TU Delft

Delft University of Technology

Human-Agent Co-Construction of Episodic Memories

Kniele, Annika; Donatelli, Lucia; Oertel, Catharine; Vossen, Piek

DOI
10.3233/FAIA250640

Publication date
2025

Document Version
Final published version

Published in
HHAI 2025 - Proceedings of the 4th International Conference on Hybrid Human-Artificial Intelligence

Citation (APA)
Kniele, A., Donatelli, L., Oertel, C., & Vossen, P. (2025). Human-Agent Co-Construction of Episodic Memories. In D. Pedreschi, M. Milano, I. Tiddi, S. Russell, C. Boldrini, L. Pappalardo, A. Passerini, & S. Wang (Eds.), *HHAI 2025 - Proceedings of the 4th International Conference on Hybrid Human-Artificial Intelligence* (pp. 228-237). (Frontiers in Artificial Intelligence and Applications; Vol. 408). IOS Press. https://doi.org/10.3233/FAIA250640

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

HHAI 2025
D. Pedreschi et al. (Eds.)
© 2025 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA250640

Human-Agent Co-Construction of Episodic Memories

Annika KNIELE [a,1], Lucia DONATELLI [a] Catharine OERTEL [b] and Piek VOSSEN [a]
[a] *CLTL, Vrije Universiteit Amsterdam*
[b] *II, TU Delft*
ORCiD ID: Annika Kniele https://orcid.org/0009-0007-4545-8065, Lucia Donatelli https://orcid.org/0000-0002-5974-7454, Catharine Oertel https://orcid.org/0000-0002-8273-0132, Piek Vossen https://orcid.org/0000-0002-8273-0132

Abstract.
We design a task to help identify how an agent can engage with information in a meaningful way through dialogue to foster collaboration. Specifically, the task involves a human and an agent sharing their memories of past events with each other, resulting in diverse information about those events. In a pilot study, we explore to what extent an LLM can be used to classify memories from the different sources as overlapping, complementary or conflicting. Knowing which of these categories a piece of information falls into will aid the agent in how to address it in dialogue, for instance to ask for further information, to adopt a shared perspective, or to agree to disagree about a conflict. We find that the LLM especially struggles with distinguishing between complementary and conflicting information, and that differing opinions about what is and is not implied by the event descriptions lead to many disagreements between the LLM and our human annotators. In future work, we will investigate to what extent conversing with the human can alleviate these issues.

Keywords. human-AI collaboration, dialogue, episodic memory, generative AI

1. Introduction

People's memories of temporally and spatially grounded past events, their *episodic* memories [1], can differ extensively from each other. From a shared trip to the beach, I may recall a dog chasing a Frisbee, while you may remember your child searching for sea shells in the sand. Through conversation, we can collaboratively reconstruct the trip, each contributing our own recollections. As AI systems—especially LLMs—rapidly improve, a key question is how well they can engage in shared remembering with humans. Possible applications of this capability include helping humans reflect on their personal lifestyle [2] and aiding those in need of mental health support [3].

When equipped with an external knowledge store, such as a knowledge graph or a vector database, AI systems can save a nearly unlimited amount of information about past events. This contrasts with humans, whose memory loses or distorts information

