The automated model of comprehension version 4.0 – Validation studies and integration of ChatGPT

Article *in* Computers in Human Behavior · February 2024

DOI: 10.1016/j.chb.2024.108154

CITATIONS
11

READS
262

5 authors, including:

Dragos-Georgian Corlatescu
Universitatea Națională de Știință și Tehnologie Politehnica București
22 PUBLICATIONS **101** CITATIONS
SEE PROFILE

Micah Watanabe
Arizona State University
16 PUBLICATIONS **139** CITATIONS
SEE PROFILE

Stefan Ruseti
Universitatea Națională de Știință și Tehnologie Politehnica București
73 PUBLICATIONS **807** CITATIONS
SEE PROFILE

Mihai Dascalu
Universitatea Națională de Știință și Tehnologie Politehnica București
359 PUBLICATIONS **3,752** CITATIONS
SEE PROFILE

Computers in Human Behavior 154 (2024) 108154

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

The automated model of comprehension version 4.0 – Validation studies and integration of ChatGPT

Dragos-Georgian Corlatescu [a], Micah Watanabe [b], Stefan Ruseti [a], Mihai Dascalu [a,c,*], Danielle S. McNamara [b]

[a] Computer Science and Engineering Department, National University of Science and Technology POLITEHNICA Bucharest, Bucharest, Romania
[b] Department of Psychology, Arizona State University, Tempe, AZ, USA
[c] Academy of Romanian Scientists, Bucharest, Romania

ARTICLE INFO

Handling Editor: Prof. Nicolae Nistor

Keywords:
Natural language processing
Reading comprehension
Automated model of comprehension
ChatGPT
Large language models

ABSTRACT

Modeling reading comprehension processes is a critical task for Learning Analytics, as accurate models of the reading process can be used to match students to texts, identify appropriate interventions, and predict learning outcomes. This paper introduces an improved version of the Automated Model of Comprehension, namely version 4.0. AMoC has its roots in two theoretical models of the comprehension process (i.e., the Construction-Integration model and the Landscape model), and the new version leverages state-of-the-art Large Language models, more specifically ChatGPT, to have a better contextualization of the text and a simplified construction of the underlying graph model. Besides showcasing the usage of the model, the study introduces three in-depth psychological validations that argue for the model's adequacy in modeling reading comprehension. In these studies, we demonstrated that AMoC is in line with the theoretical background proposed by the Construction-Integration and Landscape models, and it is better at replicating results from previous human psychological experiments than its predecessor. Thus, AMoC v4.0 can be further used as an educational tool to, for example, help teachers design better learning materials personalized for student profiles. Additionally, we release the code from AMoC v4.0 as open source in a Google Collab Notebook and a GitHub repository.

1. Introduction

The field of Learning Analytics (Hernández-de-Menéndez et al., 2022; Sghir et al., 2023; Siemens & Long, 2011) emerged from multiple research fields including, but not limited to, evaluation and assessment, educational data mining, online learning, and technology-enhanced learning (Ye, 2022). One of the research trends within the domain of LA is the development of *predictive models* of student behaviors (Joksimović et al., 2019). Predictive models of student behavior are algorithmic models that use past data to predict future performance (Geden et al., 2021). Predictive models of student behavior have been used to match students to specific tasks or lessons, identify and develop appropriate interventions for students, and forecast educational outcomes such as college retention rates (Larrabee Sønderlund et al., 2019; Namoun & Alshanqiti, 2020; Romero & Ventura, 2020). In addition, predictive models can be used to provide an understanding of the cognitive process in the learning task and thus identify the trajectory of learning and predict output variables such as scores on a reading comprehension test (Phillips et al., 2023).

The development of predictive models of students' *reading* process has been informed by theoretical models of reading found in cognitive and educational psychology (Davoudi & Moghadam, 2015; Elleman & Oslund, 2019; McNamara & Magliano, 2009). Theoretical models of reading describe the outcome of the reading process as a mental representation of the text (Davoudi & Moghadam, 2015; Elleman & Oslund, 2019; McNamara & Magliano, 2009). The readers' mental representation of a text is the set of concepts and connections between concepts retained in memory (Davoudi & Moghadam, 2015; Elleman & Oslund, 2019; McNamara & Magliano, 2009). The mental representation includes concepts and connections *explicit* in the text, as well as concepts and connections that are *inferred* by the reader (Kintsch & Welsch, 1991). For example, consider these two sentences:

* Corresponding author. Computer Science and Engineering Department, National University of Science and Technology POLITEHNICA Bucharest, Bucharest, Romania.
E-mail addresses: dragos.corlatescu@upb.ro (D.-G. Corlatescu), micah.watanabe@asu.edu (M. Watanabe), stefan.ruseti@upb.ro (S. Ruseti), mihai.dascalu@upb.ro (M. Dascalu), dsmcnama@asu.edu (D.S. McNamara).

https://doi.org/10.1016/j.chb.2024.108154
Received 26 November 2023; Received in revised form 10 January 2024; Accepted 21 January 2024
Available online 5 February 2024

performance on reading comprehension tasks (Davoudi & Moghadam, 2015; Raudszus et al., 2019). AMoC also employs these assumptions about the reading process by modeling the spreading activation through a network of concepts, wherein some concepts are explicit from the text, and others must be inferred or deduced from the context. In addition, AMoC includes parameters to simulate individual differences in readers' skill and knowledge. Finally, in this paper, a major goal is to test whether the predictions of AMoC are comparable to the observations in past experiments on readers and reading comprehension.

The development of AMoC, and thus, the predictive graph generated by AMoC, was based on two psychological models of reading comprehension. The first was the Construction Integration (CI) Model (Kintsch & Welsch, 1991) emerged in the cognitive sciences as a melding of multiple disciplines, including connectionist models and theories of discourse comprehension (Kintsch, 1998). The CI model describes comprehension in terms of spreading activation of concepts through a network, framing comprehension as a two-step process. In the first step (Construction), the reader constructs a mental representation of the text's meaning from concepts explicit in the text and generates inferences between concepts in the text. In the second step (Integration), the reader integrates the mental representation of the text into memory. Prior knowledge is an important factor at each step, as prior knowledge affords the reader (a) more appropriate activation of concepts and in turn, more accurate inference generation and (b) better integration of new concepts into memory. AMoC v4.0 reflects the CI model, by utilizing a two-step process wherein explicit textual information is first considered, followed by inferred concepts.

The second model that informed the creation of AMoC was The Landscape Model (Van den Broek et al., 1999). The Landscape Model simulates the fluctuation of activated concepts across time. Within the Landscape Model, the activation of concepts was established from four sources: reading the text directly (text-based); remembering not too distantly read concepts, meaning that a sentence read before the current one is still with some degree in the reader's mind (text-based); involuntary activating the reader's prior knowledge (inferred) also known as the cohort activation mechanism; and lastly, voluntary activation of reader's knowledge when the person deliberately stops and builds connections (thus activating concepts) in the process of comprehending the text. Similar to the Landscape Model, AMoC v4.0 dynamically models change in the reading process across time.

Both the Construction Integration Model (Kintsch & Welsch, 1991) and Landscape Model (Van den Broek et al., 1999) can be used to predict a reader's mental representation of a text based on individual and text factors. However, the classic method of prediction relies on human input, such as answering questions regarding what concepts are activated and to what extent. Another downside is that the process of activation may be forced (e.g., participants were asked, "Are you thinking about this concept?") rather than natural. Thus, the creation of automatic models of reading comprehension based on these models can significantly reduce the time and effort to generate predictive models of readers' mental representation. In turn, these models can be used to identify readers' individual skills and predict reading outcomes. This information can then be applied by teachers when selecting appropriate texts for readers across different skill levels.

2.2. The automated model of comprehension – AMoC (versions 1.0–3.0)

The Automated Model of Comprehension (AMoC), as described by Dascalu et al. (2018), was built on the Construction-Integration (Kintsch & Welsch, 1991) and the Landscape Model (Van den Broek et al., 1999) to automate the modeling of the reading comprehension process across multiple texts. AMoC simulates readers' mental representation of a text during the reading process by constructing a graph that captures the generation of concepts (nodes), inferences, and semantic connections between them.

Three versions of AMoC have been published so far, namely v1.0 (Dascalu et al., 2018), v2.0 (Corlatescu et al., 2021), and v3.0 (Corlatescu et al., 2023). The initial and subsequent iterations of AMoC utilized similar structures. AMoC generates a concept graph that incorporates both explicit textual concepts and inferred concepts and corresponds to the readers' mental representation of the text while reading. Inferred concepts were generated by leveraging a hybrid approach that combined semantic distance measures from word2vec (Mikolov et al., 2013) and WordNet (Miller, 1995). The importance of nodes within the graph was updated using a PageRank algorithm (Page et al., 1999), ensuring a comprehensive representation of the underlying information. However, notable distinctions arose between the first two versions, primarily the transition from Java (v1.0) to Python (v2.0), alongside the integration of newer models for text processing. AMoC 3.0 introduced a novel methodology by harnessing the *attention* mechanism within the Transformers (Vaswani et al., 2017). In this scenario, the *attention* concept refers to the process through which a model focuses on different parts of the input sequence when making a prediction. It works by computing attention scores between each pair of input elements (i.e., words and tokens) which determine how much attention should be paid to each element during processing, enabling the model to capture complex relationships and dependencies in the text. Concrete implementations of this idea are the GPT models which are trained to generate text that resembles human writings. The models look and pay attention to the context to complete their task. Coming back to the AMoC, the third version utilized the attention scores derived from GPT models to construct the graph by connecting nodes through edges with scores corresponding to the attention scores between words in the textual content. Furthermore, inferred concepts were generated by leveraging the capability of GPT2 (Radford et al., 2019) to generate new sentences based on the processed text, subsequently selecting concept words from these generated sentences. This approach demonstrated improved performance compared to AMoC version 2.0, particularly in tasks requiring the discrimination between texts exhibiting high and low cohesion.

2.3. Modeling comprehension using LLMs

Similar work to model the comprehension process by leveraging LLMs has been conducted by Andrus et al. (2022) as well as Patel et al. (2023). The work of Andrus et al. (2022) shares similarities with AMoC in terms of building a dynamic knowledge graph to represent essential information from the text. Their goal was to improve the long-term memory of LLMs to increase long-term coherence of narrative texts generated by the LLM (e.g., preventing the LLM from introducing conflicting information into the narrative). The authors tested their approach using a comprehension evaluation of LLMs where the input exceeded the context window of the models. They would iteratively enter one to three sentences of a long text (e.g., 10,000+ tokens) into the LLM to be summarized. This allowed the longer text to be reduced to a length which could then be used to generate novel text sections consistent with the earlier sections. Their method was similar to AMoC as it involved extracting relevant nodes from the knowledge graph and rephrasing it into coherent sentences or summarization. However, they did not include certain aspects of reading comprehension theory – in particular, generation of inferred concepts and activation over time. However, it is important to highlight that AMoC distinguishes itself from this work, given its solid foundations in comprehension theory.

The work by Patel and colleagues was comparable in terms of components of the model. Their goal was to model comprehension of math word problems. The model included both text processing powered by an LLM combined with natural language processing measures (i.e., readability metrics) to simplify difficult math problems. However, the product of the model was direct recommendations (e.g., "This problem needs to be simplified") and not a comprehensive graph of the connections between ideas in the problems.

Finally, past efforts in the field of reading comprehension to model the processes involved constructing models that predicted the

Robert drove on the bridge over the river.

Cathy paddled under the bridge in her rowboat.

From these two sentences, it can be inferred that "Robert is above Cathy." In this example, the readers' mental representation would include explicit information about Robert and Cathy, as well as the inference.

The goal of predictive models of students' reading process is to generate a graph (i.e., a predicted set of concepts and connections) that corresponds to readers' mental representation for any given text. A more accurate predictive graph can improve educational outcomes. Teachers can use predictive models of reading to better understand individual students' reading skills, identify reading trajectories, and predict performance on complex reading tasks (Farhana et al., 2022). Natural language processing (NLP) tools have been used to develop and improve predictive models of students' reading behavior Allen et al. (2015). NLP is a collection of computational techniques for automatic analysis and representation of human languages (Chowdhary & Chowdhary, 2020, pp. 603–649). NLP tools are used to parse the text for explicit concepts and connections and generate inferred concepts and connections. For example, NLP tools such as word2vec (Mikolov et al., 2013) and WordNet (Miller, 1995) have been used to generate inferences in a predictive graph that correspond to inferences in a readers' mental representation of a text (Corlatescu et al., 2021; Dascalu et al., 2018). Continued advances in NLP, specifically in Large Language Models (LLMs), have improved previous attempts to model readers' mental representation (Corlatescu et al., 2023). LLMs are advanced Artificial Intelligence (AI) systems that use deep neural networks to understand and generate human language, greatly surpassing previous language models (e.g., word2vec) in terms of the quality of the generated text and their generalizability among different tasks.

