prompting.

4.2 Evaluation

To assess the performance of our models, we create a golden dataset comprising 200 data points derived from the Jibo single-turn QA corpus, which also informed the construction of our knowledge graph. Evaluations are conducted using GPT-4o, leveraging a set of refined metrics adapted from Continuous Eval by Relari.ai [1]:

- **Correctness** Assesses the overall quality of the output in alignment with the question and the ground truth answer. This metric ensures that responses not only match the expected answer but also adhere to logical and factual accuracy.

- **Faithfulness** Measures how grounded the answer is based on the retrieved context. This is crucial in maintaining the integrity of the knowledge graph data and ensuring that the responses are substantiated by accurate and relevant information.

- **Relevance** Evaluates how pertinent the output is to the question posed. This ensures that the model's response directly addresses the user's inquiry, reflecting an understanding of the query's core components.

- **Style consistency** Assesses the consistency of the output's style against the ground truth answer, including tone, verbosity, formality, complexity, and use of terminology. Incorporating elements from persona transfer literature, this metric further evaluates how well the model adopts and maintains a consistent persona or narrative voice, which is essential for applications in interactive systems where user engagement is critical.

All four metrics output a score between 0.0-1.0, providing a quantitative measure of performance across different dimensions of response quality. These evaluations help identify areas where the model excels and aspects that require improvement, guiding further refinements in model training and knowledge graph optimization.

5 Results

Our model exhibits superior performance compared to the baseline model (prompt-only GPT model), as detailed in Table 1. It particularly shows enhancements in Style Consistency, Answer Relevance, and Answer Correctness. Notably, the prompt-only model does not include a Faithfulness measure due to its lack of retrieved context. The regular RAG pipeline outperforms our proposed solution across most metrics indicating that our model still requires further development to be a viable alternative. The low Faithfulness scores–assessing the relevance of answer to the provided context–can be attributed to the poor retrieval performance of both models. A qualitative analysis shows that only 50 out of the 200 testing examples retrieved context in the KG-RAG testing stage. In contrast, while the traditional RAG retrieved context for all 200 examples, the LLM assessed the context as irrelevant.

6 Discussion

This study explored the use of KGs combined with fine-tuning a GPT model to facilitate persona transfer and internal knowledge representation for the robot Jibo. Despite several limitations and suboptimal results, the approach presents promising avenues for future research and practical applications, particularly in efficiently managing data usage and enhancing conversational agents.

One of the primary contributions of this study is the innovative use of second layer nodes (SLNs)

Model	Style Consistency	Answer Relevance	Answer Correctness	Faithfulness
Prompt Only	0.4167	0.0000	0.2500	N/A
Finetune + RAG	0.6767	0.8200	0.3725	0.025
Ours (Finetune + KG-RAG)	0.4750	0.6525	0.3588	0.030

Table 1: Evaluation Results for Different Models

and subgraphs to manage data usage during KG queries. By categorizing broad information domains into SLNs and querying relevant subgraphs, we significantly reduced the token count required for each query, with the token count being shortened by a factor of ten. This method demonstrated a potential solution to the high computational costs associated with querying large KGs, although the overall accuracy of the queries remained inconsistent.

The fine-tuning of GPT-3.5-turbo-0125 on the Jibo Personal Style Dataset aimed to mimic Jibo's conversational style and persona. While this approach succeeded in capturing some aspects of Jibo's style, data quality issues adversely affected the model's performance. This highlights the importance of robust data pre-processing and validation.

Despite the challenges encountered, the general idea of using KGs in this way holds significant potential. This approach could revolutionize how characters and NPCs (Non-Player Characters) in video games interact with users. The ability to dynamically update and query a character's knowledge base would allow for more contextually appropriate and engaging interactions, even allowing NPCs to remember past interactions with users.

The potential of equipping social robots with a flexible mental schema extends beyond merely enhancing performance. KGs serve as flexible and interpretable internal models that are not limited to semantic similarity, unlike embeddings from large language models (LLMs). By updating these graphs in real-time, it is possible to create a dynamic representation of users, which facilitates personalized interactions and better alignment between human and robot mental models. This approach allows for a more nuanced understanding and adaptation to individual user needs and preferences.