[1] Corresponding Author: Annika Kniele, a.kniele@vu.nl

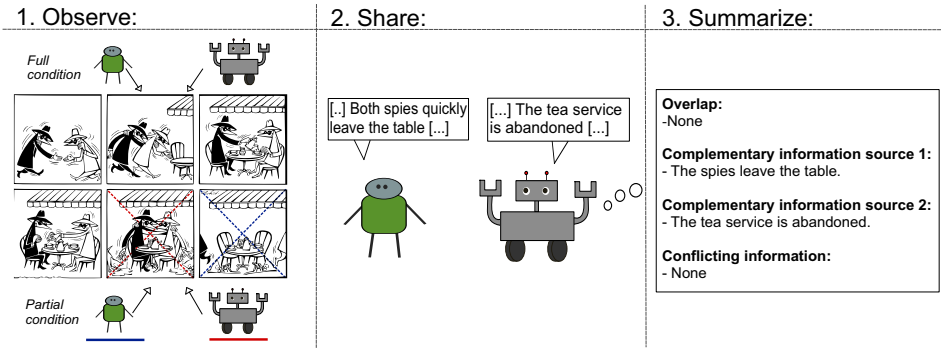

Figure 1. Human-agent memory co-construction. Step 1: The agent and the human observe the same events (or in *partial* condition: most of the same events). Step 2: They share their memories of the events with each other. Step 3: The agent summarizes the events, taking into account whether the information is overlapping, complementary, or conflicting. As part of a conversation between the human and the agent, step 2 and 3 are repeated as often as needed, or until all memories have been shared and added to the summary.

about past experiences, by for instance simply forgetting, or due to our current mood [4]. An AI system, by contrast, may make misinterpretations due to not understanding the context of a memory as well as having insufficient social intelligence, which a human is more likely to have. As both human and agent memories have strengths and weaknesses, leveraging them in hybrid collaboration will allow us to construct a more "complete" picture of past events.[2] We acknowledge that "completeness" is a context-dependent concept, meaning what is relevant to include in a "complete" picture of past events is highly dependent on the goals of the collaborative rememberers. In general, however, we define *completeness* as consisting of three aspects: First, we expect the agent memories to provide observable details the human may have forgotten. Second, the human can contribute emotional information about the memories. Third, the human and agent can both contribute their view on the events by verbalizing their memories in a certain way. If a human and an agent should collaborate, they need to figure out which parts of their memories correspond to each other and which parts are augmentations.

The aim of this paper is to introduce this task of humans and agents *co-constructing* memories of past events (Section 3.1). The task is shown in Figure 1, where the human and the agent observe the same events, but verbalize very different aspects of their memories. Further, we investigate the feasibility of using an LLM for the following important subtask: Can the agent classify the human and agent memories based on how they correspond to each other? The different types of correspondences we consider are *overlap* (Is the piece of information present in both accounts?), *complement* (Is it present in only one of the accounts and does not conflict with any information in the other account?), and *conflict* (Does it conflict with the other account?). Knowing which of these categories a piece of information falls into will aid the agent in how to address it in dialogue, to see if a conflict can be resolved, or to agree to disagree. Through an annotation task[3] we gain insights into how the categories may be interpreted differently, between the annotators and the LLM, and within pairs of annotators. This type of open-ended annotation

[2]Of course, part of the higher level of completeness expected from this hybrid collaboration is due to the human and the agent not being able to be everywhere at all times and therefore having partial memories.

[3]The code used for this pilot study is available at: https://github.com/akniele/memory_co-construction_pilot.

is what [5] call *descriptive* annotation, where surveying different beliefs is central rather than making sure annotators agree on a predefined belief.

Section 2 situates our study in previous research while Section 3 provides details on the proposed task, the requirements for our agent, the data as well as the annotation task used for evaluation. Section 4 discusses the label and error distributions in the annotation task, including examples. Section 5 concludes and provides suggestions for future work.

2. Background

2.1. Collaborative Memory in Human-Human Interaction

In human-human interaction, *collaborative memory* [6] [7] [8] refers to the act of remembering together with other people. It is seen as a social activity that can strengthen interpersonal bonds [9]. Similarly, in human-agent interaction, the agent having episodic memory and communicating about it with the human has been shown to be useful for maintaining long-term relationships with humans [10] [11] [12].