The current paper builds on the LA program to develop predictive models of students' reading behavior by leveraging new and improved LLMs, namely, ChatGPT (OpenAI, 2022). ChatGPT has already been examined in the context of LA, for example Susnjak (2023a) introduced a novel framework that combines transparent machine learning and prescriptive analytics, incorporating advancements in large language models like ChatGPT to communicate insights to learners. The research demonstrates the integration of predictive modeling and prescriptive analytics using ChatGPT to generate human-readable feedback for at-risk students. Additionally, Dai et al. (2023) explored the transformative impact of ChatGPT and generative artificial intelligence on higher education, addressing key questions about their role and implications in the evolving educational landscape. ChatGPT is a Transformer-based LLM trained using Reinforcement Learning from Human Feedback (RLHF; Ouyang et al., 2022). The Transformer (Vaswani et al., 2017) is a deep neural network that allows efficient training of a language model using a vast amount of text. During this training, the model learns word probabilities given a context and thus can generate correct outputs (both from a grammatical and logical point of view) when required. However, the answers generated by this type of model are, in some cases, not what a human would expect given a query. To address this limitation, RLHF was introduced to teach the model answers perceived by humans as being correct or preferable. In this process, human evaluators annotated multiple responses of the same model to certain questions, offering insights into whether the answer was right or wrong. The model was then additionally trained to learn human preferences and integrate them into its answers. As a result of this training, inferences generated by ChatGPT may more closely resemble human inferences (i.e., the concepts that are related to the current context are brought into the mind) compared to other NLP tools or older LLMs (Corlatescu et al., 2023). Improved generation of inferences would result in more accurate predictive models of reading processes, as the outcome of the model (i.e., the concept graph) would more closely correspond to readers' mental representation of the text.

Open-source alternatives to ChatGPT were also considered in early experiments; however, they provided suboptimal results and were not capable of adequately constructing a knowledge graph, described in detail in the following sections. The decision to leverage ChatGPT was grounded in its superior performance and robust capabilities in understanding and generating human-like text. Moreover, all newer, larger, and more powerful GPT models with more than 70B parameters are accessible via Application Programming Interface (API) calls, enabling a rapid integration within AMoC with minimal adjustments, thus increasing the flexibility of our proposed solution.

The goal of this paper is to introduce a new version of the Automated Model of Comprehension, version 4.0 (AMoC v4.0), and validate it using three studies. AMoC v4.0 simulates human reading comprehension of texts by utilizing state-of-the-art research in large language models (LLMs), namely ChatGPT, to generate knowledge graphs (i.e., a semantic network with nouns, adjectives, and adverbs as nodes, and verbs as edges) that correspond to humans' mental representation while reading. In the following sections, we provide a background on modeling reading comprehension and previous versions of AMoC. We describe our method, including an in-depth description of building and updating the AMoC v4.0 knowledge graph using ChatGPT, the parameters used to model the reader skills and an example of AMoC v4.0 "reading" a text. Next, AMoC v4.0 was tested in various scenarios to understand its capabilities and limitations. Then, we describe three psychological validation studies that were performed to compare the results of AMoC v4.0 to previous research in the field of text comprehension. Finally, we discuss the implications of our findings.

To foster transparency and facilitate further research, our model and all associated code are publicly available on GitHub (ReaderBench, 2023) and in a Colab Notebook (Corlatescu, 2023).

2. Background

2.1. Modeling reading comprehension

In psychological research, multiple models of reading comprehension have been proposed and instantiated (Davoudi & Moghadam, 2015; Elleman & Oslund, 2019; McNamara & Magliano, 2009). A reader's mental representation of a text consists of the knowledge or information they learn from the text (Kintsch & Welsch, 1991; McNamara & Magliano, 2009). Modeling reading comprehension is the process of simulating the reader's *mental representation* of a text. A predictive model of a readers' mental representation can be graphed as the set of nodes and connections that readers are expected to retain in memory after reading. The nodes and connections are described as propositions (e.g., Jill – Ate – Breakfast; Breakfast – Is – Eggs) and roughly correspond to parts of speech corresponding to content words (i.e., nouns, verbs, adjectives, adverbs); however, the parts of speech that are represented as either nodes or connections is subject to debate (Kintsch & Welsch, 1991; McNamara & Magliano, 2009). The nodes and connections in the predictive graph can represent both information explicit in the text, or information that is inferred (i.e., implicit) in the text. Each iteration of the Automated Model of Comprehension (AMoC) utilized explicit concept mapping; however, a major decision in each iteration was how to generate inferences within the model, or, in other words, how to simulate the human inference process.

In addition to the basic structure of readers' mental representation, psychological models of reading comprehension also share key assumptions about the reading process. Namely, that information in memory can be activated by related stimuli, the activation can spread to connected concepts, and this activation process is largely unconscious and automatic (McNamara & Magliano, 2009). In addition, the modeling process includes an assessment of individual differences in reading skills and text features because there are interactive effects of readers' skills and text features on comprehension (McNamara et al., 1996). Finally, the coherence or stability of a readers' mental representation in memory is predictive of their reading comprehension and

significance of various factors influencing reading comprehension, such as syntactic structure, based on metrics like scholastic achievement (Meneghetti et al., 2006; Pečjak et al., 2011). More recent efforts have included natural language processing metrics to improve the performance of the model (Nie et al., 2022). While these studies are valuable, it is essential to note that they differ from the focus of this paper, as they looked at other discrete behavioral measurements (e.g., GPA scores) to model individuals' reading comprehension scores rather than looking at features of the text to model individuals' memory of the text. Thus, AMoC represents a distinctive effort to model reading comprehension both in terms of the end product (predicting memory network versus predicting comprehension test performance) and in its use of an LLM as an instrument to build the concept graph and simulate inferences from the text.

2.4. The current research – AMoC v4.0

In this work, we present AMoC version 4.0 that builds on AMoC v1.0-v3.0 using one of the most powerful and popular LLM, ChatGPT (OpenAI, 2022). AMoC v4.0 models comprehension across numerous texts and includes new parameters to model different reading skill levels and knowledge levels. While the fundamental concepts of AMoC remained intact, refinements were implemented to effectively harness the capabilities of ChatGPT. These modifications encompassed various aspects, primarily the construction of the AMoC concept graph and the generation of inferred concepts.

ChatGPT, similar to other LLMs, must be queried for responses. This process led to the rise of a new methodology, *Prompting* (Liu et al., 2023) – i.e., the act of specifying the exact prompt or information to provide ChatGPT to enhance response quality. ChatGPT has vast applicability in education (AlAfnan et al., 2023) and learning analytics – for example, Susnjak (2023b) introduced a novel framework that combines transparent Machine Learning and prescriptive analytics, utilizing advanced language models like ChatGPT to provide human-readable feedback to at-risk learners. Moreover, OpenAI offers special attention to the educational use of ChatGPT (OpenAI, 2023b).

In contrast to AMoC v1-v3, AMoC v4.0 simulates human reading comprehension by leveraging the capabilities of ChatGPT to model the activated and inferred concepts within the readers' mental representation of the text. In the following studies, we performed prompt engineering (Radford et al., 2021) consistent with the growing body of research.

This research study aims to answer two main research questions.

RQ1. *To what extent do the inferences generated by AMoC v4.0 resemble human-generated inferences compared to the inferences generated by past versions of AMoC?*

RQ2. *To what extent does the knowledge graph in AMoC v4.0 generated using ChatGPT align with human comprehension when compared to the semantic similarity that connects concepts from the text?*

Our two research questions are focused on how the performance of AMoC v4.0 improves on past versions of AMoC due to the use of ChatGPT and to what extent our work models human comprehension in an automated manner. Our first research question concerns the generation of inferences within the concept graph that represents a human's mental representation of the text. Previous versions of AMoC used language models such as word2vec and GPT2 to generate inferences between words or ideas in the text. In AMoC v4.0, ChatGPT was used to generate inferences. Due to the extended RLHF training, AMoC v4.0 should generate inferences more similar to a human.

Our second research question focused on the representation of the connections between ideas in the concept graph. In previous versions of AMoC, the concept graph of the text (corresponding to a human's mental representation of the text) was constructed of nodes that were individual words (e.g., knight, is, heroic) and edges that linked two concepts together by assigning a similarity score (i.e., a probability of the two words co-occurring in texts). AMoC v4.0 constructs a *knowledge graph*. This knowledge graph has two types of nodes: concepts (primarily nouns: knight, dragon, castle) and properties (heroic, evil, majestic). The edges are constructed primarily with verbs that describe the relationship between the concept and property (e.g., knight, is, heroic) or concept and concept (e.g., knight, fights, dragon). Modeling comprehension as a set of relationships between concepts and properties rather than a set of semantic similarity scores reflects human comprehension – particularly in narrative texts (Zwaan et al., 1995). Thus, the knowledge graph constructed in AMoC v4.0 should better represent human comprehension than the concept graph constructed in previous versions of AMoC.

3. Method

3.1. AMoC v4.0 processing flow and step-by-step simulation

3.1.1. Preprocessing

A text is preprocessed by replacing all pronouns with corresponding reference nouns (e.g., "he" is transformed into "knight"), as the AMoC knowledge graph construction requires explicit references. In the previous versions of the model, a coreference resolution model was employed that linked a pronoun to its referenced noun(s). In order to maintain consistency throughout the AMoC modeling process, in AMoC v4.0, we performed pronoun replacement directly using ChatGPT in a sentence-by-sentence manner.

3.1.2. Knowledge graph construction

The core component of the analysis is the knowledge graph. Fig. 1 shows the step-by-step construction of the knowledge graph. The construction of the knowledge graph is conducted on a sentence-by-sentence basis. Each sentence is iteratively added to the graph. The first sentence is a special case since the knowledge graph is empty. Thus, the initial sentence analysis has a subset of operations and specified prompts for ChatGPT compared to the other sentences. The subsequent steps showcase the first sentence processing.

3.1.3. Steps 1a – 1c. Sentence 1 text base concepts

Steps 1a – 1c represent a reader's passive activation of explicit concepts and properties in the text.

Step 1a. Concepts (nouns) and properties (adjectives) are extracted from the sentence. The concepts and properties are referred to as nodes in further operations. This operation uses an open-source Python library, SpaCy version 3.2 (Honnibal & Montani, 2017). For AMoC v4.0, SpaCy is used for sentence segmentation, tokenization, lemmatization, and part of speech tagging.

Step 1b. ChatGPT is prompted to generate a list of edges between concepts or between concepts and properties. The prompt includes the extracted concepts and properties from step 1a, as well as the full text (see Appendix 1, Prompt 1)

Step 1c. The nodes and edges are added to the Knowledge graph.

3.1.4. Steps 2a – 2c. Sentence 1 inferred concepts

Steps 2a-2c represent a reader's process of inferring new concepts and properties while reading the text.

Step 2a. ChatGPT is prompted to produce concepts and properties that "are in line with the overall coherence and sense within the given text, but they are not in the text". The prompt includes the extracted concepts and properties from step 1a, as well as the full text (see Appendix 1, Prompt 2).

Step 2b. ChatGPT is prompted to generate new edges between the explicit concepts and the inferred concepts. The prompt includes the extracted concepts and properties from step 1a, the inferred concepts

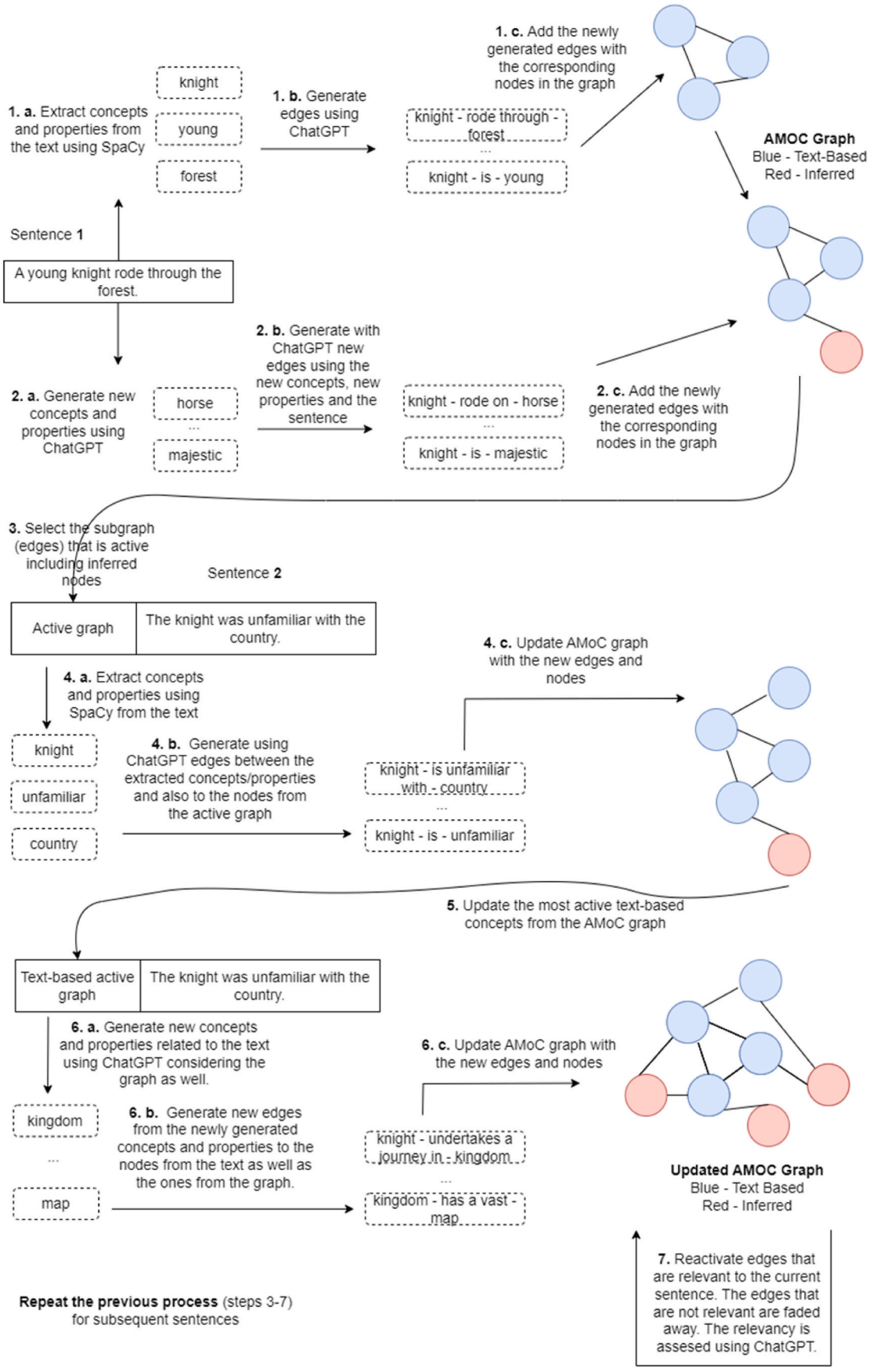

Fig. 1. AMoC v4.0 Flow.

from step 2a, as well as the full text (see Appendix 1, Prompt 3). These new edges are validated before step 2c (i.e., edges that already exist or are very similar to another edge are not added to the graph).