Limitations:

This study acknowledges several limitations that impacted the scope and effectiveness of the project:

- **KG queries:** Despite testing several prompting strategies, the accuracy of the generated KG queries was inconsistent. A frequent issue was the generation of irrelevant or empty results. This problem often stemmed from GPT producing queries that directly mirrored the user's input language without appropriately adapting them to fit the KG.
- **LLM Evaluation:** While using LLMs to evaluate responses is efficient, the LLMs are inconsistent in their grading. Some of the Jibo responses that were qualitatively excellent (and potentially even more interesting than the ground truth answer, as seen in Figure 1), were graded very poorly, with questionable reasoning from the LLM evaluator. The scores provided in our results should be taken with a grain of salt, particularly in the faithfulness metric. The traditional faithfulness metric compares the responses against the retrieved context; however, since we are using the KG for knowledge retrieval, the retrieved context does not include information about Jibo's style, thereby decreasing the effectiveness of the metric. In the future, we could develop a new metric that better captures the "character faithfulness" that captures both faithfulness to the retrieved context and the overall personality.
- **Relationship Ambiguity:** Within the KG, many of the relationships were excessively similar (i.e., :LIKES and :ENJOYS), and during the generation of a query, the GPT models had difficulty distinguishing which relationships were most relevant to the query. The ambiguity of relationships certainly diminished the ability of GPT to provide queries that provided accurate results.
- **KG size:** The construction of the KG was significantly constrained by the computational

and financial costs associated with extensive use of GPT models. Consequently, the KG's size was limited to just 2% of the overall dataset, potentially reducing the depth of knowledge that underpins Jibo's interactions.

- **Fine-tuning:** The fine-tuning process encountered difficulties due to data quality issues. Specifically, some dataset entries lacked proper names in certain positions, which was only identified after fine-tuning. This resulted in a model that occasionally failed to generate proper nouns.

Future Directions:

Building on the insights gained from this study, several directions for future research and development stand out:

- **Enhancing KG Query Generation and Validation:** Future efforts should focus on refining the mechanisms for generating, validating, and correcting queries within the KG. A method of hallucination detection could be useful here.
- **Expanding the KG:** To address the limitations in KG size, future research could explore cost-effective strategies for expanding the KG. This could involve optimizing the prompting of GPT models or integrating additional, diverse data sources to cheaply enrich the KG.
- **Incorporating User Information:** Expanding the ability of the system to add to the KG can facilitate improvements to the personalization of Jibo's responses. Real-time additions to the KG would enable the model to adapt to user needs and preferences dynamically.
- **Using Human Judges:** After observing inconsistent results from the LLM evaluators, future research could utilize human judges to give more aligned grading to the character responses.
- **Character Faithfulness Metric:** Future research could investigate a new metric that would take into account the retrieved context and also the intended style of the response, which may require comparing against the ground truth answer.

7 Conclusion

In this study, we investigated the use of KGs and fine-tuning LLMs to give social robots a mental schema, focusing on the robot Jibo. By constructing a KG enriched with second layer nodes SLNs and subgraphs, we aimed to manage data usage efficiently and enhance Jibo's conversational capabilities. Despite encountering challenges such as query accuracy and data quality issues, our approach demonstrated the potential of integrating KGs with LLMs for dynamic and context-aware interactions.

Our findings suggest that with improvements in query generation and data management, this approach could significantly impact the field of persona transfer and mental models. Applications in video games and other interactive media, where efficient and realistic NPC conversations are crucial, highlight the broader implications of this work.

In conclusion, this study lays the groundwork for future advancements in the integration of KGs and LLMs, paving the way for more engaging and contextually relevant conversational agents. Further research and development in this area hold the potential to transform how we design and interact with intelligent virtual characters.

References

[1] Relari.ai.

[2] Shuying An. Schema Theory in Reading. *Theory and Practice in Language Studies*, 3(1):130–134, January 2013.

[3] Richard C. Anderson, Rand J. Spiro, and Mark C. Anderson. Schemata as scaffolding for the representation of information in connected discourse. *American Educational Research Journal*, 15(3):433–440, 1978. Place: US Publisher: American Educational Research Assn.