A challenge we foresee is the human and the agent interpreting and verbalizing their own memories in very different ways, even when the underlying events are the same. Both perceiving stimuli from the outside world and verbalizing the experiences we have constructed from them are seen as interpretive processes [13]. This means that neither our perceptions nor how we verbalize them result in "accurate" copies of the original stimuli. The different ways one can communicate about the same situation are referred to as *construals* by [14], positing that the speaker has a choice in how they construct a situation when talking about it. For instance, they can highlight certain aspects of the situation, such as the human in Figure 1 focusing on the spies' actions, while the agent focuses on what happens to the tea service. As agents are expected to perceive the world very differently from humans, the discrepancy in how humans and agents construe events is expected to be even larger than between humans.

2.2. The Categorization Task in the Literature

The categorization task is related to Semantic Overlap Summarization [15] and Comparative Summarization [16], which focus on generating summaries containing only the overlap, or highlighting differing information, respectively. Further, the categories *overlapping*, *complementary* and *conflicting* we defined above are similar to the categories *Redundancy*, *Complement* and *Contradiction* used in Cross-Document Structure Theory [17]. Also related to this task is research on knowledge conflicts in LLMs. [18] define three categories of knowledge conflicts: context-memory, inter-context, and intra-memory conflict. Relevant for us are inter-context conflicts, which arise from conflicting information being provided to the model through a prompt. Investigation of this type of conflict can be seen as a subtask of our categorization task, with the memories of the human and the agent being the two potentially conflicting contexts presented to the LLM. [19] find that, when presented with conflicting contexts, LLMs struggle to identify exactly which information is conflicting, and to present the distinct conflicting viewpoints in their answer. Since we are asking the LLM to categorize all provided information, it is of interest which categories it will group conflicts under, if not "conflicting".

While the tasks mentioned above focus purely on static information, human-agent memory co-construction is an interactive, communicative task in which new information acquired through conversation with a human has to be dynamically integrated into the existing information. To our knowledge, leveraging LLMs and an interactive approach to combine human and agent memories to arrive at a more complete picture of real-world past events is a novel setup, connecting LLMs with Hybrid Intelligence [20] principles.

3. Methodology

3.1. Task Overview - Human-Agent Memory Co-construction

We operationalize memory co-construction between a human and an agent using three steps (see Figure 1). In step 1, an agent and a human both observe, and build memories of, the same event. In step 2, they share the memories with each other through conversation. In step 3, the agent summarizes the verbalizations of the memories, categorizing the information as overlapping, complementary, or conflicting. Steps two and three can be repeated, as the agent iteratively improves the summary by conversing with the human.

In this paper, we focus on the categorization step (step 3) as defined above. We further introduce two task conditions for this pilot study: In the *full* condition, both human and agent observe the same events. This condition can tell us whether the agent can fulfill its task under ideal conditions. In the *partial* condition, parts of the events to be observed are withheld from the agent and the human (as illustrated by the crossed-out comic panels in Figure 1), to investigate more realistic conditions, acknowledging that humans and agents rarely perceive events identically (due to for instance looking away).

3.2. Requirements for a Memory Co-construction Agent

The requirements an agent needs to fulfill to engage in our task are listed in Table 1. The first three are central to this study, as without them the agent would not be able to create a summary as described in Section 3.1. The other three requirements are needed for conversing with a human to improve the summary. A *dialogue move* as mentioned in the *Dialogue Strategy* requirement could be asking the human for more information on a previously discussed topic, or asking the human to judge if two pieces of information are conflicting. For the latter type of move, it is important that the agent is able to judge when to delegate a judgment to the human. The *Adaptivity* requirement could include adopting terms the human uses when making references, such as calling a woman playing the piano “the pianist”. This type of convention-forming is typical for human-human conversations [21] [22] and has previously been studied for dialogue systems [23] [24].