Step 2c. The inferred nodes and new edges are added to the graph and labeled as INFERRED if not present in the text.

3.1.5. Step 3. activation of concepts across time

The step-by-step process above is repeated for subsequent sentences (sentence 2 – sentence *n*). However, for subsequent sentences (i.e., not the first sentence), the constructed knowledge graph is used as part of the input. Only active concepts, properties, and edges are retained between sentences. Consistent with research on text comprehension, the explicit concepts and properties from the previous sentence are considered active. Other nodes, both explicit or inferred, are considered active if they are at a minimum distance *N* (see above: AMoC v4.0 Parameters) from the explicit concepts and properties.

Step 3. A subgraph is generated which includes only active nodes and edges and used as input in subsequent steps.

3.1.6. Steps 4a – 4c. Adding a new sentence to the knowledge graph

Once a sentence (or sentences) has been processed and a subgraph of active concepts has been created, a new sentence can be added to the graph.

Step 4a. Concepts (nouns) and properties (adjectives) are extracted from the sentence and added to the list of active nodes in the subgraph (both explicit and inferred).

Step 4b. ChatGPT is prompted to generate a list of new edges between the extracted nodes and the active nodes and edges. The prompt includes the extracted nodes from step 4a, the active nodes and edges from step 3, and the sentence text (see Appendix 1, Prompt 4). These new edges are validated before step 4c (i.e., edges that already exist or are very similar to another edge are not added to the graph).

Step 4c. The generated nodes and edges are added to the knowledge graph.

3.1.7. Step 5. creating a new text-based active graph

The *new text-based active graph* includes only the explicit nodes from the current sentence (and not the previous one). Inferred nodes are not included because readers do not typically infer words based on other inferred words (i.e., double inference).

Step 5. The active concepts are updated to include only the explicit nodes from the current sentence.

3.1.8. Steps 6a – 6c. Adding new inferences to the knowledge graph

Finally, new inferences are generated from the active, explicit nodes.

Step 6a. ChatGPT is prompted to produce concepts and properties that "are in line with the overall coherence and sense within the given text, but they are not in the text". The prompt includes the extracted concepts and properties from step 4a, the active nodes and edges from step 3, as well as the sentence text (see Appendix 1, Prompt 5).

Step 6b. ChatGPT is prompted to generate new edges between the explicit concepts and the inferred concepts. The prompt includes the extracted concepts and properties from step 4a, the active nodes and edges from step 3, the inferred concepts from step 6a, as well as the full text (see Appendix 1, Prompt 6). These new edges are validated before step 2c (i.e., edges that already exist or are very similar to another edge are not added to the graph).

Step 6c. The inferred nodes and new edges are added to the knowledge graph and labeled as INFERRED if not present in the text.

3.1.9. Step 7. reactivation

When readers proceed through the text, edges (relationships) that have already been activated can be reactivated by repeatedly encountering similar nodes and edges in the text. Similarly, the activation of an edge can decay over time if similar nodes and edges are not encountered in the text. In the knowledge graph, an edge can be reactivated in this step, which increases the activation of the nodes. Or an edge's activation can decay by a factor (parameter). When the deactivation score of an edge reaches 0, the edge becomes inactive. Similar to text preprocessing, this step uses ChatGPT to remain consistent.

Step 7. ChatGPT is prompted to name the edges that are related to the text. The prompt includes all the edges from the knowledge graph and the current sentence text (see Appendix 1, Prompt 7).

3.1.10. Finishing the knowledge graph

After completing the second sentence, steps 3a –7 are repeated iteratively for each sentence in the text until the end of the text.

3.2. AMoC v4.0 parameters

AMoC considers a set of parameters that facilitate diverse simulations of reading comprehension, catering to learners with varying proficiency levels. For all the parameters, the values can be set to any natural number, with larger numbers representing a more skilled reader (e.g., a reader who activates and retains more concepts while reading), and smaller numbers representing less skilled readers. Each of the parameters has an underlying rationale and default value in the model.

- *MaxDistance*: When the active subgraph (i.e., the subgraph given by the active concepts and properties connecting them) is extracted from the AMoC graph, the *MaxDistance* argument specifies the distance from the text-based nodes for which another node is considered active. The default value is 2, a value selected from a reading comprehension perspective. In terms of comprehension theory, a value of 1 would indicate only the edges connected with the explicit, text-based nodes are active, however readers typically have both explicit information from the sentence, and the concepts from the previous sentence activated. In contrast, a value of 3 would activate edges that are more than two steps removed from the explicit information in the sentence, which would reflect a highly skilled reader that has activated an extensive network of concepts; for example, a score of 3 in Fig. 6, would imply that the reader has activated in memory a relationship of "property" that is unrelated to the text, such as "owned".
- *MaxNewConcepts*: The number of concepts inferred from ChatGPT is limited by this variable; the default value is 3 - empirically set based on expert judgment and multiple trials of using ChatGPT; nevertheless, the values can be changed to accommodate any experimental setup.
- *MaxNewProperties*: The number of properties inferred from ChatGPT is limited by this parameter; the default value is 3 - empirically set based on expert judgment and multiple trials of using ChatGPT; nevertheless, the values can be changed to accommodate any experimental setup.
- *EdgeActiveScore*: When an edge is added to the graph or is reactivated, the active score is initialized with this value. When an edge is not related to the context of the sentence, the edge active score is decreased by 1. The edge gets deactivated if it reaches the 0 threshold. The default value is 2, an empirically obtained number based on expert judgment and multiple trials of using ChatGPT, which can be changed to experiment with different levels of information retention.
- *NumberReactivatedEdges*: This parameter is the maximum number of edges that can be reactivated by ChatGPT, given a sentence. The

default value is 15; throughout the subsequent studies, the count of retrieved edges rarely exceeded 10.

3.3. Example case

We used the text from the original Landscape Model paper (Van den Broek et al., 1999) (see Appendix 2. Landscape Model Knight Story) to simulate how AMoC v4.0 works. We showcase how the first two and the last sentences were processed step by step. The edges will be represented using the format: *node1, edge label, node2*. A node is represented in the format: word (concept/property, inferred/text-based, activation score 0 being the highest, then 1, 2, etc.).

The first sentence in the text is, "The knight rode through the forest". In steps 1a-1c, the model extracted three explicit (blue) nodes and two edges from the text, which are shown in Fig. 2.

In steps 2a-2c, the model generated eight inferred (yellow) nodes and edges. However, two edges were removed before being added to the graph as they were exactly the same as the explicit nodes/edges, or they simply had a different verb tense (e.g., Knight – rides through - forest). See Fig. 3 for the knowledge graph after the first sentence.

The second sentence is, "The knight was unfamiliar with the country". In step 3, all of the explicit and inferred nodes from step 1 were active, and thus were retained for the following sentence. In steps 4a-4c, the explicit (blue) nodes from the second sentence are added to the knowledge graph (see Fig. 4).

In steps 6a-6c, the model generated six inferred (yellow) edges. Two were removed as they were already in the graph. See Fig. 5 for the knowledge graph after the second sentence.

Finally, the model selects which edges from the graph are reactivated. In this example, the maximum activation score is two. The model only reactivated the edge, "knight, wields, sword", which retained the maximum activation of two. The other edges decayed by one. Fig. 6 shows the knowledge graph after the second sentence with an activation score for each edge.

Fig. 7 depicts the active concepts in the knowledge graph after the last sentence, "The princess married the knight". The generated edges based on the graph and the sentence were: "princess, is interested in, marriage", "knight, is, brave", "knight, is, loyal", "knight, is, chivalrous", "princess, is in, castle". The other nodes in the knowledge graph were inferred or reactivated by the model based on the sentence text and the previous graph.

4. Validation studies

In order to answer our research questions and assess the capabilities of the Automated Model of Comprehension version 4.0, three studies were conducted comparing the features of the AMoC knowledge graph to previous versions of AMoC (3.0) and past findings in reading comprehension research. While our research does not directly involve human participants, the goal of the validation studies is to compare the predictions of AMoC to past empirical results on human reading. We selected this methodology because the concept graph generated by AMoC is thought to correspond to the mental representation generated by humans while reading. Readers' mental representation predicts their performance on comprehension tasks such as multiple-choice tests. Thus, if the AMoC concept graph corresponds to human mental representation, we would expect the predictions of AMoC to show a strong relationship with the empirical results.

We include for each study details on the considered dataset, details on the employed method that considers custom configurations performed on top of the AMoC v4.0 presented in the previous section, results, and corresponding discussions. The following section includes overarching discussions, as well as limitations of our automated model.

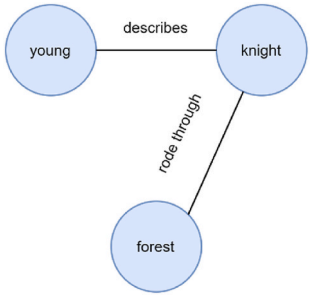

Fig. 2. First sentence graph after text analysis.

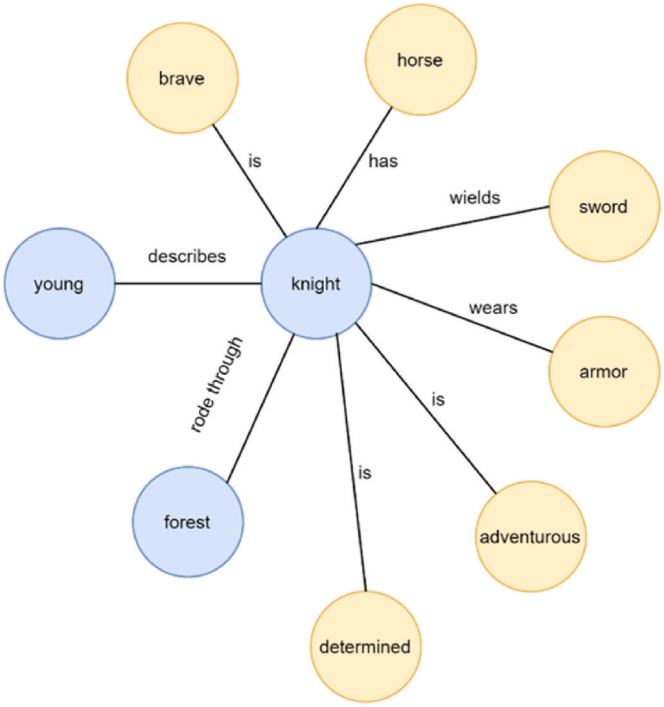

Fig. 3. First sentence graph after inference.

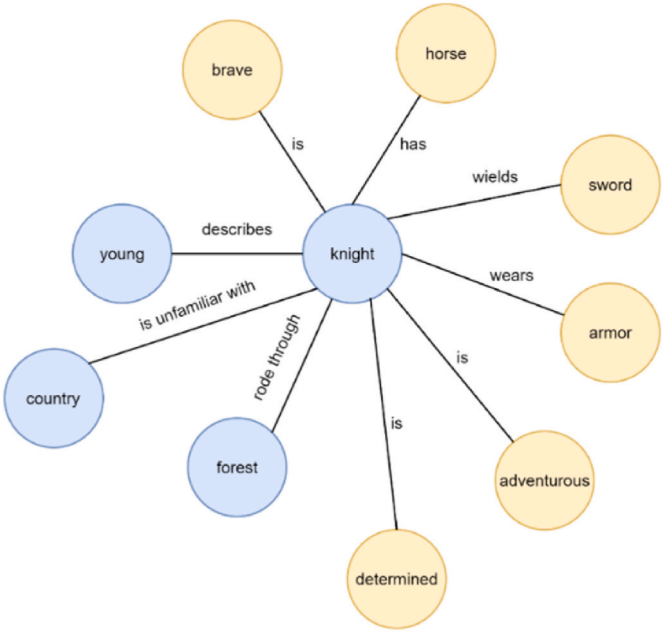

Fig. 4. Second sentence graph after text analysis.