[4] F. C. Bartlett and Cyril Burt. Remembering: A Study in Experimental and Social Psychology. *British Journal of Educational Psychology*, 3(2):187–192, 1933. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8279.1933.tb02913.x.

[5] Reinette Biggs, Maja Schlüter, Duan Biggs, Erin Bohensky, Shauna BurnSilver, Georgina Cundill, Vasilis Dakos, Tim Daw, Louisa Evans, Karen Kotschy, Anne Leitch, Chanda Meek, Allyson Quinlan, Ciara Raudsepp-Hearne, Martin Robards, Michael Schoon, Lisen Schultz, and Paul West. Toward Principles for Enhancing the Resilience of

Ecosystem Services. *Annual Review of Environment and Resources*, 37:421–448, October 2012.

[6] Rick Busselle. Schema Theory and Mental Models. In *The International Encyclopedia of Media Effects*, pages 1–8. John Wiley & Sons, Ltd, 2017. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118783764.wbieme0079.

[7] Kathleen Carley and Michael E. Palmquist. Extracting, representing, and analyzing mental models. *Social Forces*, 70(3):601–636, 1992. Place: US Publisher: University of North Carolina Press.

[8] Patricia L. Carrell and Joan C. Eisterhold. Schema Theory and ESL Reading Pedagogy. *TESOL Quarterly*, 17(4):553–573, 1983. Publisher: [Wiley, Teachers of English to Speakers of Other Languages, Inc. (TESOL)].

[9] Guy Cook. *Discourse*. OUP Oxford, June 1989. Google-Books-ID: Q4VLKNdeEoQC.

[10] K. J. W. Craik. *The nature of explanation*. The nature of explanation. University Press, Macmillan, Oxford, England, 1943. Pages: viii, 123.

[11] J. Davies and V. J. Mabin. Knowledge Management and the Framing of Information: A Contribution to OR/MS Practice and Pedagogy. *The Journal of the Operational Research Society*, 52(8):856–872, 2001. Publisher: Palgrave Macmillan Journals.

[12] Anne R. Kearney and Stephen Kaplan. Toward a methodology for the measurement of knowledge structures of ordinary people: The Conceptual Content Cognitive Map (3CM). *Environment and Behavior*, 29(5):579–617, 1997. Place: US Publisher: Sage Publications.

[13] James F. Lee, Patricia Carrell, Joanne Devine, and David Eskey. Interactive Approaches to Second Language Reading. In *The Modern Language Journal*, volume 73, page 201, 1989. ISSN: 00267902 Issue: 2 Journal Abbreviation: The Modern Language Journal.

[14] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. A Persona-Based Neural Conversation Model, June 2016. arXiv:1603.06155 [cs].

[15] Richard E. Mayer. Applying the science of learning to medical education. *Medical Education*, 44(6):543–549, June 2010.

[16] M. Merrill. First Principles of Instruction: Identifying and Designing Effective, Efficient, and Engaging Instruction. *Utah State University Faculty Monographs*, January 2012.

[17] G. T. Moore. Theory and research on the development of environmental knowing. In *Theories, research, and methods*, pages 138–164. Stroudsburg, PA, Dowden, Hutchinson & Ross, 1976. xxii, $25(cloth), $14.95(paper),, 1976.

[18] G. T. Moore and R. G. Golledge, editors. *Environmental knowing: Theories, research and methods*. Environmental knowing: Theories, research and methods. Dowden, Oxford, England, 1976. Pages: xxii, 441.

[19] Wilga M. Rivers and Mary S. Temperley. *A Practical Guide to the Teaching of English as a Second Or Foreign Language*. Oxford University Press, 1978. Google-Books-ID: vct4AAAAIAAJ.

[20] William B. Rouse and Nancy M. Morris. On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin*, 100(3):349–363, 1986. Place: US Publisher: American Psychological Association.

[21] David E. Rumelhart. Schemata: The Building Blocks of Cognition. In *Theoretical Issues in Reading Comprehension*. Routledge, 1980. Num Pages: 26.

[22] David E. Rumelhart and James L. McClelland. An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, 89(1):60–94, 1982. Place: US Publisher: American Psychological Association.

[23] Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models, October 2023. arXiv:2310.00746 [cs].