3.3. Data

To investigate to what extent an LLM can fulfill requirements two and three, we need data with a similar structure to verbalized event memories. The Mementos dataset [25], created as a benchmark for evaluating multimodal LLMs’ reasoning over image sequences, includes wordless comic strips with accompanying human descriptions of the events in the comics. For our pilot study, we use five of the comics including the descriptions. These comics are useful as a proxy for verbalized memories. First, like real-world events,

Table 1. Overview of the capability requirements for our agent

Requirement	Explanation
1. Own Memories	Having its own memories of the events in question.
2. Overlap vs. Non-Overlap	Judging which parts of its own memories correspond to what the human told it as well as what does not correspond.
3. Complement vs. Conflict	Judging which non-corresponding pieces of information can reasonably co-exist within a coherent world (i.e. do not conflict).
4. Judge Quality	Judging the summary quality, to decide whether to improve it further.
5. Dialogue Strategy	Having a strategy for choosing its next dialogue move, with the goal of improving the summary.
6. Adaptivity	Adapting its way of construing events to the human's.

the comics contain dynamic information such as causal relationships that the agent needs to interpret. Second, the comics center around people (and personified animals) and their lives, which is what a human's personal memories would likely also focus on.

In addition to the human descriptions taken from Mementos (representing the human's memories in Step 2 of Figure 1), we generate LLM descriptions of the same comics using the Gemini API [26] (representing the agent's memories), using the prompt used by [25]. We further have Gemini generate summaries detailing the overlapping, complementary, and conflicting information between the human and the LLM description (Step 3 in Figure 1). For the *partial* condition, we manually remove a random panel from each comic (not the same one for the human as for the agent). We also remove all information from the human description that can only be derived from the missing panel. Since the information in the human descriptions is in chronological order, it is simple to manually inspect it and correlate it with specific panels. Further, we generate separate LLM descriptions using only the remaining panels.[4]

3.4. Evaluation

To evaluate to what extent an LLM can categorize event information coming from agent and human memories into the categories *overlap*, *complementary 1*, *complementary 2* and *conflict*[5], we ask human annotators to identify and label errors the LLM made. Table 2 shows the possible labels. The annotation tool is a simple web interface allowing annotators to define and label spans, create relations between spans (if multiple spans are part of the same error), and provide an explanation if they chose the label *other*. The start and end of the spans is not predefined; annotators both define and label spans.

There are four annotators in total. As we have ten samples (five per condition), the annotators are assigned five samples each, such that each sample is annotated by a pair of annotators. The annotators are all non-experts, with at least basic linguistic training. They have been recruited through the authors' personal network. Before starting the task, the annotators are given annotation guidelines which they are asked to read and can ask

[4] A limitation of our study is that due to Gemini's closed source status, we cannot know if the Mementos corpus was in its pre-training data. If it was, this may have had an effect on the model's performance in the *partial* condition, as it could have used internal knowledge about the comics to better interpret the incomplete comics presented to it. Investigating this further is left for future research.

[5] Complementary 1 and 2 refer to information only present in the first event description (the human description) or the second (the LLM-generated description), respectively.

Table 2. Overview of the labels annotators can use to annotate errors.

Label Name	Explanation
Should be in xy	Used to show that a piece of information is in the wrong category of the summary. For instance, the label *should be in overlap* should be used if both source texts mention a detail that is wrongfully included in the *complementary* information section.
Missing in xy	Used for information that is present in the source texts but missing from the summary. For example, if the first source text mentions a detail that is not present in the summary at all, the label *missing in complementary 1* should be applied.
Hallucination	Used when a piece of information is present in the summary but can be found in neither of the two source texts.
Other	Used when annotators find an error but feel that none of the other labels fit. When using this label, a short explanation must be provided.

questions about. The data steward of the responsible faculty advised us that due to the non-personal nature of the annotation task, ethical approval was not needed.

4. Results

In this section, we first examine the annotation label distribution and inter-annotator agreement, followed by a closer look at the distribution of error categories and at concrete examples of each error category from the data.