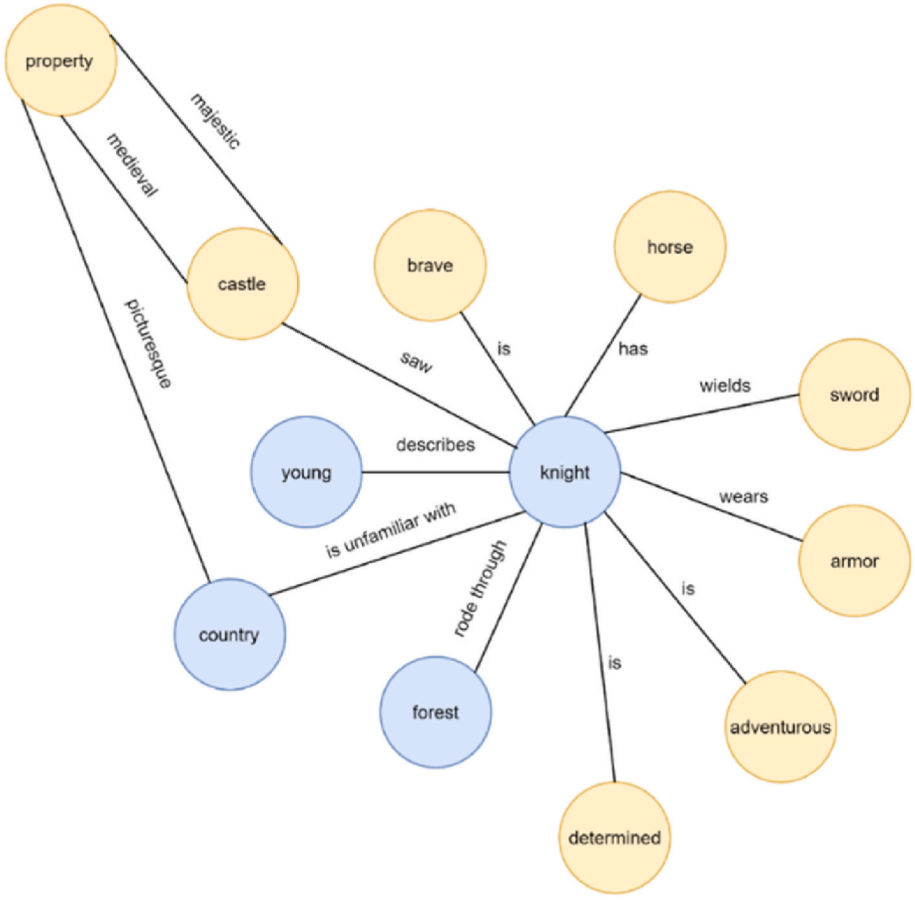

Fig. 5. Second sentence graph after inference.

4.1. Study 1 – Landscape Model simulation

Study 1 compares the activation scores generated by AMoC v4.0 to the activation scores reported in the foundational paper introducing the Landscape Model (Van den Broek et al., 1999). Both AMoC v4.0 and the Landscape Model conceptualize reading comprehension as a dynamic process unfolding across time. In addition, both models iteratively generate activation scores, which represent the predicted activation of nodes and edges in a readers' mental model of the text at each sentence of the text. As noted above, the AMoC v4.0 activation scores are generated using ChatGPT. In contrast, the activation scores in the Landscape Model paper were experimenter-generated based on assumptions and findings from text comprehension and memory research. The Landscape Model was instantiated in several experiments. For example, Van den Broek et al. (1996) gave participants a list of words after each sentence while reading and asked them to rate how strongly the words were related to the text. There was a strong correlation between participants' ratings and the predicted activation scores (r = 0.73). Previous versions of AMoC have been validated using the research on the Landscape Model. Namely, the activation scores given by AMoC v3.0 were strongly correlated with the activation scores given by the Landscape Model ($r = 0.76$; Corlatescu et al., 2023). In this experiment, we attempted to replicate past research on the Landscape Model, and test whether AMoC v4.0 produced similar activation scores to the Landscape Model. We hypothesized that the activation scores produced by AMoC v4.0 would strongly correlate with the activation scores presented in the Landscape Model paper. Consistent with our research questions, we also hypothesized that the activation scores produced by AMoC v4.0 would be more strongly correlated with the Landscape Model activation scores than those produced by AMoC v3.0.

4.1.1. Dataset

We consider the activation matrix from the original Landscape Model (Van den Broek et al., 1999) in which the concepts (i.e., nouns, verbs, adjectives) from the text or inferred were attributed an activation score by human evaluators. The scores were placed in a matrix where the columns represented the sentence index, and the lines represented the words. The values in this matrix were on a scale from 0 to 5, where 5 meant the highest activation possible, while 0 meant no activation at all. As the generation of inferred concepts differs between modeling methods and between readers, we only included inferred concepts that appeared in some form in our experiments.

4.1.2. Method

The analysis was run using AMoC v4.0 with the default parameters (i.e., a reader with average knowledge and skill). For each sentence, the activation of each word (text-based or inferred) was printed, and then the matrix was filled manually.

In order to make a clean comparison between the activation scores in the Landscape Model and the activation scores of AMoC v4.0, two changes had to be made to AMoC v4.0. First, there were some differences in the approaches meaning that the "activation" scores that AMoC v4.0 attributes to words start from 0 (most active) and go up as the activation decreases. This contrasts with the Landscape Model, where 5 is the highest activation while 0 means no activation. So, to make the results clear, we shifted AMoC v4.0 results to (5 – AMoC v4.0 score), converting the values so that they are on the same scale. Second, in the Landscape Model, verbs are treated as nodes. However, in AMoC v4.0, verbs are primarily found in the edges (similar to the CI model). Thus, verbs were assigned an activation score with the following procedure: the highest activation score of the nodes linked by the edge that the verb

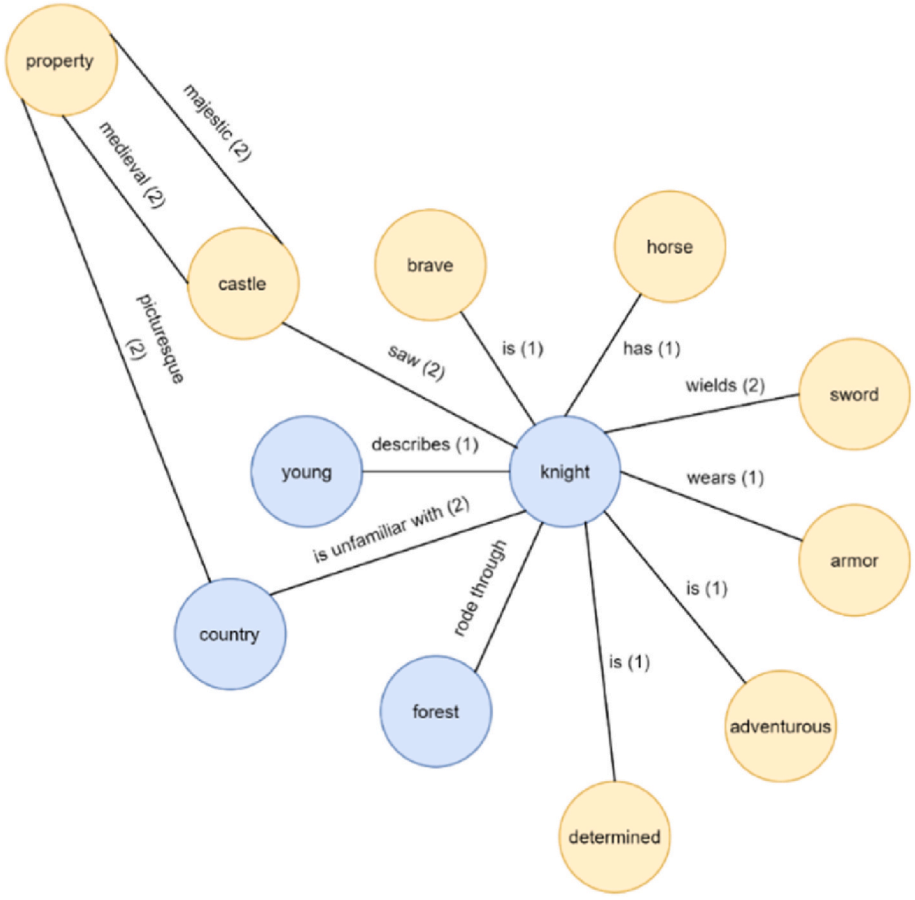

Fig. 6. Second sentence graph after reactivating/fading away step.

is part of, decayed by 0.5. For example, in the network, "knight – rode through (edge label) – forest", if "knight" has an activation score of 5 and the forest has an activation score of 4, then the verb "rode" will have an activation score of 4.5 (5–0.5 = 4.5).

4.1.3. Results

In order to test our hypothesis, a Spearman correlation was conducted between the sentence-by-sentence activation scores of the two models. The comparison was conducted at the node level, and Fig. 8 shows side-by-side the activation scores of each word as the text progresses. Note that inferred concepts not present in the text are marked in red.

Consistent with our hypothesis, there was a strong correlation between the activation scores of the Landscape Model and the activation scores of AMoC v4.0. The average correlation of word activation scores across all sentences was $r = 0.78$. The correlations by sentence are presented in Table 1.

4.1.4. Discussion

In Study 1, we examined the extent to which the activation scores predicted by the Landscape Model correlated with the activation scores generated by AMoC v4.0. We found that AMoC v4.0 and the Landscape Model provide similar activation scores on the same text. This indicates that the knowledge graph produced by produced by AMoC is consistent with the theoretical mental model generated by readers. In addition, this study replicated and improved on the previous version of AMoC, which ran the same test. The model produced by AMoC v4.0 was more strongly correlated ($r = 0.78$) to the Landscape Model activation scores than the model produced by AMoC v3.0 ($r = 0.76$; Corlatescu et al., 2023).

Visual examination of the activation values generated by AMoC and the Landscape Model suggests room for improvement both in AMoC and in the Landscape Model. For example, the Landscape Model does not predict that "dragon" will be activated in the last sentence. However, we believe the prediction from AMoC v4.0 is more likely – that the dragon is still activated at the end of the story because it was a central character. In addition, the Landscape Model has a specified decay rate of activation scores: 5, 3, 1.5, 0. The gradual decay is an assumption of the Landscape Model (Van den Broek et al., 1999) and is drawn from memory theory (Zhang & Luck, 2009). However, our results suggest that modeling memory for a concept as a sudden, instead of gradual, decay can result in similar behavioral findings. This phenomenon of sudden memory decay is known as, "Sudden Death" and has been found in studies on working memory (Donkin et al., 2015; Zhang & Luck, 2009). Further work in both comprehension theory and computational modeling of comprehension should investigate whether memory of text concepts decays gradually or suddenly.

On the other hand, AMoC v4.0 tends to remember and reactivate information that a human may not. For example, the words "rode" and "horse", are still activated/reactivated in the last sentences. While the knight was in a battle and "rode victoriously" to get the princess, the final sentences do not include explicit references to "rode" or "horse" and the knight's relationship to the princess is emphasized. In this case a human may not have the concepts "rode" or "horse" as strongly activated as the AMoC model predicts.

The final interesting observations from the AMoC matrix are that ChatGPT already thought about a princess even before it appeared in the text, given that the story already mentioned a knight and a dragon (see "princess" score for sentence 3); "sword" was also inferred given the context of the knight and the fight. This may simply be an artifact of using an LLM (e.g., the LLM training set included many texts about

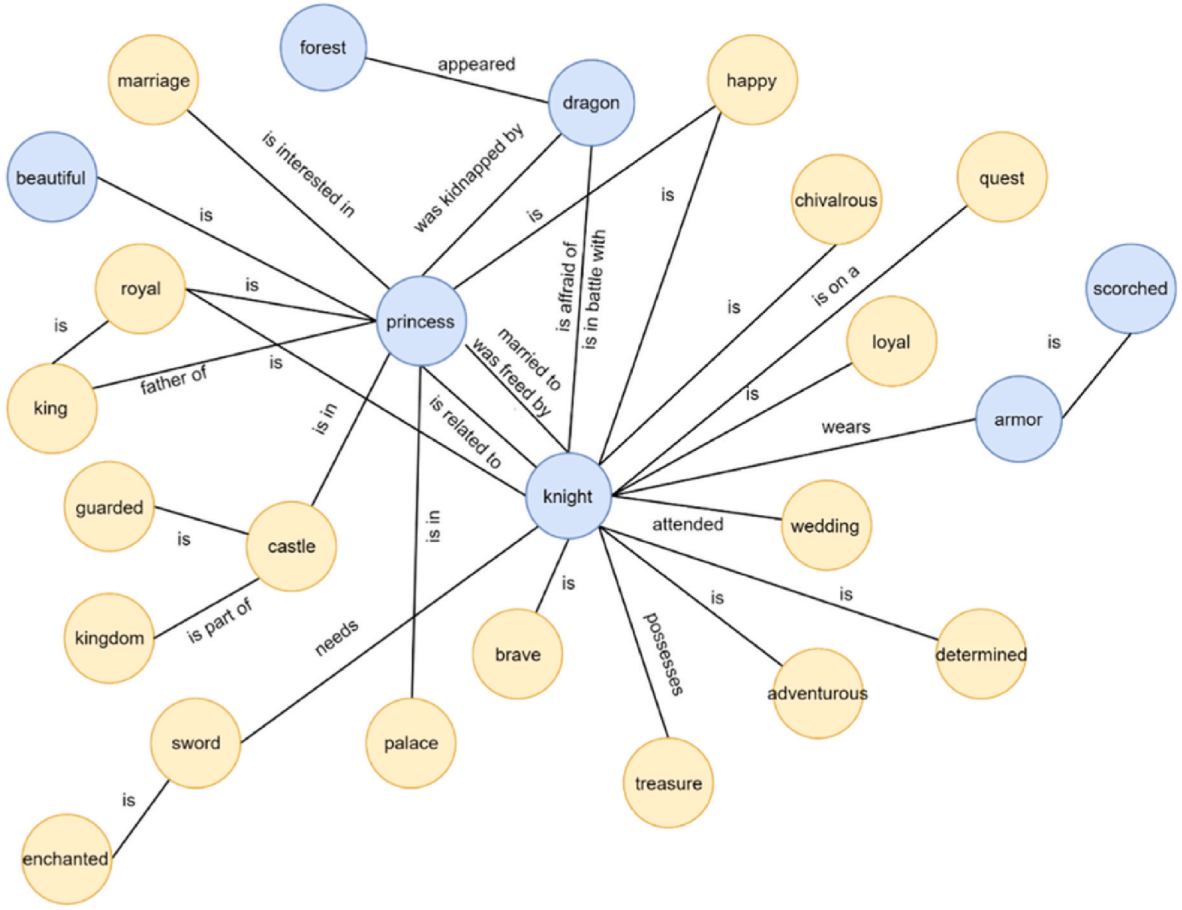

Fig. 7. Final active graph.

	AMoC v4.0												
word/sentence	1	2	3	4	5	6	7	8	9	10	11	12	13
knight	5	5	4	0	5	5	5	5	5	5	5	5	5
rode	4.5	4.5	0	0	0	0	0	0	0	4.5	4.5	0	0
horse	5	4	0	0	0	0	0	0	0	4	4	4	0
forest	5	4	3	0	4	4	4	4	0	0	0	0	0
unfamiliar	0	5	0	0	0	0	0	0	0	0	0	0	0
country	0	5	3	0	4	4	4	4	0	0	0	0	0
dragon	0	0	4	5	4	4	5	5	4	5	4	4	4
appeared	0	0	4.5	3.5	0	3.5	3.5	0	0	0	0	0	0
kidnapping	0	0	0	4.5	0	4.5	4.5	4.5	2.5	0	0	0	0
beautiful	0	0	0	5	4	0	0	0	0	0	0	0	0
princess	0	0	4	5	5	5	4	4	4	4	5	5	5
fought	0	0	0	0	0	0	0	4.5	4.5	4.5	4.5	4.5	4.5
armor	0	0	0	0	0	0	0	0	5	4	4	0	0
scorched	0	0	0	0	0	0	0	0	5	3	3	0	0
killed	0	0	0	0	0	0	0	0	0	4.5	3.5	0	0
sword	0	0	0	0	0	0	4	4	0	4	4	0	0
freed	0	0	0	0	0	0	0	0	0	0	4.5	0	0
thankful	0	0	0	0	0	0	0	0	0	0	0	4.5	3.5
married	0	0	0	0	0	0	0	0	0	0	0	0	4.5
fire	0	0	0	0	0	0	0	0	0	4	0	0	0

	Landscape												
word/sentence	1	2	3	4	5	6	7	8	9	10	11	12	13
knight	5	5	2.5	0	5	5	5	4	4	5	4	5	5
rode	5	3	1.5	0	0	0	0	0	0	0	0	0	0
horse	2	1	0	0	0	0	0	0	0	0	0	0	0
forest	5	4	2	0	0	0	0	0	0	0	0	0	0
unfamiliar	0	5	3	1.5	0	0	0	0	0	0	0	0	0
country	0	5	3	1.5	0	0	0	0	0	0	0	0	0
dragon	0	0	5	5	3	1.5	5	4	3	5	2.5	0	0
appeared	0	0	5	3	1.5	0	0	0	0	0	0	0	0
kidnapping	0	0	0	5	3	1.5	0	0	0	0	0	0	0
beautiful	0	0	0	5	2.5	3	1.5	0	0	0	0	0	0
princess	0	0	0	5	4	4	2	0	0	0	5	5	4
fought	0	0	0	0	0	0	0	5	3	3	1.5	0	0
armor	0	0	0	0	0	0	0	0	5	2.5	0	0	0
scorched	0	0	0	0	0	0	0	0	5	2.5	0	0	0
killed	0	0	0	0	0	0	0	0	0	5	3	1.5	0
sword	0	0	0	0	0	0	0	0	0	2	1	0	0
freed	0	0	0	0	0	0	0	0	0	0	5	3	1.5
thankful	0	0	0	0	0	0	0	0	0	0	0	5	3
married	0	0	0	0	0	0	0	0	0	0	0	0	5
fire	0	0	0	0	0	0	0	0	3	1.5	0	0	0

Fig. 8. AMoC v4.0 and Landscape Model Side-by-Side Activation View.

knights, dragons, and princesses), or perhaps the Landscape Model does not introduce inferred concepts early enough – it is reasonable to assume a story involving a knight and dragon would include a princess. Further research is needed to examine how the activation of inferred concepts differs between LLMs and humans.

Table 1
Spearman Correlations for the Word Activation Scores Between AMoC v4.0 and the Landscape Model by Sentence.

	Sentence Number												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Correlation	.77	.98	.66	.91	.69	.74	.68	.69	.93	.76	.68	.79	.79

4.2. Study 2 – probe word prediction

The second validation study is based on the first experiment from Keefe and McDaniel (1993). In the original experiment, 48 undergraduate introductory psychology students from Purdue University were asked to read a sentence aloud, and then they were shown a target word to read. The researchers measured the reading time (RT) in milliseconds of the target word. There were three conditions: predictive, explicit, and control. In the *predictive* condition, the sentence was related to the target word but did not explicitly contain the target word. In the *explicit* condition, the sentence contained the target word. Finally, in the *control* condition, the sentence had similar lexical elements to the predictive and explicit condition, but the sentence was not related to the target word. Table 2 presents example sentences for the target word 'sat' and the mean reading times for each condition from Keefe and McDaniel (1993). The mean reading time was reported for each condition, and the conclusion was that there was a statistical difference between the control and the other two conditions, while no statistical difference between predictive and explicit conditions was observed. This finding suggests that for both the predictive and the explicit conditions, the reader had activated the target word. However, for the control condition, the concept had not been activated.

In Study 2, we sought to test to what extent AMoC v4.0 aligns with the original results. Specifically, given the knowledge graph of a sentence, AMoC v4.0 was tested to understand how performant it is in differentiating between the three conditions. In addition, we compared the performance of AMoC v4.0 to AMoC v3.0 to see if the use of ChatGPT further improved the performance of the model. While neither AMoC v3.0 nor AMoC v4.0 require time to read, past research suggests that a word's activation in the mental model is inversely correlated to reading time. However, there is a difference between how AMoC v3.0 and AMoC v4.0 correspond to reading time. In AMoC v3.0, reading time corresponded to an activation score – that is, the relative activation of a probe word given a concept graph. In AMoC v4.0, reading time corresponds to a *relevance* score obtained from prompting ChatGPT to assess the knowledge graph and rate how well the probe word fits in the context of the graph. Consistent with our research questions, we hypothesized that the relevance scores produced by AMoC v4.0 would be significantly different between the control and predicative conditions and the control and explicit conditions. In addition, we hypothesized that the relevance scores generated by AMoC v4.0 would better correspond to the findings from Keefe and McDaniel (1993) than the activation scores produced by AMoC v3.0.

Table 2
Examples sentences and reading times (RT) as a function of condition in Keefe and McDaniel (1993).

Condition	Example Sentence Followed by the Probe 'sat'	*RT* (ms)
Control	The tired speaker moved the chair that was in his way and walked to the podium to continue his 3-h debate.	555
Predictive	After standing through the 3-h debate, the tired speaker walked over to his chair.	524
Explicit	After standing through the 3-h debate, the tired speaker walked over to his chair and sat down.	521

Note: *RT* = Average Reading Times is in milliseconds; The original target word was "sat" and was lemmatized to "sit" in the current simulation.

4.2.1. Dataset

The dataset comprised 40 sets of texts (i.e., a control, predicative, and explicit text) collected by Potts et al. (1988) and used in the experiments by Keefe and McDaniel (1993). It contained all the text categories as well as the sentence for each text. We modified the dataset by converting the text format into JSON to increase loading speed. Additionally, the target word was replaced with its lemma because the AMoC v3.0 graph only contains lemmatized words (e.g., "sat" became "sit").

4.2.2. Method

For this study, both AMoC v3.0 and AMoC v4.0 were used to generate Relevance Scores. The Relevance Scores were used to attempt a replication of the experimentally obtained results from Keefe and McDaniel (1993).

4.2.2.1. AMoC v3.0. AMoC v3.0 requires set parameters to generate a concept map. In our study, the parameters selected represented a skilled reader to reflect the experiment conducted by Keefe and McDaniel (1993), whose participants were undergraduate students from Purdue University. Thus, the parameters were.

- Transformer model: GPT2
- number of sentences remembered in their raw form: 1
- maximum active concepts in the reader's mind: 5
- maximum dictionary expansion or how many new concepts can be inferred: 7
- attention score threshold for new concepts: 0.3
- generation imagination factor (1 less imaginative, 4 more imaginative): 1
- weight decay percentage of each edge from sentence to sentence: 0.1
- whether to use AOE scores to scale the importance of a node: True

Similar to how the participants in the Keefe and McDaniel (1993) experiment read the target word, the model was allowed to "read" the target word as well. The target word was added as an additional sentence, meaning that the word was introduced to the graph connected to the other nodes using the attention scores, but there was no further inference in the model. The activation score used in the analyses was computed as the sum of activation scores of the edges connected to the node.

4.2.2.2. AMoC v4.0. In order to obtain a relevance score for the target word, we included an extra step in the AMoC v4.0 process. ChatGPT was prompted to assign a score (1–4) of relevance or connection to the story in the graph (see Appendix 1, Prompt 8). The four possible scores were.

1 – no connection or relevance to one or more ideas in the graph or the graph's story
2 – little connection or relevance to one or more ideas in the graph or the graph's story
3 – a clear connection or relevance to one or more ideas in the graph or the graph's story
4 – a strong connection or relevance to one or more ideas in the graph or the graph's story

4.2.3. Results

Two Linear Mixed Effects models (Pinheiro & Bates, 2000) were conducted predicting AMoC v3.0 activation scores and AMoC v4.0

relevance scores. A pairwise comparison between the conditions was performed using Tukey Contrasts for Multiple Comparisons of Means. The results shown in Table 3 provide the scores obtained by both AMoC v3.0 and AMoC v4.0, where the z (Z-score) is a measure of how many standard errors a parameter estimate is away from its null hypothesis value, and p-value indicates the probability of observing a Z-score as extreme as the one obtained if the null hypothesis were true.

4.2.4. Discussion

In this study, we sought to test to what extent AMoC v4.0 produced results consistent with the findings of Keefe and McDaniel (1993). Specifically, we were interested in how well AMoC v4.0 differentiated between sentences where a target word is activated by the text, either explicitly or implicitly, and sentences where the target word is not activated by the text. We generated activation scores for AMoC v3.0 and relevance scores for AMoC v4.0 for different sentence/target word pairs and conducted a linear mixed effects model to test differences in activation scores. Consistent with our hypothesis, we found that the AMoC v4.0 activation scores were significantly different based on the condition with the activation scores for the target word higher in the predictive and explicit conditions than in the control condition.

The first observation that can be drawn from this study is that the AMoC v3.0 model has an explicit bias. That is, if a word appears in a text, it will have higher activation than in any situation where it does not. Thus, in every scenario, the *explicit* category has a significantly higher activation in comparison to the other conditions. This finding is inconsistent with the findings of Keefe and McDaniel (1993) which suggest that the target word should have sufficient activation from context in the predicative condition. However, both *explicit* and *predictive* activation scores from the AMoC v3.0 graph were statistically different from the *control* scores, which corresponds to the findings of Keefe and McDaniel (1993).

In contrast, AMoC v4.0 does not have an explicit bias, and the results from AMoC v4.0 replicate the results obtained by Keefe and McDaniel (1993). This finding is consistent with our hypotheses and research questions that AMoC v4.0 more closely resembles human inference generation and mental representation of text compared to AMoC v3.0.

Although this study showed the increased capabilities of AMoC v4.0 in comparison to AMoC v3.0, the results can be considered as a validation for the AMoC v3.0 model as well since the general meaning of the text is preserved, even with an explicit bias.

4.3. Study 3 - recall of critical textual information

Our final study was based on the work of O'Brien et al. (1998) who compared two views of the memory representations while reading: the memory-based text processing view and the here-and-now view. In their experiments, there were three types of texts: consistent, inconsistent, and qualified. The consistent version of the text described a trait of the protagonist that was in line with a target action taken by them later on. In contrast, the inconsistent version described a target action that was not in line with the protagonist's described trait. Lastly, the qualification version was similar to the inconsistent version, but it included a statement that limited the circumstances under which the described trait of the protagonist was applicable.

We tested AMoC with four of the five experiments presented in the paper by O'Brien et al. (1998). The difference between the experiments was the qualified versions of the story, and we named them with respect to the number of the experiment (qualified1, qualified2, qualified3, and qualified4). Each example also has an introduction, a filler text that does not contain relevant information for the decision, two target sentences, a closing statement, and a comprehension question that needs to be assessed. The qualification revolved around the circumstances in which the main character (Mary) would adhere to her vegetarian diet. Appendix 3. Study 3 example shows an example from O'Brien et al. (1998).

The O'Brien et al. (1998) experiments involved undergraduate students from the University of New Hampshire. They were given texts in a random manner, and their task was to read these texts that were displayed on a screen sentence by sentence. During this time, the researchers measured the reading times for target sentences. According to theoretical models of comprehension, the consistent condition should have the shortest reading times, followed by the qualification conditions, and the inconsistent condition should have the longest reading times. This implies that the consistent scenario will have a lower reading time than the inconsistent case. Table 4 shows the results from the O'Brien et al. (1998) paper, where the participants' mean reading times (in seconds) are presented. There were two target sentences for which reading time was recorded. The first target sentence introduced conflicting information, whereas the second was consistent with the narrative. This difference is reflected in the results, as the mean reading times for the second target sentence were not significantly different across conditions. O'Brien and colleagues did not find statistical significance for each comparison; however, the overall pattern of results supported their hypothesis that consistent information would be read faster than inconsistent information (see Table 4).