4.1. Label distribution

Since each example was annotated by one of the two pairs of annotators, we calculated the span overlap between each pair as a proxy for inter-annotator agreement. Since the annotators had to define the error spans themselves, the number of negative examples (spans without an error) is undefined. Because of this, the balanced F-measure [27], a variant of the F-measure addressing this issue, is used. We use a lenient F-measure where only one token needs to overlap for a positive match. Averaging over the two pairs of annotators, the score is 0.18.[6] This low score suggests this is indeed a very subjective, and cognitively demanding, task. To check whether certain errors were simply overlooked by annotators, we asked one of the annotators to inspect the annotations of their partner and added any new agreements to the annotator's original annotations. This gave us a score of 0.38. While this is still a fairly low score, it indicates that at least some of the disagreements are due to annotators overlooking certain errors. Most disagreements that persisted seem to be due to annotators seeing different things as implied.

In Figures 2 and 3 the distribution of labels chosen by the annotators are shown, per condition. As we can see, the *other* label is most commonly used by annotators in both conditions, indicating that our label set may not have been sufficient to cover the range of error phenomena found by the annotators. The second and third most common labels in both conditions are *missing in complementary 1* and *should be conflict*. The latter label's frequency suggests that the model struggled to identify conflicting information, while the former label could point to the model judging information from the first (i.e. the human) description to be unimportant. This will be further discussed in Section 4.2.

[6] Since there were so few spans that overlapped, we decided not to report the label agreement.

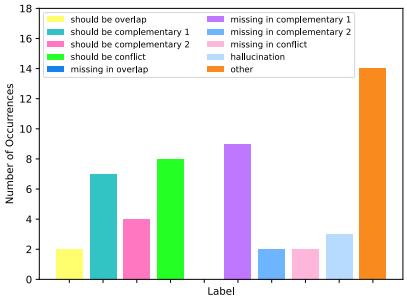

Figure 2. Distribution for *full* condition

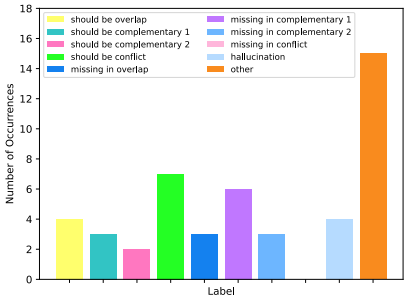

Figure 3. Distribution for *partial* condition

Adding up the label counts per condition (excluding *other*), 33 errors were annotated in the full data and 32 in the partial data. Due to this rather small difference, we did not distinguish between the conditions in the error categorization below.

4.2. Error categories

The errors labeled by the annotators were classified into seven categories, see Table 3. We focus on the most frequent error types, *category boundaries* (39 occurrences), *importance* (21 occurrences) and *implied* information (10 occurrences) and on how they impact the feasibility of the proposed task.

First, the *category boundaries* label was given to any error that arose from the annotator and the model making different assumptions about how to delineate the categories (i.e. overlap, complement and conflict). An example is that of the two spies in Figure 1, who "walk to the table, shoulder to shoulder" in the first description, but "grapple and wrestle" each other to the table in the second one. The model classified these different ways of approaching the table as complementary, while both annotators agreed on them being conflicting. The question is whether such examples pose an issue for our agent, as choosing among the two options is rather subjective. One can very well imagine that the "grappl[ing] and wrestl[ing]" involved the two spies walking "shoulder to shoulder".

Next, another error category that led to different opinions is the *importance* of individual pieces of information, and whether they therefore should be mentioned in the summary. While different types of information received this error category, a lot of the cases concern information about the participants' internal states, or what Lieberman [28] calls "seeing minds" (as opposed to "seeing matter", so visually seeing objects): Phrases like "to his surprise", "seeing the lady asleep" and "both have their own schemes" were annotated as missing from the summary. This type of information is important for understanding the stories that the comics tell. However, in the Hybrid Intelligence framework, understanding context and people are considered to be human strengths, not strengths of AI. Therefore, it may be acceptable for the agent to be weaker in this respect.

Finally, the third error category, *implied information*, includes errors that arise from differences in whether implied information was taken into account. Consider Figure 1. The human states that "both spies quickly leave the table", while the agent remembers "the tea service [being] abandoned". The LLM put this information in the respective complementary section, while the annotator saw it as overlap, arguing that the tea service

Table 3. An overview of the error categories

Name	Explanation
Category Boundaries	Errors that arise from the annotator and the model making different assumptions about where the categories (i.e. overlap, complement and conflict) start and end.
Implied Information	Errors that arise from implied information being taken into account or not.
Importance	Errors involving judgments of how important a piece of information is, and whether it therefore should be mentioned in the summary.
Hallucinations	Made-up information and simple misclassifications.
(Temporal) Ordering	Errors regarding the (temporal) ordering in which the model presented information in the summary.
Repetition	Errors surrounding repeated information, usually between categories.
Reference	Errors arising from referring expressions being used in an unclear way.

being abandoned implies that the spies have left, and that it is likely that the spies, when leaving the table, would not take the tea service with them from the cafe.

We cannot know exactly how people are interpreting the information in the event descriptions and the summary, since this depends on their world knowledge and personal experiences. Similarly, we do not know details about the LLM's training data.

For all three error types mentioned so far, we hypothesize that using the interactive capabilities of the proposed agent might alleviate them, as it allows the agent to inquire about where the human would draw the category boundaries, what they see as being implied, and about mental states of participants.

5. Conclusion

We define the task of human-agent memory co-construction, in which a human and an agent discuss their memories of past events to arrive at a more complete understanding of those events. We detail agent requirements and carry out a pilot annotation study to gain insights into the kind of errors an LLM makes when trying to combine human and agent memories, classifying the information as overlapping, complementary, or conflicting. We find that the largest sources of errors are the LLM struggling with the boundaries between these categories, not always knowing when a piece of information is important enough, and whether to take implied information into account.

In future work, we would like to extend this pilot study to further investigate the different error types and how they might be prevented. We also want to scale up the annotations to see whether our hypothesis, that incomplete data is more difficult to summarize for a model, might still hold true. Further, we would like to test the agent in an interactive setup with a human participant, to see to what extent co-construction of memories can compensate for their respective weaknesses that were discussed in this study.

6. Acknowledgements

This publication is part of the project 'Hybrid Intelligence: augmenting human intellect' (https://hybrid-intelligence-centre.nl) with project number 024.004.022 of the research programme 'Gravitation' which is (partly) financed by the Dutch Research Council (NWO). We are grateful to Maya Nachesa, Stella Verkijk, Esther Weinberg, and Ino van de Wouw for their annotation efforts.

References

[1] Tulving E, et al. Episodic and semantic memory. Organization of memory. 1972;1(381-403):1.

[2] Kocielnik R, Xiao L, Avrahami D, Hsieh G. Reflection companion: a conversational system for engaging users in reflection on physical activity. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2018;2(2):1-26.

[3] Ma Z, Mei Y, Su Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In: AMIA Annual Symposium Proceedings. vol. 2023; 2024. p. 1105.

[4] Schacter DL. The seven sins of memory: insights from psychology and cognitive neuroscience. American psychologist. 1999;54(3):182.

[5] Röttger P, Vidgen B, Hovy D, Pierrehumbert J. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In: Carpuat M, de Marneffe MC, Meza Ruiz IV, editors. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics; 2022. p. 175-90. Available from: https://aclanthology.org/2022.naacl-main.13/.

[6] Dixon RA. Memory: Collaborative. In: Smelser NJ, Baltes PB, editors. International Encyclopedia of the Social Behavioral Sciences. Oxford: Pergamon; 2001. p. 9570-2. Available from: https://www.sciencedirect.com/science/article/pii/B0080430767014960.