In Study 3, we sought to explore to which extent these findings are replicable using AMoC v4.0. In the O'Brien et al. (1998) experiment, differences in reading time were only found for the first target sentence. Thus, our experiment focused on the performance of AMoC v4.0 in differentiating between the *probability* scores of the first target sentence. We hypothesized that the probability scores generated by AMoC v4.0 would be significantly different as a function of condition, such that the target sentence would have a higher probability score in the consistent condition than in the inconsistent condition. In addition, we tested the performance of AMoC v4.0 to the performance of AMoC v3.0 and hypothesized that AMoC v4.0 would more closely resemble the findings of O'Brien et al. (1998) compared to AMoC v3.0.

4.3.1. Dataset

The corpus was comprised of 22 texts, and each text had multiple versions (see Table 5). Each text version contained an introduction, an elaboration that related to the target sentence in one of three ways (consistent, qualification, inconsistent), a filler text, two target sentences, and a closing paragraph. In the corpus, there was a consistent and inconsistent version for each of the 22 texts, with only a subset of the texts having qualification versions (see Table 5). The initial dataset was

Table 3
Between-group comparisons of the LME model predicting AMoC score ratings.

Between-Group Comparisons	AMoC v3.0		AMoC v4.0	
	z	p	z	p
Explicit/Control	8.39	<.001**	5.00	<.001**
Predictive/Control	1.99	<.001**	3.87	<.001**
Predictive/Explicit	−6.39	<.001**	−1.129	0.25

Table 4
Participants mean reading times (in seconds) from O'Brien et al. (1998).

Sentence by Experiment	Passage condition		
	Consistent	Qualification	Inconsistent
Experiment 1			
Sentence 1	1.731	1.864	2.042
Sentence 2	1.704	1.852	1.908
Experiment 2			
Sentence 1	1.749	1.851	2.040
Sentence 2	1.792	1.794	1.896
Experiment 3			
Sentence 1	1.840	1.950	2.037
Sentence 2	1.787	1.882	1.912
Experiment 4			
Sentence 1	1.903	2.015	2.162
Sentence 2	1.891	1.923	2.102

Table 5
Number of texts with each type of elaboration.

Type of elaborations	Number of texts
Consistent & Inconsistent	22
Qualified 1 & Qualified 2	18
Qualified 3 & Qualified 4	15

transformed into a JSON format to reduce processing time.

4.3.2. Method

For each text, the concept and knowledge graphs created consisted of the introduction sentence and one of the sentences by condition (*consistent*, *inconsistent*, *qualified 1*, *qualified 2*, *qualified 3*, *qualified 4*), the filler sentences were not added to the graphs. The only target sentence considered was the first one (*Target Sentence 1*).

In Study 2, we directly compared activation scores to the reading times of the target word. However, the testing input is a sentence in the current study; as such, a different comparison was employed, later described in detail.

4.3.2.1. AMoC v3.0 probability scores. The GPT2 model employed in AMoC v3.0 assigns a probability to each token in its dictionary, indicating the likelihood of it being the next token in the sequence. Token probability can be extended to word probability as each word can contain one or multiple tokens. For a single-token word, the word probability equals the token probability. For multiple-token words, the word probability is computed as the product of the tokens inside the word.

Using this capability, GPT2 was used to assign a score for each type of text, having as input the active concepts in the AMoC v3.0 concept graph and computing the probability of each word from the target sentence iteratively. For example, if the target sentence had n words, the (n-2) word score was computed using the AMoC v3.0 graph concepts and all the (n-1) words in the target sentence. Then, the final score was computed as the product of all the probabilities. One consideration was whether all words or only content words should be used in the computation of the final score. In this paper we included only the probabilities of the content words in the computation of the final score. However, the results were tested with both configurations and the pattern was the same.

Given the difficulty of the task, the AMoC v3.0 model used parameters that provided it with high knowledge and high comprehension skills.

- Transformer model: GPT2
- number of sentences lookback: 2
- maximum active concepts: 7
- maximum dictionary expansion: 9
- attention score threshold: 0.3
- generation imagination factor: 1
- weight decay percentage: 0.1
- use AOE: True

4.3.2.2. AMoC v4.0 probability scores. In order to generate probability scores from AMoC v4.0, ChatGPT prompts were used to directly evaluate the relationship between the knowledge graph and the target sentence. ChatGPT was prompted to list the edges that supported the target sentence, as well as the ones that contradict it (see Appendix 1, Prompt 9). For example, if the target sentence was, "Mary ordered a salad", the edge in the knowledge graph "Mary - is - vegetarian" supports the statement. On the contrary, if the target sentence was "Mary ordered a burger" the same edge contradicts the statement. Three values were obtained from the lists that ChatGPT produced. The total number of edges that support, the total number of edges that contradict, and the difference between the number of supporting edges and contradicting edges. Using these numbers, we analyzed if there was a significant statistical difference between the types of texts. The AMoC v4.0 used the default parameters.

4.3.3. Results

A Linear Mixed Effects models (LME) was conducted predicting differences in the AMoC v3.0 probability scores from text conditions. The predictors were the word probabilities given by the previously mentioned algorithm and the predicted value was the text class. Table 6 presents the results.

A Linear Mixed Effects model was conducted to predict the differences in text category by the number of edges that either supported the target sentence, contradicted the target sentence, or the difference between the number of supporting and contradicting edges (see Table 7).

Even though the statistical tests have more argumentation value and represent the correct test to run, we also show the mean values for every situation described in the experiment in Table 8. Recall that all the texts had consistent and inconsistent text types, but only a part had the combination of qualified 1 and 2, as well as the qualified 3 and 4. Refer to Table 5 to see how many texts were considered for each category.

4.3.4. Discussion

Study 3 sought to replicate the findings of O'Brien et al. (1998) using AMoC v3.0 and AMoC v4.0. We generated probability scores from AMoC v3.0 and AMoC v4.0 and tested if the probability scores differed as a function of text condition (consistent, inconsistent, qualified). The current experiment was the most difficult experiment for AMoC v3.0 and AMoC v4.0 which helps us better understand the models' capabilities, as well as their limitations.

In the case of AMoC v3.0, no configuration of the test or the model seemed to replicate the results of O'Brien et al. (1998). Even though, in theory, this was the easiest test case since the information required to correctly assess the difference between the statements was "close" in terms of processing steps. We believe the reason behind this issue is the nature of the text – the discriminating information was mentioned only once, and as the text progresses, the information may fade away in the AMoC concept graph. For example, in the sample showcased in Appendix 3. Study 3 example, the information that Mary is a vegetarian is mentioned only once, and the minimum distance between the information and the target sentence is two sentences. Thus, while the vegetarian does not remain in focus for multiple sentences in the AMoC model, a human can retain this specific information with more ease. As such, we believe that we have reached the limit of AMoC v3.0, meaning that the knowledge is lost using this method, and some indirect inference of that knowledge is not sufficient.

AMoC v4.0 achieved a higher performance in this study. The results showed a clear difference between the type consistent and the type inconsistent. The LME test showed a significant statistical difference between the two text types, no matter what metric was considered: number of support edges, number of contradict edges, or the difference between the two. Other than that, results do not exhibit statistical differences between texts except for a few pairings. The difference between the number of supporting edges and the number of contradicting edges seemed to differentiate between consistent and the other types of texts.

Table 8 provides additional insights into the capabilities of AMoC v4.0. Even though the means are not statistically sufficient for a comparison between texts, the results are in line with the expectations.

Table 6
AMoC v3.0 Study 3 Elaboration Comparison Results.

Pair	z	p
inconsistent – consistent	1.55	0.12
qualified1 – inconsistent	−1.01	0.30
qualified2 – inconsistent	0.64	0.52
qualified3 – inconsistent	0.76	0.44
qualified4 – inconsistent	0.39	0.69

Table 7
AMoC v4.0 Study 3 Elaboration Comparison Results.

Type1	Type2	Support		Contradict		Difference	
		t	p	t	p	t	p
C	I	3.659	.001**	2.069	.044*	−2.979	.007**
C	Q1	2.155	.045*	−0.668	.513	−2.080	.037*
C	Q2	1.900	.075	−1.599	.128	−2.172	.044*
C	Q3	3.506	.003**	−1.394	.185	−2.736	.010*
C	Q4	1.960	.070	−0.062	.951	−1.424	.176
I	Q1	−1.083	.294	2.221	.033*	2.377	.029*
I	Q2	1.425	.172	1.430	.161	1.567	.136
I	Q3	1.123	.280	0.366	.717	0.553	.584
I	Q4	−1.996	.066	1.454	.168	1.682	.103

Concretely, when comparing consistent and inconsistent, the number of supporting edges for the consistent case is higher than in the inconsistent case; reversely, the number of contradicting edges for the inconsistent case is higher than for the consistent case. Additionally, the qualified texts should be somewhere in the middle between the two extremes, consistent and inconsistent. And this can clearly be observed from the means as well. The relationship between them in the case of support is *consistent* > *qualified* > *inconsistent*, while the contradiction is *consistent* < *qualified* < *inconsistent*. We consider this result as being a good indicator of the quality of the model, a quality that is higher than in the case of AMoC v3.0.

5. General discussion

In this study, we introduced an updated version of AMoC, namely v4.0, which combines past research in comprehension theory with recent advancements in NLP, specifically LLMs – i.e., ChatGPT - to generate graphs reflecting the reader's mental representation of the text. We sought to answer research questions on to which extent the use of ChatGPT improved on previous versions of AMoC. Specifically, we were interested in evaluating to what extent conceptualizing the graph as a set of relationships between concepts and using ChatGPT to generate the inferences in the graph more closely resembled humans' mental representation of texts compared to previous versions of AMoC. We tested our research questions by comparing the performance of different versions of AMoC to empirical results obtained in research on reading comprehension that involved human subjects.

In this paper, we first described the processes of AMoC v4.0 and demonstrated them in a use case. The use case immediately demonstrates the advantages of using ChatGPT, as the extracted edges from the text contained all the textual information: "young, describes, knight" and "knight, rode through, forest." Furthermore, the inferred edges were closely aligned with the story. The inferences generated by ChatGPT correctly included "horse" because the knight is "riding". Additionally, common *characteristics* of knights are brought to the front: he has a sword, he has armor, and he is brave, adventurous, and determined.

While inspecting the second sentence from the use case, we observe similar patterns: the edge generated from the text makes sense. The inferred edges are related to the text, but the last three, "castle, majestic, property", "country, picturesque, property", and "castle, medieval, property" are not following the rule of having a verb or a verb phrase as an edge. Probably these edges were best rephrased as, for example, the castle is majestic. This may be a limitation of the model: sometimes ChatGPT will generate an edge that may contain some errors in terms of the format. Additionally, there will always be some randomness using ChatGPT. Even with the temperature of 0, the ChatGPT parameter that should make the results somehow consistent, the generated concepts differ from one run to the other, but, as we saw in our multiple studies, they tend to be consistent in terms of relationship with the story. For the reactivating stage (and fading away, respectively), ChatGPT tends to dismiss a big portion of the edges given the current text. For example, only the knight who wields a sword was chosen to be reactivated while the other previous edges faded away. ChatGPT tends not to assume things unless they are specified clearly; for example: if the text says that "Jim is a teacher." and elaborates on this idea, then if ChatGPT is asked "Is Jim a mechanic?", the answer will be something in the lines of "I don't know, he can be." In contrast, people will tend to answer "No, because he is a teacher." This behavior may impact the assessment of which edges are relevant to the text, meaning that the process is stricter. Given this strictness, we selected the parameter of truly deactivating an edge of being 2, and when an edge fades away, this score drops by 1. When it reaches 0, it becomes deactivated. In other words, the information is not deactivated right away, but it may linger in memory for one more sentence if, of course, it is not reactivated afterward.

When reaching the last sentence of the use case, the graph had considerably grown since the model went through the full text. The generated edges are related to the text, and the reactivated edges are in line with the whole story presented until the final sentence.

To answer the proposed research questions, three validation studies were conducted testing AMoC v4.0 and AMoC v3.0 on established findings in reading research. Study 1 was based on the Landscape Model simulation, and we found that the automated activation scores produced by AMoC v4.0 have a strong correlation with the human activation scores of the Landscape Model and outperform AMoC v3.0 by a slight margin. The main difference between the activation scores of AMoC v4.0 and the Landscape model is that AMoC v4.0 retains concepts for more sentences than the Landscape model. For example, the dragon loses activation quickly in the Landscape Model when the text does not explicitly mention the dragon. In addition, AMoC v4.0 may activate inferred concepts, such as the sword for the knight, earlier than the Landscape Model. Overall, there is a substantial overlap between the two models, which suggests that AMoC v4.0 can accurately approximate the mental representation of the text predicted by the Landscape Model.

The second study was based on the work of Keefe and McDaniel (1993). The knowledge graph built by AMoC v4.0 is a good replica of the information presented in a text. Additionally, AMoC v3.0 had an explicit bias such that the explicit text condition outperformed the predicative text condition, inconsistent with the findings of Keefe and McDaniel (1993). In contrast, AMoC v4.0 showed no difference between the explicit and predicative text conditions, suggesting that the AMoC v4.0 model more closely aligns with the work of Keefe and McDaniel (1993) and better corresponds to human text comprehension than AMoC v3.0.