[7] Gyollai D. Collaborative inhibition: a phenomenological perspective. Review of Philosophy and Psychology. 2024:1-19.

[8] Rajaram S, Pereira-Pasarin LP. Collaborative memory: Cognitive research and theory. Perspectives on psychological science. 2010;5(6):649-63.

[9] Harris CB, Paterson HM, Kemp RI. Collaborative recall and collective memory: What happens when we remember together? Memory. 2008;16(3):213-30.

[10] Sánchez ML, Correa M, Martínez L, Ruiz-del Solar J. An episodic long-term memory for robots: The bender case. In: RoboCup 2015: Robot World Cup XIX 19. Springer; 2015. p. 264-75.

[11] Kasap Z, Magnenat-Thalmann N. Building long-term relationships with virtual and robotic characters: the role of remembering. The Visual Computer. 2012;28:87-97.

[12] Campos J, Kennedy J, Lehman JF. Challenges in exploiting conversational memory in human-agent interaction. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems; 2018. p. 1649-57.

[13] Chafe WL. The time of the sign: A semiotic interpretation of modern culture. MacCannell D, MacCannell JF, editors. Bloomington, MN: Indiana University Press; 1982.

[14] Langacker RW. Foundations of cognitive grammar: Volume I: Theoretical prerequisites. vol. 1. Stanford university press; 1987.

[15] Bansal N, Akter M, Santu SKK. Semantic overlap summarization among multiple alternative narratives: An exploratory study. In: Proceedings of the 29th International Conference on Computational Linguistics; 2022. p. 6195-207.

[16] Lerman K, McDonald R. Contrastive Summarization: An Experiment with Consumer Reviews. In: Ostendorf M, Collins M, Narayanan S, Oard DW, Vanderwende L, editors. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Boulder, Colorado: Association for Computational Linguistics; 2009. p. 113-6. Available from: https://aclanthology.org/N09-2029.

[17] da Cruz Souza JW, Di Felippo A. Characterization of temporal complementarity: fundamentals for multi-document summarization. Alfa: Revista de Linguistica. 2018;62(1):121-47.

[18] Xu R, Qi Z, Guo Z, Wang C, Wang H, Zhang Y, et al. Knowledge Conflicts for LLMs: A Survey. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics; 2024. p. 8541-65. Available from: https://aclanthology.org/2024.emnlp-main.486/.

[19] Wang Y, Feng S, Wang H, Shi W, Balachandran V, He T, et al. Resolving Knowledge Conflicts in Large Language Models. In: First Conference on Language Modeling; 2024. Available from: https://openreview.net/forum?id=ptvV5HGTNN.

[20] Akata Z, Balliet D, De Rijke M, Dignum F, Dignum V, Eiben G, et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer. 2020;53(8):18-28.

[21] Brennan SE, Clark HH. Conceptual pacts and lexical choice in conversation. Journal of experimental psychology: Learning, memory, and cognition. 1996;22(6):1482.
[22] Hawkins RX, Frank M, Goodman ND. Convention-formation in iterated reference games. Cognitive Science. 2017.
[23] Shi Z, Sen P, Lipani A. Lexical Entrainment for Conversational Systems. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics; 2023. p. 278-93.
[24] Kruijt J. The impact of common ground on Referring Expressions in Human-Robot interaction. Vrije Universiteit Amsterdam; 2025.
[25] Wang X, Zhou Y, Liu X, Lu H, Xu Y, He F, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. arXiv preprint arXiv:240110529. 2024.
[26] Team G, Georgiev P, Lei VI, Burnell R, Bai L, Gulati A, et al.. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context; 2024. Available from: https://arxiv.org/abs/2403.05530.
[27] Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. Journal of the American medical informatics association. 2005;12(3):296-8.
[28] Lieberman MD. Seeing minds, matter, and meaning: The CEEing model of pre-reflective subjective construal. Psychological Review. 2022;129(4):830.