Finally, the third study tested the capabilities of AMoC v4.0 on the findings of O'Brien et al. (1998). They found that reading time was

Table 8
AMoC v4.0 Means per Elaboration Type.

Type	C/I		C/I/Q1/Q2		C/I/Q3/Q4	
	Support Mean	Contradict Mean	Support Mean	Contradict Mean	Support Mean	Contradict Mean
C	8.09	7.27	6.89	5.39	8.47	7.47
I	2.95	18.50	2.72	20.4	2.73	15.67
Q1	–	–	4.00	6.72	–	–
Q2	–	–	3.94	10.83	–	–
Q3	–	–	–	–	3.87	13.33
Q4	–	–	–	–	4.60	7.67

longer in the cases where contradictions or partial contradictions (i.e., qualifications) appeared, even when the contradicting information preceded the target sentence by 2 or more sentences. In our study, AMoC v3.0 was unable to differentiate between any of the text conditions, suggesting that the performance of AMoC v3.0 does not correspond to human performance in this task. In contrast, AMoC v4.0 successfully identified differences between consistent and inconsistent texts. However, there were mixed results for comparisons with the qualified texts. Overall, the findings from all three validation studies suggest that while the predictive modeling process of AMoC v4.0 generates graphs that more closely align with human mental representations than AMoC v3.0, there are still areas for improvement.

Overall, AMoC v4.0 showcased the potential use of semantic networks and cognitive processes for constructing meaning from textual content. From a theoretical perspective, AMoC v4.0 deepens our comprehension of the intricate web of semantic connections within cognitive processes. It illuminates the processes of knowledge construction and the making of meaning, which are at the core of learning. Moreover, AMoC v4.0 presents a longitudinal view of cognitive processes across the text and models reading comprehension given specific parameters as input, thus facilitating the simulation of various conditions.

6. Limitations

There were two primary limitations of this study. The first was the stochastic nature of ChatGPT. Two runs where the temperature is set to 0 in ChatGPT will provide similar, but not identical results. However, in the aggregate, patterns in ChatGPT responses can be observed. Additionally, the edges generated by ChatGPT can be poorly formatted, or the information is not present where it should be. The first one can be dealt with, in theory, programmatically: for example, if a generated edge does not have three components (i.e., node1, edge label, and node2) then discard the edge. The other ones are harder to spot, as we saw in the example of “castle, medieval, property”. Better language models, including open-source alternatives that will be released in the future, may diminish these problems. The second limitation was that our study was conducted without human participants. While our aim was to compare the predictions of AMoC to the actual behavior of human readers, we elected to compare the predictions of AMoC to previously published empirical findings in reading research. However, testing of human readers through protocols such as concept mapping or sorting tasks (Lawless et al., 1998) may demonstrate that there may be fundamental differences (or similarities) in the way humans and AMoC v4.0 generate representations of text.

One concern is that teachers or researchers must pay for API calls in order to use AMoC v4.0 at scale. While we considered utilizing open-source LLM models (one example is GPT-Neox - Black et al., 2022) for AMoC v4.0, the results were not satisfactory for two reasons. First, while the other LLMs perform comparably in general chat capabilities, in our testing, ChatGPT outperformed other LLMs in the concrete scenarios encountered when building AMoC v4.0. Second, open-source LLM models require a powerful GPU to perform inference tasks, a fact that would render AMoC v4.0 unusable for teachers or researchers who do not have access to this kind of equipment. In comparison, ChatGPT can be run from any PC that has an internet connection, with an affordable cost (0.002$/1000 tokens, approximately 750 words for English) for API calls. In addition, at the moment of writing, the newest version of ChatGPT, GPT-4 (OpenAI, 2023a), has been released and is outperforming the 3.5 version in benchmarks. The model was in closed beta when we conducted our experiments, and it was 10 times more expensive to use. Nevertheless, experimenting with newer versions can be easily done using the provided notebook (Corlatescu, 2023) and an OpenAI API key.

7. Conclusions and future work

The current study argues how psychological theory about reading comprehension can be combined with advancements in NLP to develop predictive models of students’ reading behavior. Furthermore, the study provides initial evidence in the form of validation studies that the predictions generated by an LLM-based reading comprehension model correspond to empirical results testing humans’ reading comprehension. AMoC v4.0 both provides an understanding of the cognitive processes behind reading and generates predictions about the trajectory of student learning. Thus, AMoC v4.0 has important practical applications within the field of education. First, AMoC v4.0 can create personalized profiles for students by understanding how they process information and the concepts triggered during reading. Educators can then utilize these profiles to select learning materials for individual students, fostering more engaging and effective learning experiences. Additionally, prompt engineering may be used to simulate other personas while using ChatGPT. For example, previous research has found that ChatGPT will produce different outputs if it is prompted to act as a specific field expert (White et al., 2023). This way, AMoC v4.0 can be fine-tuned to reflect contextual differences in education locales (e.g., schools, ages/grades, demographic differences). Second, the individualization process can also be approached by curriculum designers. AMoC v4.0 can generate predictions about students’ reading comprehension at a large scale, enabling designers to test the expected comprehension and retention of educational materials such as course content and textbooks. Finally, the graphs generated by AMoC v4.0 provide data for reading comprehension researchers. Both the intermediate and final graphs can be examined to uncover how specific aspects of grammatical style (e.g., word repetition across sentences) or language (e.g., increased use of connectives) affect reading comprehension.

The current version of AMoC may also be used to simulate long-term memory while reading. Currently, AMoC simulates the memory of readers for a single, short text (i.e., approximately 500 words or fewer). However, many existing texts in schools (e.g., textbooks, novels) have word counts far exceeding the current length of texts tested in AMoC. The abbreviated, representational nature of the AMoC graphs may allow AMoC to retain a greater number of concepts over a longer section of texts. This may have implications both for reading comprehension theory and modeling, as well as the “memory” of LLMs. Current modeling of reading comprehension typically takes place on single, short texts (Davoudi & Moghadam, 2015; Elleman & Oslund, 2019). Thus, predictive modeling of longer texts may reveal new discoveries about reading comprehension processes and memory for long texts. In addition, current LLMs retain all the prompts (input) in a conversation up to a specific length. After the conversation exceeds the specified length, LLMs drop the information from the beginning of the conversation. While saving the whole history of the conversation may be beneficial for LLMs to give accurate responses, the retention factor cannot be scaled to infinity. A conceptual graph such as the one generated by AMoC may be a more suitable solution, as it allows the LLM to retain information about the conversation in a subgraph – rather than retaining the complete text of the conversation. Augmenting LLMs with a “memory” based on conceptual graphs may be an important step in equipping LLMs with a memory similar to humans.

Furthermore, AMoC v4.0 can be used in other experiments for cognition. As noted in the limitations section, an initial replication study could utilize different comprehension tasks to better compare the mental representations generated by humans to the predictive graphs generated by AMoC. The second research direction consists of evaluating LLMs in the context of rebuilding the initial text starting from the final knowledge graph or the intermediate knowledge graphs. Similar to how students are tested in a classroom when they are asked to summarize a text that they have just read, ChatGPT can be asked to recreate the story or to summarize the story based on the AMoC v4.0 concept graphs. This task would test whether the concept graphs could be utilized as long-term

memory inputs for LLMs. Finally, an alignment study could be conducted matching AMoC v4.0's parameters to readers with varying reading skill levels based on predicted and actual reading comprehension.

Overall, AMoC v4.0 is a significant step forward in the use of LLMs to generate predictive models of students' reading behavior. AMoC v4.0 generates graphs that more closely correspond to readers' mental representation of texts, which in turn allows for better prediction of reading comprehension outcomes compared to previous models of student reading. The success of AMoC v4.0 provides researchers with a deeper understanding of the cognitive processes in reading tasks and better identification of students' individual learning trajectories.

CRediT authorship contribution statement

Dragos-Georgian Corlatescu: Writing – original draft, Visualization, Software. **Micah Watanabe:** Writing – original draft, Resources, Methodology, Data curation. **Stefan Ruseti:** Writing – review & editing, Methodology, Conceptualization. **Mihai Dascalu:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Danielle S. McNamara:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Ministry of Research, Innovation, and Digitalization, project CloudPrecis, Contract Number 344/390020/ September 06, 2021, MySMIS code: 124812, within POC, the Institute of Education Sciences (NSF R305A130124, R305A190063), the U.S. Department of Education, and the National Science Foundation (NSF REC0241144; IIS-0735682). The opinions expressed are those of the authors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chb.2024.108154.

Appendix 1. Prompts

This additional material presents the prompts used in building AMoC as well as evaluating it. Every prompt will have a placeholder in the format **info** that is replaced programmatically for each specific case.

1 New relationships - the first sentence

I want to build a knowledge graph using the provided text. The graph should consist of two types of nodes: concept nodes and property nodes. Concepts nodes represent objects or persons from the story and are generally represented by nouns in the text. Property nodes describe the concepts nodes and are generally represented by adjectives in the text. An edge connects a concept to another concept or a concept to a property, and it is described by a relationship between the connected nodes.

The format for representing the graph is as follows: ('concept1′, 'relation (edge)', 'concept2′) or ('concept1′, 'relation (edge)', 'property1′).

I already extracted the nodes from the text and these ones you should use:

nodes_from_text

I want you to tell me the relationships (edges) between the nodes given the text. The text is:

text

List them as a Python list and do not provide additional explanations.

2 Infer concepts and properties – the first sentence

I want to build a knowledge graph using the provided text. The graph should consist of two types of nodes: concept nodes and property nodes. Concepts nodes represent objects or persons from the story and are generally represented by nouns in the text. Property nodes describe the concepts nodes and are generally represented by adjectives in the text. An edge connects a concept to another concept or a concept to a property, and it is described by a relationship between the connected nodes.

The format for representing the graph is as follows: ('concept1′, 'relation (edge)', 'concept2′) or ('concept1′, 'relation (edge)', 'property1′).

I already extracted the nodes from the text:

nodes_from_text

PromptsGenerate a list of concepts and properties that are in line with the overall coherence and sense within the given text, but they are not in the text. The text is:

text

List them in the following format:

{

"concepts": [concept1, concept2, …],

"properties": [property1, property2, …]

}

3 Generate new inferred relationships – the first sentence

I want to build a knowledge graph using the provided text. The graph should consist of two types of nodes: concept nodes and property nodes. Concepts nodes represent objects or persons from the story and are generally represented by nouns in the text. Property nodes describe the concepts nodes and are generally represented by adjectives in the text. An edge connects a concept to another concept or a concept to a property, and it is described by a relationship between the connected nodes.

The format for representing the graph is as follows: ('concept1′, 'relation (edge)', 'concept2′) or ('concept1′, 'relation (edge)', 'property1′).

I already extracted the nodes from the text:

nodes_from_text

And I already have additional concepts and properties that are in line with the overall coherence and sense within the given text:

concepts: *concepts*

properties: *properties*

Create relationships between the nodes from the text and the new concepts and the new properties (as you see fit). The text is:

text

Provide them in the following format:

{

"concept_relationships": [concept_relation1, concept_relation2, …],

"property_relationships": [property_relation1, property_relation2, …]

}

4 New relationships – general case

I want to build a knowledge graph using the provided text. The graph should consist of two types of nodes: concept nodes and property nodes. Concepts nodes represent objects or persons from the story and are generally represented by nouns in the text. Property nodes describe the concepts nodes and are generally represented by adjectives in the text. An edge connects a concept to another concept or a concept to a property, and it is described by a relationship between the connected nodes.

The format for representing the graph is as follows: ('concept1′, 'relation (edge)', 'concept2′) or ('concept1′, 'relation (edge)', 'property1′).

I already extracted the nodes from the text and these ones you should use:

nodes_from_text

I also have the knowledge graph:

Nodes (with their types, and a score of how central they are in the story (0 is most central, 1 less central, etc.)):

nodes_from_graph

Edges (relationships between the nodes):

edges_from_graph

I want you to tell me the relationships (edges) between the nodes from the text themselves. And also between the nodes from the text and the other nodes from the graph (here prioritize the relationships based on the score). The text is:

text

List them as a Python list and do not provide additional explanations.

5 Infer concepts and properties – general case

I want to build a knowledge graph using the provided text. The graph should consist of two types of nodes: concept nodes and property nodes. Concepts nodes represent objects or persons from the story and are generally represented by nouns in the text. Property nodes describe the concepts nodes and are generally represented by adjectives in the text. An edge connects a concept to another concept or a concept to a property, and it is described by a relationship between the connected nodes.

The format for representing the graph is as follows: ('concept1′, 'relation (edge)', 'concept2′) or ('concept1′, 'relation (edge)', 'property1′).

I already extracted the nodes from the text:

nodes_from_text

I also have the current knowledge graph:

Nodes (with their types, and a score of how central they are in the story (0 is most central, 1 less central, etc.)):

nodes_from_graph

Edges (relationships between the nodes):

edges_from_graph

Generate a list of concepts and properties that are in line with the overall coherence and sense within the given text and the knowledge graph, but they are not in the text. The text is:

text

List them in the following format and explain the role of the text and the knowledge graph in the decision-making process:

{

"concepts": [concept1, concept2, …],

"properties": [property1, property2, …]

}

6 Generate new inferred relationships – general case

I want to build a knowledge graph using the provided text. The graph should consist of two types of nodes: concept nodes and property nodes. Concepts nodes represent objects or persons from the story and are generally represented by nouns in the text. Property nodes describe the concepts nodes and are generally represented by adjectives in the text. An edge connects a concept to another concept or a concept to a property, and it is described by a relationship between the connected nodes.

The format for representing the graph is as follows: ('concept1′, 'relation (edge)', 'concept2′) or ('concept1′, 'relation (edge)', 'property1′).
I already extracted the nodes from the text:
nodes_from_text
I also have the current knowledge graph:
Nodes (with their types, and a score of how central they are in the story (0 is most central, 1 less central, etc.)):
nodes_from_graph
Edges (relationships between the nodes):
edges_from_graph
And I already have additional concepts and properties that are in line with the overall coherence and sense within the given text:
concepts: *concepts*
properties: *properties*
Create relationships between the nodes from the text and the new concepts and the new properties (as you see fit). The text is:
text
Provide them in the following format:
{
"concept_relationships": [concept_relation1, concept_relation2, …],
"property_relationships": [property_relation1, property_relation2, …]
}

7 Extract relevant edges

You have the following edges from a knowledge graph in the format: node - edge - node.
edges
As you can see, they are numbered. Tell me what edges are related to/support/contradict the following text.
text
Provide them in the following format (a list of numbers):
[1, 2, 3, …]

8 Keefe Experiment Prompt

You have the following edges from a knowledge graph in the format: node - edge - node.
edges_from_graph
Using the graph and the story that the graph says, for the word "*word*" assign a score between 1 and 4 with the following meaning.

1 no connection or relevance to one or more ideas in the graph but pay attention to the story as well
2 little connection or relevance to one or more ideas in the graph but pay attention to the story as well
3 a clear connection or relevance to one or more ideas in the graph but pay attention to the story as well
4 a strong connection or relevance to one or more ideas in the graph but pay attention to the story as well

9 O'Brien Experiment Prompt

You have the following edges from a knowledge graph in the format: node - edge - node.
edges
Tell me what edges su
and provide the response in the following format (the number representing the number of the edge):
{
"support": [1, 2, 3, …],
"contradict": [1, 2, 3, …]
}

Appendix 2. Landscape Model Knight Story

A young knight rode through the forest. The knight was unfamiliar with the country. Suddenly, a dragon appeared. The dragon was kidnapping a beautiful princess. The knight wanted to free the princess. The knight wanted to marry the princess. The knight hurried after the dragon. The knight and the dragon fought for life and death. Soon, the knight's armor was completely scorched. At last, the knight killed the dragon. The knight freed the princess. The princess was very thankful to the knight. The princess married the knight.

Appendix 3. Study 3 example

Table 9
Sample from O'Brien et al. (1998).

Entry Keys	Text
Introduction	Today, Mary was meeting a friend for lunch. She arrived early at the restaurant and decided to get a table. After she sat down, she started looking at the menu.
Consistent Elaboration	This was Mary's favorite restaurant because it had fantastic junk food. Mary enjoyed eating anything that was quick and easy to fix. In fact, she ate at McDonalds at least three times a week. Mary never worried about her diet and saw no reason to eat nutritious foods.

(*continued on next page*)

Table 9 (*continued*)

Entry Keys	Text
Inconsistent Elaboration	This was Mary's favorite restaurant because it had fantastic health food. Mary, a health nut, has been a strict vegetarian for ten years. Her favorite food was cauliflower. Mary was so serious about her diet that she refused to eat anything which was fried or cooked in grease.
Qualified Elaboration. Experiment 1	This was Mary's favorite restaurant because it had fantastic health food. Mary, a health nut, has been a strict vegetarian for ten years. Her favorite food was cauliflower. Mary was so serious about her diet that she refused to eat anything which was fried or cooked in grease. Nevertheless, Mary never stuck to her diet when she dined out with friends.
Qualified Elaboration. Experiment 2	This was Mary's favorite restaurant because it had fantastic health food. Mary, a health nut, has been a strict vegetarian for ten years. Her favorite food was cauliflower. Mary was so serious about her diet that she refused to eat anything which was fried or cooked in grease. Nevertheless, Mary never stuck to her diet when she dined out with friends because she enjoyed eating meat occasionally.
Qualified Elaboration. Experiment 3	As she was waiting, Mary recalled that this had been her favorite restaurant because it had fantastic health food. Mary recalled that she had been a health nut and a strict vegetarian for about ten years. Back then, her favorite food had been cauliflower. At that time, Mary had been so serious about her diet that she had refused to eat anything which was fried or cooked in grease.
Qualified Elaboration. Experiment 4	As she was waiting, Mary recalled that this had been her favorite restaurant because it had fantastic health food. Mary recalled that she had been a health nut and a strict vegetarian for about ten years but she wasn't anymore. Back then, her favorite food had been cauliflower. At that time, Mary had been so serious about her diet that she had refused to eat anything which was fried or cooked in grease.
Filler	After about 10 min, Mary's friend arrived. It had been a few months since they had seen each other. Because of this they had a lot to talk about and chatted for over a half hour. Finally, Mary signaled the waiter to come take their orders. Mary checked the menu one more time. She had a hard time deciding what to have for lunch.
Target sentence 1	Mary ordered a cheeseburger and fries.
Target sentence 2	She handed the menu back to the waiter.
Closing.	Her friend didn't have as much trouble deciding what she wanted. She ordered and they began to chat again. They didn't realize there was so much for them to catch up on.
Comprehension Question	Was Mary meeting her husband for lunch?

References

AlAfnan, M. A., Dishari, S., Jovic, M., & Lomidze, K. (2023). Chatgpt as an educational tool: Opportunities, challenges, and recommendations for communication, business writing, and composition courses. *Journal of Artificial Intelligence and Technology, 3*(2), 60–68.

Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Are you reading my mind? Modeling students' reading comprehension skills with natural language processing techniques. In *Proceedings of the fifth international conference on learning analytics and knowledge*. NY, USA: Poughkeepsie.

Andrus, B. R., Nasiri, Y., Cui, S., Cullen, B., & Fulda, N. (2022). Enhanced story comprehension for large language models through dynamic document-based knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*. Virtual Event.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., & Phang, J. (2022). *Gpt-neox-20b: An open-source autoregressive language model. arXiv preprint arXiv:2204.06745*.

Chowdhary, K., & Chowdhary, K. (2020). *Natural language processing*. Fundamentals of artificial intelligence.

Corlatescu, D.-G. (2023). *Automated model of comprehension version 4.0, notebook*. Retrieved July 17th from https://colab.research.google.com/drive/1xiNHhKbidwDOOs_JSDfZh-CXkNr4jGvw?usp=sharing.

Corlatescu, D.-G., Dascalu, M., & McNamara, D. S. (2021). *Automated model of comprehension V2.0*. Utrecht, The Netherlands: International Conference on Artificial Intelligence in Education.

Corlatescu, D.-G., Watanabe, M., Ruseti, S., Dascalu, M., & McNamara, D. S. (2023). *The automated model of comprehension version 3.0: Paying attention to context*. Tokyo, Japan: International Conference on Artificial Intelligence in Education.

Dai, Y., Liu, A., & Lim, C. P. (2023). Reconceptualizing ChatGPT and generative AI as a student-driven innovation in higher education. *Procedia CIRP, 119*, 84–90. https://doi.org/10.1016/j.procir.2023.05.002

Dascalu, M., Paraschiv, I.-C., McNamara, D. S., & Trausan-Matu, S. (2018). *Towards an automated model of comprehension (AMoC)*. Leeds, UK: European Conference on Technology Enhanced Learning.

Davoudi, M., & Moghadam, H. R. H. (2015). Critical review of the models of reading comprehension with a focus on situation models. *International Journal of Linguistics, 7*(5), 172–187.

Donkin, C., Nosofsky, R., Gold, J., & Shiffrin, R. (2015). Verbal labeling, gradual decay, and sudden death in visual short-term memory. *Psychonomic Bulletin & Review, 22*, 170–178.

Elleman, A. M., & Oslund, E. L. (2019). Reading comprehension research: Implications for practice and policy. *Policy Insights from the Behavioral and Brain Sciences, 6*(1), 3–11.

Farhana, E., Rutherford, T., & Lynch, C. F. (2022). Predictive student modelling in an online reading platform. In *Proceedings of the AAAI conference on artificial intelligence*. Virtual Event.

Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., & Lester, J. (2021). Predictive student modeling in game-based learning environments with word embedding representations of reflection. *International Journal of Artificial Intelligence in Education, 31*, 1–23.

Hernández-de-Menéndez, M., Morales-Menendez, R., Escobar, C. A., & Ramírez Mendoza, R. A. (2022). Learning analytics: State of the art. *International Journal on Interactive Design and Manufacturing, 16*(3), 1209–1230. https://doi.org/10.1007/s12008-022-00930-0

Honnibal, M., & Montani, I. (2017). Spacy 2: Natural language understanding with bloom embeddings. *Convolutional Neural Networks and Incremental Parsing, 7*(1).

Joksimović, S., Kovanović, V., & Dawson, S. (2019). The journey of learning analytics. *HERDSA Review of Higher Education, 6*, 27–63.

Keefe, D. E., & McDaniel, M. A. (1993). The time course and durability of predictive inferences. *Journal of Memory and Language, 32*(4), 446–463.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge university press.

Kintsch, W., & Welsch, D. M. (1991). The construction-integration model: A framework for studying memory for text. In *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock*. (pp. 367–385). Lawrence Erlbaum Associates, Inc.

Larrabee Sønderlund, A., Hughes, E., & Smith, J. (2019). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology, 50*(5), 2594–2618.

Lawless, C., Smee, P., & O'Shea, T. (1998). Using concept sorting and concept mapping in business and public administration, and in education: An overview. *Educational Research, 40*(2), 219–235.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1–35.

McNamara, D. S., Kintsch, E., Songer, N.-B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1–43.

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychology of Learning and Motivation, 51*, 297–384.

Meneghetti, C., Carretti, B., & De Beni, R. (2006). Components of reading comprehension and scholastic achievement. *Learning and Individual Differences, 16*(4), 291–301.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representation in vector space*. Scottsdale, AZ: Workshop at ICLR.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences, 11*(1), 237.

Nie, Y., Deacon, H., Fyshe, A., & Epp, C. D. (2022). Predicting reading comprehension scores of elementary school students. In A. Mitrovic, N. Bosch, A. I. Cristea, & C. Brown (Eds.), *Proceedings of the 15th international conference on educational data mining* (pp. 158–171). https://doi.org/10.5281/zenodo.6852952

O'Brien, E. J., Rizzella, M. L., Albrecht, J. E., & Halleran, J. G. (1998). Updating a situation model: A memory-based text processing view. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(5), 1200.

OpenAI. (2022). *Introducing ChatGPT*. Retrieved May 5th from https://openai.com/blog/chatgpt.

OpenAI. (2023a). GPT-4 technical report. https://doi.org/10.48550/arXiv.2303.08774.

OpenAI. (2023b). *Teaching with AI*. OpenAI. Retrieved September 1st from https://openai.com/blog/teaching-with-ai.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., & Ray, A. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*.

Patel, N., Nagpal, P., Shah, T., Sharma, A., Malvi, S., & Lomas, D. (2023). Improving mathematics assessment readability: Do large language models help? *Journal of Computer Assisted Learning, 39*(3), 804–822. https://doi.org/10.1111/jcal.12776

Pečjak, S., Podlesek, A., & Pirc, T. (2011). Model of reading comprehension for 5th grade students. *Studia Psychologica, 53*(1), 53–67.
Phillips, R. V., van der Laan, M. J., Lee, H., & Gruber, S. (2023). Practical considerations for specifying a super learner. *International Journal of Epidemiology, 52*(4), 1276–1285.
Pinheiro, J.-C., & Bates, D.-M. (2000). Linear mixed-effects models: Basic concepts and examples. *Mixed-effects models in S and S-Plus*, 3–56.
Potts, G. R., Keenan, J. M., & Golding, J. M. (1988). Assessing the occurrence of elaborative inferences: Lexical decision versus naming. *Journal of Memory and Language, 27*(4), 399–415.
Radford, A., Kim, J.-W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Clark, J. (2021). *Learning transferable visual models from natural language supervision. International conference on machine learning, 139*, 8748–8763.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.
Raudszus, H., Segers, E., & Verhoeven, L. (2019). Situation model building ability uniquely predicts first and second language reading comprehension. *Journal of Neurolinguistics, 50*, 106–119.
ReaderBench. (2023). *Automated model of comprehension version 4.0*. Retrieved July 17th from https://github.com/readerbench/AMoCv4.
Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data Mining and Knowledge Discovery, 10*(3), e1355.
Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in predictive learning analytics: A decade systematic review (2012–2022). *Education and Information Technologies, 28*(7), 8299–8333. https://doi.org/10.1007/s10639-022-11536-0
Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review, 46*(5), 30.
Susnjak, T. (2023a). Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and ChatGPT. *International Journal of Artificial Intelligence in Education*, 1–31.
Susnjak, T. (2023b). Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and ChatGPT. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00336-3
Van den Broek, P., Risden, K., Fletcher, C. R., & Thurlow, R. (1996). A "landscape" view of reading: Fluctuating patterns of activation and the construction of a stable memory representation. *Models of understanding text*, 165–187.
Van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The landscape model of reading: Inferences and the online construction of a memory representation. *The construction of mental representations during reading*, 71–98.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.-N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need. Advances in neural information processing systems*. CA, USA: Long Beach.
White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D.-C. (2023). *A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382*.
Ye, D. (2022). The history and development of learning analytics in learning, design, & technology field. *TechTrends, 66*(4), 607–615.
Zhang, W., & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. *Psychological Science, 20*(4), 423–428.
Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science, 6*(5), 292–297.